

Normal mucosal fibroblast remodeling during carcinogenesis promotes initial submucosal invasion of early-stage colorectal carcinoma

Shuichi Tsukamoto

Kobe University

Takaaki Nakanishi

Kobe University

Takayuki Kodama

Kobe University

Manabu Shigeoka

Kobe University

Yu-ichiro Koma

Kobe University

Shingo Sakashita

National Cancer Center

Genichiro Ishii

National Cancer Center Hospital East

Masafumi Horie

`mhorie@med.kobe-u.ac.jp`

Kobe University

Article

Keywords:

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10972596/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Article title: Normal mucosal fibroblast remodeling during carcinogenesis promotes initial submucosal invasion of early-stage colorectal carcinoma

Authors: Shuichi Tsukamoto, MD, PhD (kodama-827a@outlook.jp)^{1,2}, Takaaki Nakanishi, MD (nthawk@med.kobe-u.ac.jp)^{1,3}, Takayuki Kodama, MD, PhD (takodama@med.kobe-u.ac.jp)¹, Manabu Shigeoka, MD, PhD (mshige@med.kobe-u.ac.jp)⁴, Yu-ichiro Koma, MD, PhD (koma@med.kobe-u.ac.jp)^{1,5}, Shingo Sakashita, MD, PhD (ssakashi@east.ncc.go.jp)⁶, Genichiro Ishii, MD, PhD (gishii@east.ncc.go.jp)², and Masafumi Horie, MD, PhD (mhorie@med.kobe-u.ac.jp)¹

Affiliations:

1: Division of Molecular and Genomic Pathology, Department of Pathology, Kobe University, Graduate School of Medicine, Kobe, Hyogo, Japan

2: Department of Pathology and Clinical Laboratory, National Cancer Center Hospital East, Kashiwa, Chiba, Japan.

3: Department of Diagnostic Pathology, Kobe University Hospital, Kobe, Hyogo, Japan

4: Division of Pathology for Regional Communication, Department of Social/Community Medicine and Health Science, Kobe University, Graduate School of Medicine, Kobe, Hyogo, Japan

5: Division of Histopathology and Cytology, Department of Health Science, Kobe University, Graduate School of Medicine, Kobe, Hyogo, Japan

6: Division of Pathology, Exploratory Oncology Research and Clinical Trial Center, National Cancer Center, Kashiwa, Chiba, Japan.

Corresponding author contact details

Masafumi Horie, MD, PhD

Affiliation: Division of Molecular and Genomic Pathology, Department of Pathology, Kobe University, Graduate School of Medicine, Kobe, Hyogo 650-0017, Japan

Email: mhorie@med.kobe-u.ac.jp

Phone number: +81 78-382-5465

Abstract

Fibroblast remodeling during the onset of colorectal cancer invasion remains poorly understood. Flatter early-stage colorectal carcinomas (eCRCs) develop submucosal invasion at smaller tumor sizes than exophytic eCRCs, suggesting that fibroblast remodeling process differs according to tumor morphology. We investigated how cancer-associated fibroblast (CAF) states change during initial submucosal invasion and whether this process differs according to the morphology of eCRCs. Public single-cell RNA sequencing, spatial transcriptomic, and single-cell chromatin accessibility datasets were integrated, and the identified CAF states were mapped onto eCRC tissues using immunohistochemistry and RNA in situ hybridization. We identified normal mucosal fibroblasts (NMFs), two *PDPN*-high, *ACTA2*-low CAFs (PodCAFs1 and PodCAFs2), and matrix CAFs (mCAFs1). PodCAFs1 exhibited a favorable-prognosis-associated profile, whereas PodCAFs2 and mCAFs1 expressed programs tumor-promoting programs. Integrated analyses indicated a spatial and transcriptional progression from NMFs through PodCAFs1 to PodCAFs2 and mCAFs1. Histological mapping showed that NMF-, PodCAFs1-, and PodCAFs2/mCAFs1-associated profiles predominated in normal mucosa, carcinoma in situ, and submucosal invasion, respectively. Carcinoma in situ regions with PodCAFs2/mCAFs1-associated profiles were more extensive in flatter

eCRCs. These findings reveal spatially organized fibroblast remodeling during early colorectal cancer invasion and suggest that earlier progression toward later CAF states may contribute to the greater invasive propensity of flatter eCRCs.

Introduction

Colorectal cancer (CRC) is the third most frequently diagnosed cancer and the second leading cause of cancer-related death worldwide. According to GLOBOCAN 2024, CRC accounted for 9.9% of newly diagnosed cancers and 9.4% of cancer deaths globally.¹

Although chemotherapy, radiotherapy, and immunotherapy have improved outcomes for selected patients, complete removal of disease before regional or distant spread remains central to curative treatment. Advances in colonoscopy and endoscopic resection techniques, particularly endoscopic mucosal resection and endoscopic submucosal dissection, have enabled less invasive and potentially curative removal of many early-stage CRCs (eCRCs). A better understanding of the biological mechanisms governing CRC invasion could inform strategies to prevent or delay invasive progression, thereby increasing the proportion of lesions amenable to curative endoscopic treatment.

Cancer-associated fibroblasts (CAFs) are major cellular components of the tumor microenvironment (TME) and contribute to extracellular matrix (ECM) remodeling,

cancer cell invasion, immune regulation, and treatment resistance.²⁻⁴ CAFs produce abundant ECM to increase tissue stiffness and create physical and biochemical barriers to drug delivery and immune-cell infiltration. ECM remodeling can also promote cancer cell motility and invasiveness through processes associated with epithelial–mesenchymal transition (EMT).⁵ In addition, CAFs modulate local immune responses through cytokines and cell-cell interactions. In CRC, for example, interleukin-1 receptor 1 (IL1R1)-positive inflammatory CAFs recruit M2-like macrophages and T helper 17 cells and contribute to an immunosuppressive microenvironment.⁶ In rectal cancer, enrichment of interleukin-1-activated inflammatory CAFs has been associated with resistance to neoadjuvant chemoradiotherapy and cancer-cell survival.⁷

CAFs are heterogeneous populations rather than a single, uniformly tumor-promoting cell type. Normal fibroblasts or some CAF populations can restrain tumor progression or support tissue homeostasis, indicating that their functions depend on cellular state, tissue context, and disease stage^{8,9}. This functional diversity has stimulated efforts to therapeutically reprogram CAF phenotypes rather than eliminate CAFs indiscriminately. For example, the synthetic retinoid tamibarotene, also known as Am80, can promote a transition from alpha-smooth muscle actin (α SMA)-expressing CAFs toward Meflin-expressing fibroblasts, thereby improving drug delivery and restraining tumor growth in

preclinical models.¹⁰ Defining the cellular origins, spatial distributions, and regulatory programs of individual CAF populations is therefore essential for understanding their biological and translational significance.

Most studies of the TME have focused on advanced cancers, in which fibrotic or desmoplastic stroma is readily apparent. However, stromal cell populations can already acquire disease-associated phenotypes in dysplastic or non-invasive lesions, suggesting that the TME develops progressively during the early phases of carcinogenesis.¹¹⁻¹³

Defining the accompanying changes in stromal cell states, particularly the transition of normal fibroblasts into functionally distinct CAF populations, could provide insight into the eCRC invasion and identify biological processes that might eventually be targeted to prevent or delay invasive progression.

In our previous study, we found that normal mucosal fibroblasts (NMFs), including pericryptal fibroblasts adjacent to colorectal epithelial crypts, contribute to the CAF population in eCRC.¹⁴ Following the development of adenoma or carcinoma in situ (CIS), a subset of these fibroblasts acquired podoplanin (PDPN) expression with a concomitant reduction in alpha-smooth muscle actin (α SMA) expression. Because this expression pattern was closely associated with neoplastic lesions, we termed these cells PDPN-positive CAFs (PodCAFs). Although their potential diagnostic value was demonstrated,

the phenotypic diversity, biological functions, and regulatory mechanisms of PodCAFs during eCRC progression remained unclear.

In this study, we combined histological evaluation of human eCRC tissues with integrated single-cell RNA sequencing (scRNA-seq), spatial transcriptomics (ST), and single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq). We aimed to define the heterogeneity and spatial organization of NMF-derived CAF populations, to characterize their changes during the progression from non-neoplastic mucosa through CIS to submucosally invasive CRC, and to identify the transcriptional and epigenetic programs accompanying their inferred state transitions. Our findings delineate distinct NMF-derived CAF populations associated with CIS and submucosal invasion and provide a framework for understanding how stromal remodeling accompanies the acquisition of invasive properties in CRC.

Results

Stromal PDPN and α SMA expression patterns differ according to the macroscopic morphology of eCRC

We previously reported that α SMA was expressed in NMFs, particularly pericryptal fibroblasts, but showed reduced expression or a more limited stromal distribution in CIS,

whereas PDPN was absent from normal mucosa but expressed in most CIS lesions.¹⁴

During that study, PDPN and α SMA showed reciprocal stromal expression patterns according to tumor morphology: flatter lesions generally exhibited higher PDPN and lower α SMA expression, whereas exophytic lesions showed the opposite pattern (Figure 1a). Across 45 eCRC cases, PDPN and α SMA expression in CIS region were inversely correlated (Pearson's $r = -0.50$, $p < 0.001$, Figure 1b). This observation prompted us to investigate the association between macroscopic morphology and CAF marker expression in eCRC. To further evaluate this association, we established four macroscopic classification systems for eCRC, as described in the Methods section (Figure 2a). Stromal PDPN expression was significantly higher in Sessile or Flat lesions than in Pedunculated lesions, in Flat lesions than in Purely Elevated lesions, in Non-polypoid lesions than in Polypoid lesions, and in Single-based Sessile lesions than in Pedunculated or Granular lesions (Figure 2b).

Macroscopic morphology has been reported to be associated with distinct biological and clinical characteristics of colorectal cancer, and flatter lesions tend to exhibit SM inv.^{15,16}

We therefore analyzed 349 eCRCs resected by endoscopy or surgery and assessed clinicopathological parameters, including tumor size, morphology, and the presence of SM inv (Table 1).

The frequency of SM inv was significantly higher in Flat, Single-based Sessile, and Non-polypoid lesions than in their respective counterpart morphology groups, whereas a similar trend was observed for Sessile or Flat lesions compared with Pedunculated lesions, in line with the reports by other researchers (Table 2) ^{15,16}. To examine the relationship between tumor size and SM inv, cases were grouped into 10-mm tumor size bins and stratified according to macroscopic morphology (Figure 2c-f). Across all four classification systems, lesions with flatter morphology showed a greater proportion of SM inv at smaller tumor sizes.

Predicted probabilities of SM inv were estimated using logistic regression models incorporating tumor size, classification group, and their interaction term (Figure 2g). A significant interaction between morphology and tumor size was identified only in the Sessile or Flat versus Pedunculated comparison. Therefore, size-specific odds ratios (ORs) were calculated for this classification. Sessile or Flat lesions showed higher odds of SM inv than Pedunculated lesions at smaller tumor sizes, with ORs of 3.09 at 10 mm and 2.06 at 15 mm (Figure 2h, Supplementary Data S1). The OR decreased with increasing tumor size, approached unity at approximately 20-25 mm, and reached 0.27 at 40 mm.

For the remaining three macroscopic classification systems, no significant interaction

between morphology and tumor size was detected; therefore, single size-adjusted ORs were estimated. Flat lesions had significantly higher odds of SM inv than Purely Elevated lesions (OR 4.46, 95% CI 2.53-8.05; $p < 0.001$). Likewise, Single-based Sessile lesions and Non-polypoid lesions showed higher odds of SM inv than their respective counterpart groups, Pedunculated or Granular lesions and Polypoid lesions (OR 2.26, 95% CI 1.42-3.64; $p < 0.001$ and OR 2.14, 95% CI 1.36-3.39; $p = 0.001$, respectively; Figure 2i, Supplementary Data S2). Since stromal PDPN expression was enriched in flatter lesions and these lesions exhibited a higher frequency of SM inv, we next investigated the potential association between PDPN-positive CAFs (PodCAFs) and the early invasive behavior of eCRC.

scRNA-seq analysis identifies molecularly distinct CAF populations in CRC

To characterize the molecular and functional features of CAFs, we analyzed three publicly available CRC single-cell RNA sequencing (scRNA-seq) datasets: GSE132465, GSE144735, and GSE178341^{17,18}. Fibroblast populations were extracted from each dataset and integrated using an anchor-based approach. A total of 12,696 fibroblasts, including 4,596 CAFs, were subjected to downstream analyses (Figure 3a and Supplementary Figure S1a).

After normalization and dimensionality reduction, unsupervised clustering identified 18 fibroblast clusters and 13 CAF clusters (Figure 3b and Supplementary Figure S1b). Cluster-specific differentially expressed genes (DEGs) were identified using the FindAllMarkers function. Representative marker genes included *COL11A1*, *COL1A1*, and *POSTN* in cluster 0, *SOX6* and *CXCL14* in cluster 1, and *MMP3* and *WNT5A* in cluster 2 (Figure 3c and Supplementary Data S3).

To annotate the CAF clusters, we applied the CAF classification framework proposed by Cord and colleagues using the AUCell package.¹⁹ Cluster 1 could not be assigned unambiguously using AUCell alone and was subsequently designated PodCAFs1 based on its high *PDPN* and low *ACTA2* expression, a defining characteristic of PDPN-positive CAFs in our previous study.¹⁴ Although cluster 2 showed enrichment for the tumor-like CAFs (tCAFs) signature, it was designated PodCAFs2 because it also showed high *PDPN* and low *ACTA2* expression. Cluster 3 showed a high reticular CAF (rCAFs) enrichment score but was defined as *intermediate_fibro* because it showed minimal expression of the reticular-cell markers such as *CCL19*, *CCL21*, and *CXCL13* and was abundant in normal colorectal tissues (Supplementary Figure S1b). Finally, clusters 0–12 were designated as mCAFs1, PodCAFs1, PodCAFs2, *intermediate_fibro*, iCAFs1, mCAFs2, dCAFs, iCAFs2, ifnCAFs, vCAFs, hsp_tCAFs, apCAFs1, and apCAFs2, respectively (Figure 3d

and Supplementary Figure S1c, d). Normal fibroblast clusters were designated NF1–NF10 according to the number of cells in each cluster (Figure 3d). From here, we focused on molecular profiles of PodCAFs1/2 and their neighboring clusters, mCAFs1/2 (Figure 3e).

Although the relative abundance of CAF clusters varied among the original datasets, mCAFs1 and PodCAFs1/2 consistently constituted major CAF populations across all three datasets (Supplementary Figure S1e). We next compared the identified CAF populations with those reported previously. PodCAFs2 showed high *WNT5A* expression and transcriptional similarities to the *PDPN*^{high}/*WNT5A*^{high} CAF population described as “inflammatory fibroblast (InfFib)” by Harada and colleagues, who implicated this population in the development of colitis-associated carcinoma.²⁰ In addition, PodCAFs1 and intermediate_fibro cells showed high expression of *SOX6* and *ADAMDEC1*, respectively, corresponding to the *SOX6*^{high} and *ADAMDEC1*^{high} CAF populations reported by Ku and colleagues to have tumor-restraining roles.²¹

To investigate the relationship between CAFs and NMFs, we examined the expression of mucosal fibroblast marker genes across all fibroblast populations. Crypt-top markers, including *PDGFRA* and *CXCL14*, were strongly expressed in NF4 and PodCAFs1; therefore, NF4 was designated as the NMF population. In contrast, crypt-bottom markers,

including *RSPO3* and *SFRP1*, were strongly expressed in NF1, NF2, and iCAFs1/2. Among the CAF populations, PodCAFs2 showed the highest *WNT5A* expression (Supplementary Figure S1f).

Finally, the expression of PodCAFs-associated genes, including *CXCL14* and *WNT5A*, was positively correlated with each other across CAFs. Consistent with the immunohistochemical findings shown in Figure 1 and our previous observations,¹⁴ the expression of PodCAFs-associated genes was inversely correlated with *ACTA2* expression across CAFs (Supplementary Figure S1g).

Distinct prognostic and functional gene-expression profiles of PodCAFs1, PodCAFs2, and mCAFs1

We focused on PodCAFs1, PodCAFs2, and mCAFs1 because they were major CAF populations across the integrated datasets. We compared established favorable- and poor-prognosis-associated CAF marker gene expressions and a calculated module score for the published low-risk gene set (LR gene set) (Supplementary Figure S2)²¹. *PRRX1* and *TNC*^{22,23} were highly expressed in mCAFs1, PodCAFs2, dCAFs, and iCAFs1/2. By contrast, *CXCL14* and the LR gene set score were higher in intermediate_fibro and PodCAFs1. *LRRC15*²⁴ showed low expression in PodCAFs1, PodCAFs2, and mCAFs1, whereas

Ptll6^{25,26} was predominantly expressed in iCAFs1/2 (Figure 4a). *PRRX1* expression was inversely correlated with the LR gene-set score across individual cells (Pearson's $r = -0.49$, $p < 0.001$, Figure 4a). These profiles distinguished PodCAFs1 from PodCAFs2 and mCAFs1.

Hallmark GSEA showed significant positive enrichment of EMT in mCAFs1 (NES 2.42; FDR 0.0279) and of hypoxia, inflammatory response, and glycolysis (NES 2.13, 1.82, and 1.67; FDR 0.0120, 0.0120, and 0.036, respectively) (Figure 4b, c, d and Supplementary Data S5).

Gene Ontology enrichment mapping identified programs for response to external stress and anatomical structure development in mCAFs1 and for responses to chemical stimuli, toxic substances, metals, and bacterial products in PodCAFs2 among the positively enriched processes (Figure 4e, f and Supplementary Data S6).

PodCAFs1 showed no significantly enriched Hallmark gene sets, and its limited significantly enriched Gene Ontology terms did not form a coherent functional module (Supplementary Data S6). Thus, its transcriptional characteristics were not readily captured by the pathway-level signatures examined.

Integration of scRNA-seq and spatial transcriptomic data reveals distinct

distributions of CAF states in eCRC

Since the scRNA-seq datasets were derived predominantly from advanced CRCs, the distribution of the identified fibroblast states within eCRC tissues remained unclear. We therefore integrated GSE132465 with six eCRC Visium ST datasets using DeepCOLOR to deconvolve the spatial transcriptomic data and infer the distributions of individual scRNA-seq cells across histologically defined regions (Figure 5a). The Visium sections were segmented into normal mucosa, submucosa and muscularis propria, adenoma, CIS, or SM inv (Supplementary Figure S3a). Within each major cell-type category, cells in the upper 10% of DeepCOLOR spatial-assignment proportions for each tissue region were selected as top10% cells. Selection was independent among regions, allowing the same cell to be assigned to more than one region. Overall, 29,052 cell–region assignments met this criterion (Supplementary Figure S3b). Co-localization clusters 0, 1, 2, 3, 5, and 8 contained both epithelial cells and fibroblasts and were selected for subsequent analyses (Figure 5b, c, Supplementary Figure S3c, and Supplementary Data S7). NF4 and PodCAFs1 cells among top10% cells constituted large proportions of fibroblasts assigned to adenoma and CIS, whereas PodCAFs2, mCAFs1, and mCAFs2 cells among top10% cells were more prominent in SM inv. Other top10% CAF populations were uncommon in these eCRC regions (Figure 5d).

Top10% stromal cells in the selected co-localization clusters showed significantly higher expression of genes characteristic of PodCAFs1, PodCAFs2, or mCAFs1, including *WNT5A*, *POSTN*, and *PDPN*, than did top10% stromal cells in the remaining clusters (Figure 5e). In a representative Visium section, NF4 was distributed from normal mucosa toward the CIS-SM inv interface, and PodCAFs1 was concentrated near this interface. PodCAFs2 and mCAFs1 showed higher spatial-assignment scores within SM inv, consistent with imputed expression of their marker genes (Figure 5f). The preceding immunohistochemical, scRNA-seq, and ST findings suggested a continuum of fibroblast states from NF4/NMF through PodCAFs1 to mCAFs1. To examine this continuum, we performed pseudotime trajectory analyses using the scRNA-seq data. Monocle 3 inferred a pseudotime trajectory originating from NF4 in the scRNA-seq dataset. The inferred trajectory extended from NF4 to PodCAFs1 and subsequently branched toward PodCAFs2, mCAFs1, and mCAFs2 (Figure 5g). *PDPN* expression increased during the earlier part of the trajectory, corresponding to the progression from NF4 toward PodCAFs1. *ACTA2* and *POSTN* expression initially decreased and subsequently increased during later pseudotime, coinciding with the progression from PodCAFs1 and PodCAFs2 toward mCAFs1. *CXCL14* expression decreased during intermediate pseudotime, whereas *FAP* and *MMP3* expression emerged or increased near the middle

of the trajectory, corresponding to the progression from PodCAFs1 toward PodCAFs2 and mCAFs1 (Figure 5h).

Bulk RNA-seq data from TNMplot provided an independent comparison with a relatively early invasive setting. *PDPN*, *WNT5A*, *PRRX1*, *MMPI*, *MMP3*, and *FAP* expression was higher in stage I CRC than in normal colorectal tissue, whereas *CXCL14* was lower and *ACTA2* and *POSTN* differed less markedly (Supplementary Figure S4). These patterns were broadly consistent with the scRNA-seq and ST findings.

DeepCOLOR predicted ligand-mediated interactions involving PodCAFs1, PodCAFs2, and mCAFs1. *CXCL14* was among the predicted mediators of PodCAFs1 interactions with carcinoma and immune cells, whereas PodCAFs2- and mCAFs1-derived chemokines, transforming growth factor (TGF- β) superfamily members, and ECM-associated molecules were predicted to interact with carcinoma, endothelial, immune, and SPP1-expressing macrophage populations (Supplementary Figure S3d, e).

Integrated scRNA-seq and scATAC-seq analyses support a continuous trajectory from NF4/NMF to mCAFs1

Using GSE201336 scATAC-seq dataset,²⁷ we examined the corresponding changes in chromatin accessibility and investigate the regulatory changes associated with these

fibroblast states. Initial cell clusters were broadly annotated as epithelial, stromal, T or natural killer, B, plasma, monocyte or dendritic, and endothelial cells based on gene activity scores for representative marker genes. The stromal cell cluster was extracted and re-clustered. Cells with gene activity for *SI00B*, representing glial or neuronal cells, and those with gene activity for *CDH5*, representing endothelial cells, were excluded. After these exclusions, 20,334 fibroblasts were retained for subsequent analyses (Supplementary Figure S5a, b). The re-clustered scATAC-seq fibroblast dataset was integrated with the scRNA-seq fibroblast dataset. Fibroblast annotations were transferred to the scATAC-seq cells by comparing scATAC-seq-derived gene activity scores with scRNA-seq gene-expression profiles (Figure 6a, b). The integrated analysis identified an ordered series of NF4, PodCAFs1, PodCAFs2, and mCAFs1 states, which was used to construct an scATAC-seq pseudotime trajectory from NF4 to mCAFs1. No scATAC-seq population was annotated as mCAFs2 (Figure 6a).

ChromVAR analysis identified distinct TF motif-accessibility profiles across the fibroblast states. FOX-family motifs predominated in NF4, whereas PodCAFs1 showed a broader profile including selected FOX, AP-1, RUNX, and BACH motifs. AP-1 and BACH motifs were strongly enriched in PodCAFs2 and mCAFs1, together with RUNX, BATF, and NFAT motifs in mCAFs1 (Supplementary Figure S6).

Across the pseudotime, FOX-family motif deviation scores decreased, whereas NF- κ B-, AP-1-, BACH-, and C/EBP-associated scores began to increase in PodCAFs1 and reached higher levels in PodCAFs2 and mCAFs1 (Figure 6c).

To prioritize candidate regulators, motif deviation scores were correlated with expression of the corresponding TF genes and evaluated in relation to state-dependent changes in motif accessibility (Figure 6d). FOSL1, FOSL2, JUND, and BACH1 showed particularly large changes together with positive motif-expression correlations. FOXF1, FOXF2, MSC, CEBPA, and RUNX family members also showed positive correlations but smaller accessibility changes.

We next examined chromatin accessibility and marker-gene activity along the scATAC-seq pseudotime trajectory. Peak-to-gene linkage analysis identified accessible chromatin peaks whose accessibility was significantly associated with the activity of the linked genes. We focused on linked peaks showing marked accessibility changes between NF4 and PodCAFs1 or between PodCAFs1 and mCAFs1. Among genomic regions linked to *PDPN*, the peak at chr1:13,593,375–13,593,875 remained accessible throughout pseudotime, whereas peaks at chr1:13,583,491–13,583,991, chr1:13,605,103–13,605,603, and chr1:13,616,117–13,616,617 gained accessibility during later pseudotime. These changes accompanied increased *PDPN* gene activity. *ACTA2* gene activity decreased

from NF4/NMF to PodCAFs1 and subsequently increased toward mCAFs1, accompanied by corresponding accessibility changes at the linked peak at chr10:88,952,570–88,953,070. *FAP* gene activity remained low throughout most of the NF4/NMF and PodCAFs1 states and increased toward PodCAFs2 and mCAFs1. Increased accessibility of linked peaks at chr2:162,244,448–162,244,948 and chr2:162,244,974–162,245,474 preceded the increase in *FAP* gene activity. JASPAR motif analysis identified candidate transcription factor-binding motifs within these peaks. By contrast, accessibility of peaks linked to *CXCL14* decreased during later pseudotime. The peak at chr5:135,578,797–135,579,297 showed a marked decline in accessibility that paralleled the decrease in *CXCL14* gene activity (Figure 6e; Supplementary Figures S7 and S8). TF footprinting showed higher flanking accessibility for FOXF1, FOXF2, and MSC motifs in NF4, and for FOSL1, FOSL2, BACH1, and JUND motifs in PodCAFs2 and mCAFs1 (Figure 6f). Together, these analyses identified state-associated TF programs and prioritized candidate regulators along the inferred trajectory.

Histological marker profiles across eCRC progression correspond to the inferred ordering of fibroblast states

Stromal areas positive for PDPN, α SMA, POSTN, and FAP were visually estimated in

normal mucosa, CIS, and SM inv (Figure 7a, b). PDPN was generally low in normal mucosa and high in CIS and SM inv, though slightly decreased in SM inv. α SMA was high in normal mucosa, variable in CIS, and high in SM inv. POSTN was detected in normal mucosa, lower in CIS, and high in SM inv, whereas FAP was low in normal mucosa and CIS but high in SM inv.

CXCL14 expression was detected in small rounded mucosal fibroblasts in normal mucosa and CIS. In SM inv, round inflammatory cells were positive, whereas spindle-shaped stromal cells showed no detectable immunoreactivity. Because positive fibroblasts could not be reliably distinguished from round inflammatory cells for area-based quantification, CXCL14 expression was evaluated qualitatively (Supplementary Figure S9a). These marker profiles were broadly concordant with the inferred ordering of NF4, PodCAFs1, and mCAFs1.

To distinguish PodCAFs2- and mCAFs1-associated profiles, serial sections were examined for FAP, POSTN, MMP3, and *WNT5A* expression. Spatially corresponding FAP and POSTN expression was focal in CIS but widespread in SM inv. Most positive regions showed little MMP3 or *WNT5A* expression and were consistent with an mCAFs1-associated profile. Small pericryptal regions near initial invasion additionally showed MMP3 expression and *WNT5A* signals, consistent with a PodCAFs2-associated profile

(Supplementary Figure S9b).

FAP was absent or low in most normal and CIS stroma but prominent around shallowly invasive glands and throughout deeper SM inv stroma (Figure 7c). The horizontal extents of carcinoma, FAP-positive stroma, and SM inv were quantified as L(carcinoma), L(FAP), and L(SM inv), respectively (Figure 7d). L(FAP)/L(carcinoma) was positively correlated with L(SM inv)/L(carcinoma) (Pearson's $r = 0.7880$, $p < 0.001$, Figure 7e) and was significantly greater than L(SM inv)/L(carcinoma) (Figure 7f). FAP-positive stromal regions extending into CIS were designated FAP-positive CIS (FAPposCIS).

The relative extent of FAPposCIS, L(FAPposCIS)/L(carcinoma), was significantly greater in Flat than in Purely Elevated eCRCs. Similar directional differences were observed in the other three macroscopic classification systems, although these did not reach statistical significance (Figure 7g).

Collectively, the histological distribution of marker-defined fibroblast populations corresponded to the ordering inferred from scRNA-seq, Visium ST, and scATAC-seq analyses. An mCAFs1-associated profile was present in focal CIS regions and became widely distributed throughout SM inv stroma, whereas PodCAFs2-associated profiles were confined to small pericryptal regions near initial invasion (Figure 8).

Discussion

Our findings identify differences in SM inv and CAF marker-expression profiles between flatter and exophytic eCRCs. Integrated scRNA-seq, Visium ST, and scATAC-seq analyses supported an inferred ordering of fibroblast states from NF4/NMF through PodCAFs1 to PodCAFs2 and mCAFs1. Fibroblast-state remodeling in flatter eCRCs had progressed further along this ordering at a smaller tumor size, suggesting a microenvironmental basis for their distinct invasive behavior.

Previous studies have reported both genomic and clinical differences between flatter and exophytic eCRCs. Flatter eCRCs have been reported to show frequent loss of heterozygosity at chromosome 3p and relatively infrequent activating *KRAS* mutations, whereas activating *KRAS* mutations are more frequent in exophytic eCRCs.^{28,29} Clinically, flatter eCRCs are known to develop SM inv at a smaller tumor size than exophytic lesions,^{15,16} an association reproduced across four macroscopic classification systems in our cohort of 349 eCRCs. Despite these established genomic and clinical differences, corresponding differences in the TME remain largely unexplored.

Flatter eCRCs showed broader stromal PDPN expression and lower α SMA expression than exophytic ones, consistent with the previously reported retention of α SMA in exophytic eCRCs and its reduction in flatter ones.³⁰ These findings indicate that flatter

and exophytic eCRCs differ in their stromal profiles as well as in their genomic and clinical features. Because cancer invasion depends on interactions between neoplastic cells and stromal populations, including tumor-associated macrophages^{31,32} and CAFs,^{33,34} we hypothesized that different CAF profiles resulted in different SM inv propensity of eCRC. We therefore used single-cell and spatial omics analyses to resolve the CAFs' (epi)genetic and functional heterogeneity underlying these histological observations.

Among several scRNA-seq based CAFs classification systems,^{35,36} we applied the classification developed by Cords and colleagues in breast cancer and subsequently evaluated in other cancers, including CRC.³⁷ PodCAFs1 did not correspond closely to these classes but expressed NMF-associated genes, such as *PDGFRA* and *CXCL14*, suggesting a CAF state characteristic of the colorectal tissue context. CXCL14-expressing CAFs with potential tumor-restraining effects have recently been reported in CRC and were characterized by a low-risk gene set including *CXCL14* and *ADAMDECI*.²¹ In our data, their co-expression characterized intermediate_fibro, whereas PodCAFs1 expressed *CXCL14* with little or no *ADAMDECI*. Although the effects of CXCL14 are context dependent,^{38,39} PodCAFs1 expressed other low-risk genes and retained NMF-like features. Together with evidence that normal fibroblasts can

restrain tumor progression,⁸ these findings suggest that PodCAFs1 represent a potentially tumor-restraining, neoplasia-associated state of NMFs. Their functional effects nevertheless remain to be determined. By contrast, *PRRX1*, a regulator of tumor-promoting CAF programs,²² was expressed at low levels in PodCAFs1 but at high levels in PodCAFs2 and mCAFs1. PodCAFs2 expressed factors associated with the senescence-associated secretory phenotype, including *CXCL8*, *MMP3*, and *IL11*,^{40,41} whereas mCAFs1 expressed abundant collagen genes and TGF- β superfamily members. These programs have been implicated in ECM remodeling, cancer cell invasion, and EMT.⁴²⁻⁴⁵ PodCAFs1, PodCAFs2, and mCAFs1 therefore exhibited transcriptional profiles consistent with different roles during eCRC progression.

Integration with Visium ST data using DeepCOLOR⁴⁶ enabled CAF states defined predominantly from advanced CRC samples to be examined in eCRC tissues. NF4 and PodCAFs1 were prominent in adenoma and CIS, whereas PodCAFs2 and mCAFs1 were more abundant in SM inv. Other CAF states, including iCAFs, vCAFs, ifnCAFs, and hsp_tCAFs, were uncommon in eCRC and transcriptionally distinct from these populations. These states might emerge during more advanced remodeling or arise from other cellular sources.

Integrated scRNA-seq and scATAC-seq analyses supported an inferred trajectory from

NF4 through PodCAFs1 and PodCAFs2 to mCAFs1. Although PodCAFs1 showed little *FAP* expression, accessibility of a *FAP*-linked peak had already increased, suggesting an intermediate state in which part of the chromatin program associated with later CAF states precedes robust *FAP* expression. Along the same trajectory, FOX-family motif deviation scores decreased, whereas motifs associated with AP-1, BACH, and C/EBP family members increased toward PodCAFs2 and mCAFs1. Concordant motif accessibility, TF expression, and footprinting profiles implicated these TF families in later CAF-associated programs, although the contributions of individual members remain unresolved.

Few studies have examined CAF changes during progression from non-neoplastic state to initial submucosally invasive carcinoma. A strength of the present study is the direct mapping of molecularly defined fibroblast states onto human eCRC tissues, where the largely preserved mucosal architecture allows progression from non-neoplastic mucosa to initial invasion to be traced histologically and interpreted in relation to neoplastic glands. Our histological mapping showed spatially ordered but overlapping CAF profiles across normal mucosa, CIS, and SM inv. Corresponding POSTN and FAP expression was focal in CIS but widespread in SM inv stroma. Most positive regions showed little MMP3 or *WNT5A* expression and were consistent with mCAFs1, whereas small pericryptal regions near initial invasion additionally expressed MMP3 and *WNT5A*, consistent with

PodCAFs2. Harada and colleagues described hypoxia-induced PDPN- and WNT5A-positive inflammatory fibroblasts, termed InfFibs, that promoted CRC development and were broadly distributed along the luminal surface of advanced CRCs.²⁰ Although InfFibs appear transcriptionally similar to PodCAFs2, the PodCAFs2-associated profile was localized near initial invasion in our eCRC tissues, suggesting that its distribution and contribution to tumor progression might expand with disease progression and desmoplastic remodeling. By contrast, mCAFs1 resembled the tumor-promoting THBS2-expressing matrix CAFs reported by Liu and colleagues.⁴² The broad distribution of mCAFs1-associated profiles in POSTN- and FAP-positive stroma suggests that ECM production by this population might contribute to an invasion-permissive environment through TGF- β signaling and collagen-mediated alterations in epithelial polarity.^{47,48}

Flatter eCRCs had a greater horizontal extent of FAP-positive CIS than Purely Elevated lesions. This finding suggests that intramucosal stromal remodeling is more advanced in some flatter lesions and provide a possible microenvironmental explanation for their tendency to develop SM inv at a smaller tumor size. Moreover, the remodeling may have a critical role in initiation SM inv of eCRC.

This study has several limitations. The upstream signals and specific TFs that regulate the inferred fibroblast-state changes were not determined, although integrated chromatin-

accessibility, motif–expression correlation, and footprinting analyses identified AP-1, BACH, and C/EBP family members as candidate regulators. In addition, the cross-sectional tissue analyses and absence of perturbation experiments precluded direct assessment of fibroblast lineage relationships and causal functions. Finally, we focused on NMF-derived fibroblast states and did not evaluate other potential CAF sources, including adipocytes,⁴⁹ pericytes,⁵⁰ and bone-marrow-derived stromal cells.⁵¹

Defining the signals that regulate these fibroblast states could identify opportunities to preserve tumor-restraining stromal properties or prevent CAF programs associated with SM inv. Such knowledge might ultimately provide a basis for delaying invasion and increasing the proportion of colorectal neoplasms amenable to curative endoscopic treatment.

Methods

eCRC samples for evaluating submucosal invasion development

We retrospectively identified 349 sporadic eCRC lesions from 310 patients that were unrelated to inflammatory bowel disease and had been resected endoscopically or surgically at Kobe University Hospital between January 2010 and March 2025. Histological types were restricted to well- or moderately differentiated adenocarcinoma, diagnosed based on severe cytological and/or structural atypia. All specimens were fixed

in 10% neutral-buffered formalin, embedded in paraffin, sectioned at a thickness of 4 μ m, and evaluated histologically.

For this study, eCRC was defined as carcinoma in situ (CIS) or carcinoma invading the submucosa, arising in the colorectum, corresponding to pathological tumor stages pTis and pT1, respectively, in the eighth edition of the Union for International Cancer Control TNM classification.⁵² CIS corresponds to high-grade dysplasia in the fifth edition of the WHO Classification of Tumors of the Digestive System.⁵³

The macroscopic type of each lesion was recorded according to the ninth edition of the Japanese Classification of Colorectal, Appendiceal, and Anal Carcinoma.⁵⁴ To capture the morphological contrast between flatter and more exophytic lesions, eCRCs were grouped using four prespecified binary macroscopic classification systems: (1) Purely Elevated, comprising lesions with only 0-I and/or 0-IIa components, versus Flat; (2) Polypoid, comprising lesions containing a 0-I component, versus Non-Polypoid; (3) Pedunculated or Granular, comprising 0-Ip or 0-Isp lesions or lesions with multiple 0-Is and/or 0-IIa components, versus Single-based Sessile; and (4) Pedunculated, comprising 0-Ip or 0-Isp lesions, versus Sessile or Flat. In each system, the former the more exophytic morphology and the latter category represented the flatter morphology. For all 349 eCRCs, anatomical location, macroscopic classification, tumor size, presence or absence of submucosal

invasion, and resection method were recorded.

Statistical modeling of submucosal invasion according to eCRC macroscopic classification

Logistic regression was used to assess the associations between eCRC macroscopic classification and submucosal invasion (SM inv). The binary outcome was the presence or absence of SM inv. For each of the four macroscopic classification schemes, morphology was included as the primary explanatory variable and tumor size, expressed in mm, as a continuous covariate. Two models were fitted for each classification scheme: a main-effects model including morphology and tumor size, and an interaction model additionally including a morphology-by-tumor-size interaction term. The interaction model was compared with the corresponding main-effects model using a likelihood-ratio test.

When there was no evidence of a morphology-by-tumor-size interaction, the association between morphology and SM inv was reported as a tumor-size-adjusted odds ratio (OR) with a 95% confidence interval (CI). When evidence of an interaction was detected, the association was not summarized using a single OR; instead, model-predicted probabilities of SM inv and size-specific ORs with 95% CIs were calculated across the observed range

of tumor sizes. All tests were two-sided, and $p < 0.05$ was considered statistically significant. Analyses were performed using R v.4.3.3, and plots were generated using ggplot2 v.3.5.2.

RNAscope in situ hybridization

RNA in situ hybridization was performed using the RNAscope 2.5 HD Reagent Kit-RED (Advanced Cell Diagnostics [ACD], Newark, CA, USA; catalog number 322350), a positive-control probe (targeting *PPIB*; ACD; catalog number 313901), and a target probe against *WNT5A* (ACD; catalog number 604921), in accordance with the manufacturer's protocol. After staining, whole-slide images were acquired using a NanoZoomer virtual slide scanner (Hamamatsu Photonics, Shizuoka, Japan) and reviewed using NDP.view 2 software (Hamamatsu Photonics).

Immunohistochemistry

Immunohistochemical analyses were performed using the same case series in which stromal PDPN expression had been evaluated in our previous study.¹⁴ Of the 73 eCRC cases in this series, 45 were additionally examined for α SMA expression. Antibodies used for immunohistochemistry are as follows: a mouse monoclonal antibody against PDPN (clone D2-40; dilution 1:4; Agilent Technologies, Santa Clara, CA, USA); a mouse monoclonal antibody against α SMA (clone 1A4; dilution 1:1; Agilent Technologies).

Both antibodies were applied using a VENTANA BenchMark ULTRA automated staining system (Roche, Basel, Switzerland). Stromal PDPN and α SMA expression was evaluated as described below, and the proportions of the positive stromal areas were compared.

Of these 45 cases, 27 were additionally examined by immunohistochemistry using antibodies against C-X-C motif chemokine ligand 14 (CXCL14; rabbit polyclonal; dilution 1:100; catalog number 10468-1-AP; Proteintech, Rosemont, IL, USA), matrix metalloproteinase 3 (MMP3; mouse polyclonal; dilution 1:500; catalog number 66338-1-Ig; Proteintech), periostin (POSTN; rabbit polyclonal; dilution 1:800; catalog number ab14041; Abcam, Cambridge, UK), and fibroblast activation protein (FAP; rabbit monoclonal; clone SP355; dilution 1:75; catalog number ab227703; Abcam).

Immunoreaction was performed using a Leica Bond-MAX automated staining system (Leica Biosystems, Wetzlar, Germany), with 3,3'-diaminobenzidine (Dako Cytomation, Glostrup, Denmark) as the chromogen. Whole-slide images were acquired using a NanoZoomer virtual slide scanner (Hamamatsu Photonics, Shizuoka, Japan) and reviewed using NDP.view 2 software (Hamamatsu Photonics).

Immunohistochemistry image analyses

The non-neoplastic and neoplastic regions of each tissue section were evaluated in their entirety by a single gastrointestinal pathologist (ST). For PDPN and FAP, stromal areas

showing greater staining intensity than the corresponding normal colorectal lamina propria were considered positive because PDPN and FAP expression was absent or low in the normal lamina propria. For α SMA and POSTN, stromal areas showing staining intensity equivalent to or greater than that in the normal lamina propria were considered positive because these proteins were consistently expressed in normal mucosal fibroblasts (NMFs). The proportion of the positive stromal area was visually estimated using the total evaluable stromal area in each region as the denominator. For each case, stromal marker expression was evaluated separately in the normal mucosa, CIS, and SM inv regions. Within each region, expression was classified as high when the positive area constituted 50% or more of the evaluable stromal area and as low when it constituted less than 50%. A length-based method was additionally developed to quantify the extent of FAP immunohistochemistry images. Because the carcinoma and FAP-positive stromal regions could be distributed discontinuously throughout a tissue section, the horizontal length of each discrete region was measured using NDP.view 2. For a target region x , the total horizontal length, $L(x)$, was calculated as the sum of the individual segment lengths:

$$L(x) = \sum_{i=1}^n L_i(x)$$

where $L_i(x)$ denotes the horizontal length of the i -th discrete segment of the target region.

For analyses normalized to the horizontal extent of the carcinoma, the relative extent of

region x stroma was calculated as: $\frac{L(x)}{L(Carcinoma)}$

Immunoreactivity for CXCL14 and MMP3 was assessed qualitatively because positive fibroblasts could not be reliably distinguished from round inflammatory cells for area- or length-based quantification. In desmoplastic stroma, CXCL14 expression was evaluated specifically in spindle-shaped stromal cells. MMP3 expression in CAFs was assessed by comparing its distribution in spindle-shaped pericryptal stromal cells with the distributions of FAP and POSTN expression and with *WNT5A* RNA signals detected by RNAscope on serial sections.

Processing and integration of scRNA-seq data

Preprocessing and analysis of scRNA-seq data were performed using R v.4.3.3 and Seurat v.5.1.0. Three publicly available colorectal cancer scRNA-seq datasets, GSE132465, GSE144735, and GSE178341, were downloaded from the Gene Expression Omnibus (GEO).

GSE132465 and GSE144735 had been preprocessed and annotated by Lee and colleagues.¹⁷ Cells annotated as “Myofibroblasts”, “Stromal 1”, “Stromal 2”, or “Stromal 3” in the Cell_subtype metadata field were extracted as fibroblasts. GSE178341, deposited by Pelka and colleagues,¹⁸ was obtained as a count matrix with the associated metadata. Cells with more than 200 and fewer than 10,000 detected genes and a

mitochondrial transcript proportion of 15% or less were retained for subsequent analyses.

After preprocess, unsupervised clustering was performed using a resolution of 0.5.

Stromal cells with high *DCN* and *COL1A1* expression were retained, whereas endothelial cells with high *CD34* and *CLDN5* expression and pericytes with high *MCAM*, *PDGFRB*, *RGS5*, and *CSPG4* expression were excluded.

Among the retained fibroblasts, clusters predominantly derived from tumor tissues were classified as cancer-associated fibroblasts (CAFs). Fibroblast and CAF datasets were separately integrated using anchor-based integration approach, allowing the identification of mutual nearest neighbors and effective correction of batch-specific variation.

One fibroblast cluster contained cells from both normal and tumor tissues in GSE178341 but consisted almost entirely of normal-tissue-derived cells in GSE132465 and GSE144735. This cluster was designated intermediate fibroblasts (intermediate_fibro). The final integrated dataset comprised 3391, 3269, and 6036 fibroblasts from GSE132465, GSE144735, and GSE178341, respectively. Of these, 1434, 864, and 2298 cells, respectively, were classified as CAFs or intermediate_fibro.

Differentially expressed gene analysis was performed using the RNA assay. Cluster marker genes were identified using FindAllMarkers with the Wilcoxon rank-sum test, with only.pos = TRUE, min.pct = 0.25, and logfc.threshold = 0.25. To exclude genes with

minimal expression, raw counts were averaged across all cells within each cluster, and only genes with a mean raw count greater than one in the corresponding cluster were retained.

Functional annotation of CAFs using AUCell

CAF clusters were functionally annotated using AUCell v.1.24.0. AUCell uses ranked gene-expression profiles to calculate a gene set enrichment score for individual cell.¹⁹

CAF subtype-specific gene sets were derived from the differentially expressed gene lists reported by Cords and colleagues.³⁷ AUCell enrichment scores for each CAF signature, including matrix CAFs (mCAFs), inflammatory CAFs (iCAFs), and other reported CAF subtypes, were calculated for individual cells and summarized for each cluster. Mean AUCell enrichment scores were visualized using rose plots, and clusters were provisionally annotated according to the CAF signatures with the highest scores.

Gene set enrichment analysis of scRNA-seq data

For each CAF cluster, both upregulated and downregulated genes were identified against all other CAF clusters using FindAllMarkers with the Wilcoxon rank-sum test, with `only.pos = FALSE`, `min.pct = 0.25`, and `logfc.threshold = 0.25`. Raw counts were averaged

across all cells within each cluster, and genes with a mean raw count of 1 or less in the corresponding cluster were excluded. The remaining genes were ranked in descending order according to their average log2 fold change.

Gene set enrichment analysis (GSEA)⁵⁵ was performed using the GSEA function in clusterProfiler v.4.10.1, with minimum and maximum gene-set sizes of 10 and 500, respectively. Gene sets with a false discovery rate-adjusted p-value below 0.05 were considered significantly enriched. Gene sets were obtained from the Molecular Signatures Database (MSigDB) v2026.1.Hs^{56,57}. Hallmark gene sets were used to assess broad biological programs, whereas Gene Ontology Biological Process gene sets were used to examine more specific biological processes.

Tissue segmentation and preprocessing of Visium spatial transcriptomic datasets

Six publicly available Visium ST datasets of eCRC were obtained from the DDBJ Sequence Read Archive (DRA, accession number DRA016520 and DRA016537) and Genomic Expression Archive (GEA, accession number E-GEAD-619 and E-GEAD-622). The datasets were loaded and processed using Python v.3.10.16 and Scanpy v.1.11.0. High-resolution tissue images were manually annotated using QuPath v.0.6.0 by a gastrointestinal pathologist (ST). Tissue regions were segmented into normal mucosa,

submucosa and muscularis propria, adenoma, CIS, or SM inv. The correspondence between individual Visium spots and the segmented tissue regions was exported as a comma-separated values file and added to the metadata of each Visium dataset. The six datasets were subsequently merged for integrative analysis.

Integration of the scRNA-seq and the Visium data using DeepCOLOR package

The GSE132465 scRNA-seq dataset analyzed using Seurat was converted to AnnData format using SeuratDisk v.0.0.0.9021. The converted scRNA-seq dataset was integrated with the six Visium ST datasets using DeepCOLOR v.0.0.6, a Python package based on a variational autoencoder that projects scRNA-seq and ST data into a shared latent space to infer spatial co-localization patterns.⁴⁶

The integration was performed according to the DeepCOLOR tutorial (https://github.com/kojikoji/deepcolor/blob/master/tutorial/deepcolor_tutorial.ipynb; accessed July 30, 2026) and the method described by Hashimoto and colleagues.⁵⁸

For each cell in the scRNA-seq dataset, DeepCOLOR calculated the proportion of its aggregated spatial assignments attributable to each histologically annotated tissue region. Cells were stratified using the six major cell-type categories provided in the Cell_Type metadata field of GSE132465: epithelial, T, B, stromal, mast, and myeloid cells. Within

each category, cells in the upper 10% of the spatial-assignment proportions for each tissue region were selected and designated “top10% cells”. Selection was performed independently for each tissue segment; therefore, the same cell could be selected for more than one segment.

Putative ligand–receptor-mediated cell–cell communication was inferred using `deepcolor.calculate_proximal_cell_communication`, and the results were visualized using river plots. Co-localization clusters containing epithelial and stromal cells were identified using `deepcolor.analyze_pair_cluster`. Clusters containing both carcinoma cells and fibroblasts were selected, and their top10% cells were used in downstream analyses.

Analysis of bulk RNA-sequencing data using TNMplot

Bulk RNA-sequencing data were analyzed using the TNMplot web platform, which enables comparisons of gene expression between normal and cancer tissues and across cancer stages.⁵⁹ Expression data and stage-wise statistical comparisons for selected genes in normal colorectal tissues and colorectal cancers of stages I–IV were obtained from TNMplot (<https://tnmplot.com/>; accessed July 30, 2026).

Processing and integration of scATAC-seq data

The GSE201336 single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) dataset was downloaded from the Gene Expression Omnibus. This dataset comprised non-tumor colorectal mucosa, colorectal adenomatous polyps, and colorectal carcinomas from eight patients with familial adenomatous polyposis, four patients with sporadic colorectal cancer, and two healthy donors.²⁷

Fragment files were imported into R and processed using ArchR v.1.0.3, and an Arrow file was generated for each sample. Quality control, preprocessing, dimensionality reduction, and initial clustering were performed as described by Becker and colleagues,²⁷ with minor modifications.

The scATAC-seq fibroblast dataset was integrated with the integrated scRNA-seq fibroblast dataset generated from GSE132465, GSE144735, and GSE178341 using the ArchR `addGeneIntegrationMatrix` function. The resulting `GeneIntegrationMatrix` was used to transfer fibroblast annotations and impute gene expression in the scATAC-seq dataset. Peak calling, marker-peak identification, chromVAR motif deviation analysis, and peak-to-gene linkage analysis were performed according to the ArchR manual (<https://www.archrproject.com/bookdown/index.html>; accessed July 30, 2026).⁶⁰

Transcription factor binding prediction

Genomic sequences encompassing selected linked peaks were retrieved from the GRCh38 reference genome and screened for motifs corresponding to candidate transcription factors prioritized by the integrated motif-accessibility and gene-expression analysis. Motif matching was performed using the JASPAR2026 CORE Human collection with the default relative profile-score threshold of 80%.⁶¹ Sequence logos for matched motifs were obtained from the JASPAR website.

Ethics statement

The clinicopathological evaluation and experiments using eCRC tissue samples were conducted in accordance with the Declaration of Helsinki and were approved by the Institutional Review Board of Kobe University (approval number B230125). Given the retrospective nature of the study, the requirement for written informed consent was waived by the Institutional Review Board, and participants were provided with an opportunity to opt out of the study.

Statistics

Statistical analyses for clinicopathological and histological evaluations were performed using R v.4.3.3. Correlation between two factors were tested using the Pearson correlation

coefficient. Between-group differences were assessed using the two-sided Wilcoxon rank-sum test for unpaired data and using the two-sided Wilcoxon signed-rank test for paired data. The associations between eCRC macroscopic classification and SM inv frequency were tested using Logistic regression analysis or using the two-sided Pearson χ^2 test. A two-sided p-value < 0.05 was defined significant.

Data Availability

The processed data in the present study are available from S.T. or the corresponding author H.M. upon request. Visium ST data were downloadable from the DDBJ Sequence Read Archive (DRA) (<https://www.ddbj.nig.ac.jp/dra/index-e.html>) and the count data are stored in the Genomic Expression Archive (GEA) (<https://www.ddbj.nig.ac.jp/gea/index-e.html>) with the associated accession number DRA016520 and DRA016537 for raw data and the associated accession number E-GEAD-619 and E-GEAD-622 for count data. scRNA-seq datasets (GSE132465, GSE144735, and GSE178341) and scATAC-seq dataset (GSE201336) are available from Gene Expression Omnibus.

Code availability

All statistical and bioinformatic analysis were conducted using conventional codes. No new codes for analyses were created.

References

- 1 Sung, H. *et al.* Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA: a cancer journal for clinicians* **76**, e70090, doi:10.3322/caac.70090 (2026).
- 2 Augsten, M. Cancer-associated fibroblasts as another polarized cell type of the tumor microenvironment. *Frontiers in oncology* **4**, 62, doi:10.3389/fonc.2014.00062 (2014).
- 3 Ishii, G. New insights into cancer pathology learned from the dynamics of cancer-associated fibroblasts. *Pathology international* **74**, 493-507, doi:10.1111/pin.13461 (2024).
- 4 Kobayashi, H. *et al.* Cancer-associated fibroblasts in gastrointestinal cancer. *Nature reviews. Gastroenterology & hepatology* **16**, 282-295, doi:10.1038/s41575-019-0115-0 (2019).
- 5 Szabo, P. M. *et al.* Cancer-associated fibroblasts are the main contributors to epithelial-to-mesenchymal signatures in the tumor microenvironment. *Scientific reports* **13**, 3051, doi:10.1038/s41598-023-28480-9 (2023).
- 6 Koncina, E. *et al.* IL1R1(+) cancer-associated fibroblasts drive tumor development and immunosuppression in colorectal cancer. *Nature communications* **14**, 4251, doi:10.1038/s41467-023-39953-w (2023).
- 7 Nicolas, A. M. *et al.* Inflammatory fibroblasts mediate resistance to neoadjuvant therapy in rectal cancer. *Cancer cell* **40**, 168-184.e113, doi:10.1016/j.ccell.2022.01.004 (2022).
- 8 Bissell, M. J. & Hines, W. C. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nature medicine* **17**, 320-329, doi:10.1038/nm.2328 (2011).
- 9 Sahai, E. *et al.* A framework for advancing our understanding of cancer-associated fibroblasts. *Nature reviews. Cancer* **20**, 174-186, doi:10.1038/s41568-019-0238-1 (2020).
- 10 Iida, T. *et al.* Pharmacologic conversion of cancer-associated fibroblasts from a protumor phenotype to an antitumor phenotype improves the sensitivity of pancreatic cancer to

- chemotherapeutics. *Oncogene* **41**, 2764-2777, doi:10.1038/s41388-022-02288-9 (2022).
- 11 Nakayama, H., Enzan, H. & Yasui, W. Expression of podoplanin/D2-40 in pericryptal stromal cells in superficial colorectal epithelial neoplasia. *Medical molecular morphology* **46**, 20-23, doi:10.1007/s00795-012-0004-x (2013).
 - 12 Shigeoka, M. *et al.* Intraepithelial CD163(+) macrophages in tongue leukoplakia biopsy: A promising tool for cancer screening. *Oral diseases* **26**, 527-536, doi:10.1111/odi.13269 (2020).
 - 13 Urakami, S. *et al.* Biological and clinical significance of the YKL-40/osteopontin-integrin β 4-p70S6K axis induced by macrophages in early oesophageal squamous cell carcinoma. *The Journal of pathology* **261**, 55-70, doi:10.1002/path.6148 (2023).
 - 14 Tsukamoto, S. *et al.* Podoplanin Expression in Early-Stage Colorectal Cancer-Associated Fibroblasts and Its Utility as a Diagnostic Marker for Colorectal Lesions. *Cells* **13**, doi:10.3390/cells13201682 (2024).
 - 15 Shimoda, T. *et al.* Early colorectal carcinoma with special reference to its development de novo. *Cancer* **64**, 1138-1146, doi:10.1002/1097-0142(19890901)64:5<1138::aid-cncr2820640529>3.0.co;2-a (1989).
 - 16 Medawar, E. *et al.* Clinical management of patients with colorectal intramucosal carcinoma compared to high-grade dysplasia and T1 colorectal cancer. *Gastrointestinal endoscopy* **101**, 1211-1221.e1215, doi:10.1016/j.gie.2024.11.021 (2025).
 - 17 Lee, H. O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nature genetics* **52**, 594-603, doi:10.1038/s41588-020-0636-z (2020).
 - 18 Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734-4752.e4720, doi:10.1016/j.cell.2021.08.003 (2021).
 - 19 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature methods* **14**, 1083-1086, doi:10.1038/nmeth.4463 (2017).
 - 20 Harada, A. *et al.* Hypoxia-induced Wnt5a-secreting fibroblasts promote colon cancer progression. *Nature communications* **16**, 3653, doi:10.1038/s41467-025-58748-9 (2025).
 - 21 Ku, J. *et al.* Identification of a unique subpopulation of mucosal fibroblasts in colorectal cancer with tumor-restraining characteristics. *Molecules and cells* **48**, 100263, doi:10.1016/j.mocell.2025.100263 (2025).
 - 22 Lee, K. W. *et al.* PRRX1 is a master transcription factor of stromal fibroblasts for myofibroblastic lineage progression. *Nature communications* **13**, 2793, doi:10.1038/s41467-022-30484-4 (2022).
 - 23 Ilie, S. M. *et al.* Cancer-associated fibroblasts and their prognostic role in colorectal cancer: review and meta-analysis. *Frontiers in oncology* **15**, 1635055,

- doi:10.3389/fonc.2025.1635055 (2025).
- 24 Qi, L. *et al.* Targeting LRRC15 in Cancer-Associated Fibroblasts Modifies the Extracellular Matrix and Enhances Tumor Immune Responses to Suppress Lung Cancer Progression. *Cancer research* **86**, 2377-2394, doi:10.1158/0008-5472.can-25-2871 (2026).
 - 25 Elyada, E. *et al.* Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer discovery* **9**, 1102-1123, doi:10.1158/2159-8290.cd-19-0094 (2019).
 - 26 Yang, D., Liu, J., Qian, H. & Zhuang, Q. Cancer-associated fibroblasts: from basic science to anticancer therapy. *Experimental & molecular medicine* **55**, 1322-1332, doi:10.1038/s12276-023-01013-0 (2023).
 - 27 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. *Nature genetics* **54**, 985-995, doi:10.1038/s41588-022-01088-x (2022).
 - 28 Yashiro, M. *et al.* Genetic pathways in the evolution of morphologically distinct colorectal neoplasms. *Cancer research* **61**, 2676-2683 (2001).
 - 29 Hasegawa, H. *et al.* K-ras gene mutations in early colorectal cancer ... flat elevated vs polyp-forming cancer. *Oncogene* **10**, 1413-1416 (1995).
 - 30 Yao, T. & Tsuneyoshi, M. Significance of pericryptal fibroblasts in colorectal epithelial tumors: a special reference to the histologic features and growth patterns. *Human pathology* **24**, 525-533, doi:10.1016/0046-8177(93)90165-d (1993).
 - 31 Tanigawa, K. *et al.* S100A8/A9 Induced by Interaction with Macrophages in Esophageal Squamous Cell Carcinoma Promotes the Migration and Invasion of Cancer Cells via Akt and p38 MAPK Pathways. *The American journal of pathology* **192**, 536-552, doi:10.1016/j.ajpath.2021.12.002 (2022).
 - 32 Tsukamoto, S. *et al.* Matrix Metalloproteinase 9 Induced in Esophageal Squamous Cell Carcinoma Cells via Close Contact with Tumor-Associated Macrophages Contributes to Cancer Progression and Poor Prognosis. *Cancers* **15**, doi:10.3390/cancers15112987 (2023).
 - 33 Miyako, S. *et al.* Periostin in Cancer-Associated Fibroblasts Promotes Esophageal Squamous Cell Carcinoma Progression by Enhancing Cancer and Stromal Cell Migration. *The American journal of pathology* **194**, 828-848, doi:10.1016/j.ajpath.2023.12.010 (2024).
 - 34 Neri, S. *et al.* Podoplanin-expressing cancer-associated fibroblasts lead and enhance the local invasion of cancer cells in lung adenocarcinoma. *International journal of cancer* **137**, 784-796, doi:10.1002/ijc.29464 (2015).

- 35 Pramod, S., Lavon, H., Scherz-Shouval, R. & Sherman, M. H. Heterogeneity and plasticity of cancer-associated fibroblasts. *Nature cancer* **7**, 583-596, doi:10.1038/s43018-026-01146-x (2026).
- 36 Han, C., Liu, T. & Yin, R. Biomarkers for cancer-associated fibroblasts. *Biomarker research* **8**, 64, doi:10.1186/s40364-020-00245-w (2020).
- 37 Cords, L. *et al.* Cancer-associated fibroblast classification in single-cell and spatial proteomics data. *Nature communications* **14**, 4294, doi:10.1038/s41467-023-39762-1 (2023).
- 38 Parikh, A. *et al.* Malignant cell-specific CXCL14 promotes tumor lymphocyte infiltration in oral cavity squamous cell carcinoma. *Journal for immunotherapy of cancer* **8**, doi:10.1136/jitc-2020-001048 (2020).
- 39 Augsten, M. *et al.* Cancer-associated fibroblasts expressing CXCL14 rely upon NOS1-derived nitric oxide signaling for their tumor-supporting properties. *Cancer research* **74**, 2999-3010, doi:10.1158/0008-5472.can-13-2740 (2014).
- 40 Basisty, N. *et al.* A proteomic atlas of senescence-associated secretomes for aging biomarker development. *PLoS biology* **18**, e3000599, doi:10.1371/journal.pbio.3000599 (2020).
- 41 Zhou, J. *et al.* Aging-associated interleukin-11 drives the molecular mechanism and targeted therapy of idiopathic pulmonary fibrosis. *European journal of medical research* **30**, 542, doi:10.1186/s40001-025-02755-5 (2025).
- 42 Liu, Z. *et al.* THBS2-producing matrix CAFs promote colorectal cancer progression and link to poor prognosis via the CD47-MAPK axis. *Cell reports* **44**, 115555, doi:10.1016/j.celrep.2025.115555 (2025).
- 43 Nishina, T. *et al.* Interleukin-11-expressing fibroblasts have a unique gene signature correlated with poor prognosis of colorectal cancer. *Nature communications* **12**, 2281, doi:10.1038/s41467-021-22450-3 (2021).
- 44 Okusha, Y. *et al.* The intranuclear PEX domain of MMP involves proliferation, migration, and metastasis of aggressive adenocarcinoma cells. *Journal of cellular biochemistry* **119**, 7363-7376, doi:10.1002/jcb.27040 (2018).
- 45 Hosono, M. *et al.* CXCL8 derived from tumor-associated macrophages and esophageal squamous cell carcinomas contributes to tumor progression by promoting migration and invasion of cancer cells. *Oncotarget* **8**, 106071-106088, doi:10.18632/oncotarget.22526 (2017).
- 46 Kojima, Y. *et al.* Single-cell colocalization analysis using a deep generative model. *Cell systems* **15**, 180-192.e187, doi:10.1016/j.cels.2024.01.007 (2024).
- 47 Gudjonsson, T. *et al.* Normal and tumor-derived myoepithelial cells differ in their ability

- to interact with luminal breast epithelial cells for polarity and basement membrane deposition. *Journal of cell science* **115**, 39-50, doi:10.1242/jcs.115.1.39 (2002).
- 48 Kirkland, S. C. Type I collagen inhibits differentiation and promotes a stem cell-like phenotype in human colorectal carcinoma cells. *British journal of cancer* **101**, 320-326, doi:10.1038/sj.bjc.6605143 (2009).
- 49 Bochet, L. *et al.* Adipocyte-derived fibroblasts promote tumor progression and contribute to the desmoplastic reaction in breast cancer. *Cancer research* **73**, 5657-5668, doi:10.1158/0008-5472.can-13-0530 (2013).
- 50 Hosaka, K. *et al.* Pericyte-fibroblast transition promotes tumor growth and metastasis. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E5618-5627, doi:10.1073/pnas.1608384113 (2016).
- 51 Ishii, G. *et al.* Bone-marrow-derived myofibroblasts contribute to the cancer-induced stromal reaction. *Biochemical and biophysical research communications* **309**, 232-240, doi:10.1016/s0006-291x(03)01544-4 (2003).
- 52 Brierley, J. D., Gospodarowicz, M. K. & Wittekind, C. (Wiley-Blackwell, Chichester, 2016).
- 53 WHO Classification of Tumors Editorial Board. *Digestive System Tumors*. 5th edn, Vol. 1 (International Agency for Research on Cancer, 2019).
- 54 Japanese Society for Cancer of the Colon and Rectum. *Japanese Classification of Colorectal, Appendiceal, and Anal Carcinoma*. 9th edn, (Kanehara & Co., Ltd., 2018).
- 55 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
- 56 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).
- 57 Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)* **27**, 1739-1740, doi:10.1093/bioinformatics/btr260 (2011).
- 58 Hashimoto, M. *et al.* Spatial and single-cell colocalization analysis reveals MDK-mediated immunosuppressive environment with regulatory T cells in colorectal carcinogenesis. *EBioMedicine* **103**, 105102, doi:10.1016/j.ebiom.2024.105102 (2024).
- 59 Bartha, Á. & Györfy, B. TNMplot: An enhanced platform for pharmacological target identification through cross-stage and pan-cancer gene expression analysis. *British journal of pharmacology* **183**, 2648-2659, doi:10.1111/bph.70390 (2026).
- 60 Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics* **53**, 403-411, doi:10.1038/s41588-021-

00790-6 (2021).

- 61 Ovek Baydar, D. *et al.* JASPAR 2026: expansion of transcription factor binding profiles and integration of deep learning models. *Nucleic acids research* **54**, D184-d193, doi:10.1093/nar/gkaf1209 (2026).

Acknowledgements

We would like to thank Yumi Hashimoto and Nobuo Kubo for their technical support.

This study was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (JSPS) (24K23318 and 26K18603) and by the Japan Agency for Medical Research and Development (AMED) (JP256f0137011)

Author Contributions

S.T. conceived the study, prepared tissue samples, performed histological evaluation and bioinformatics analysis, acquired funding, and wrote the original draft. M.H. acquired funding and supervised the study. T.N. provided eCRC clinicopathological data. S.S. provided GPU-equipped computer for DeepCOLOR analysis. T.N., T.K., M.S., Y.-i.K, S.S, G.I. and M.H. helped data curation, discussed the results with S.T, and reviewed the original draft. All authors have read and agreed to the published version of the manuscript.

Role of Funders

The funders (Japan Society for the Promotion of Science and Japan Agency for Medical Research and Development) played no role in the study design, data acquisition, data analysis, and reporting.

Declaration of interests

All authors declare no competing interests.

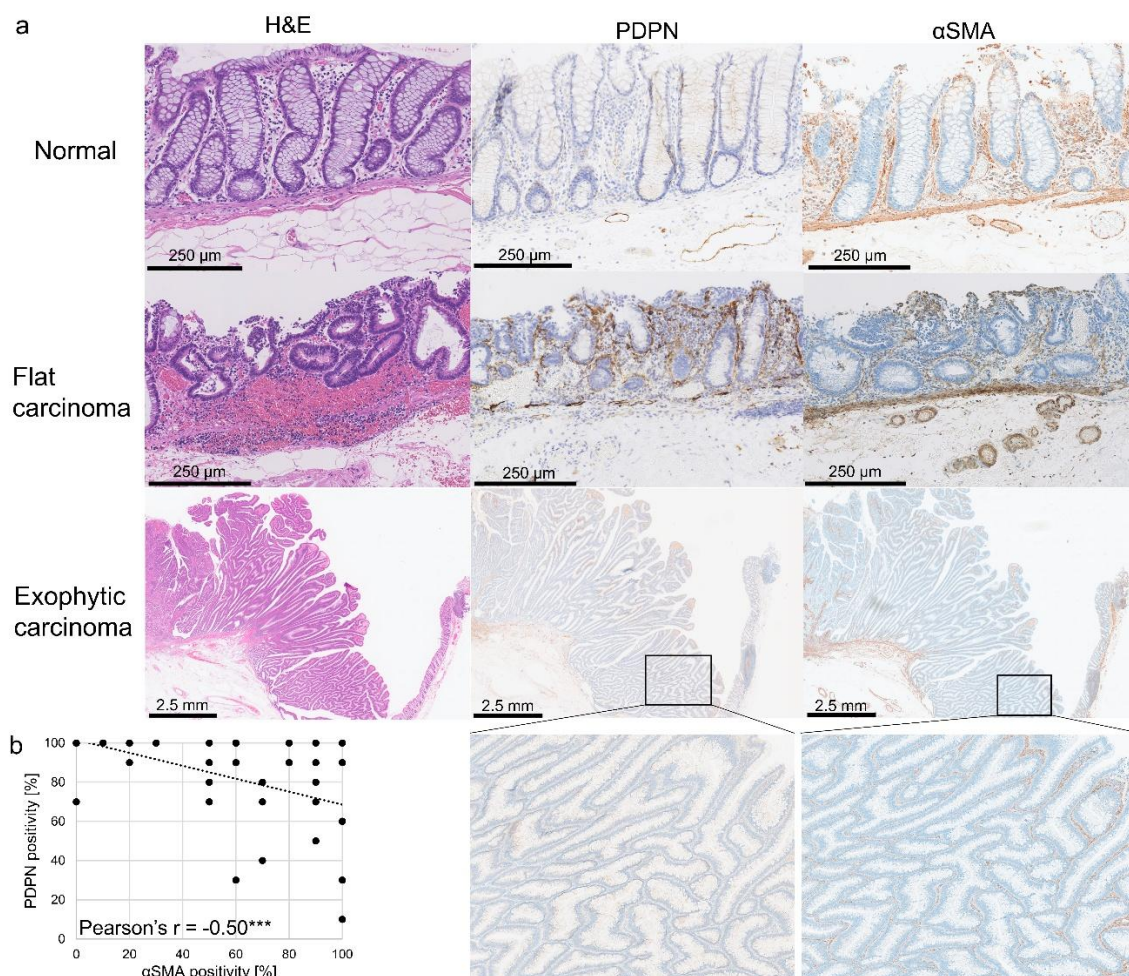


Figure 1. Histological features and stromal PDPN and α SMA expression in early colorectal carcinoma (eCRC).

a: Representative H&E and immunohistochemical images of normal colorectal mucosa and flat and exophytic eCRCs. Enlarged views of the boxed regions are shown below. b: Stromal PDPN expression was inversely correlated with α SMA positivity in 45 eCRCs (Pearson's $r = -0.50$). Scale bars, 250 μ m and 2.5 mm, as indicated. H&E, haematoxylin-eosin staining; PDPN, podoplanin; α SMA, alpha-smooth muscle actin; ***, $p < 0.001$.

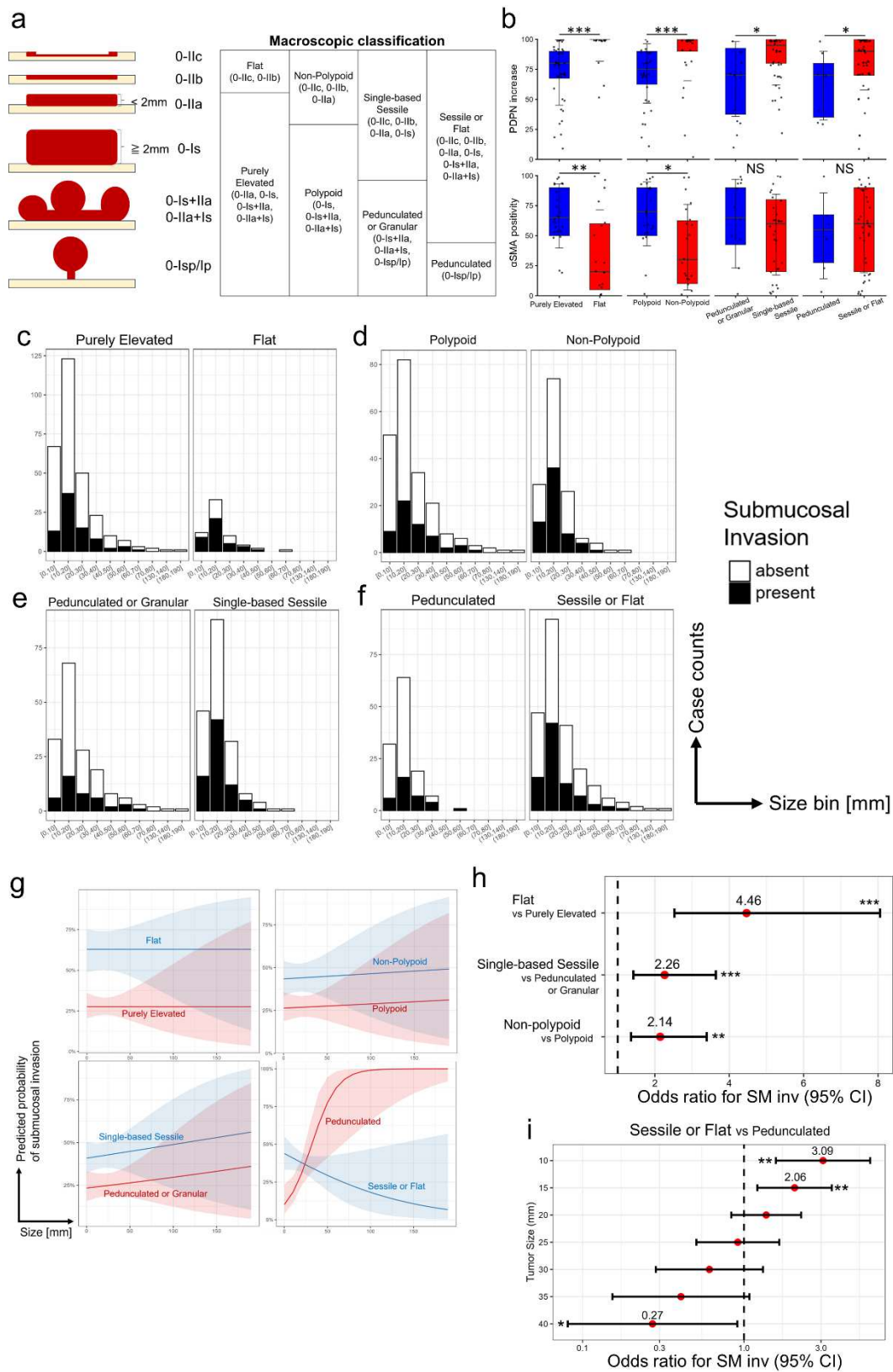


Figure 2. CAF marker expression and risk of submucosal invasion according to the macroscopic classifications of eCRCs.

a: Schematic representation of four macroscopic classifications schemes for eCRC. b: PDPN and α SMA expression according to eCRC macroscopic classification. Box plots show the median (center line) and interquartile range (box); individual tumors are shown as dots. Error bars indicate mean $\pm$ SD. c-f: Numbers of tumors with (black) or without (white) submucosal invasion, stratified by morphology and tumor size in 10-mm increments. g: Model-predicted probabilities of submucosal invasion derived from logistic regression; shaded areas indicate 95% confidence intervals (CIs). h: Tumor size-adjusted ORs for submucosal invasion in Flat versus Purely Elevated, Non-Polypoid versus Polypoid, and Single-based Sessile versus Pedunculated or Granular classifications. Error bars indicate 95% CIs. i: Size-specific odds ratios (ORs) for submucosal invasion in Sessile or Flat versus Pedunculated eCRCs. NS, not significant; *: $p < 0.05$; **: $p < 0.01$; and ***: $p < 0.001$.

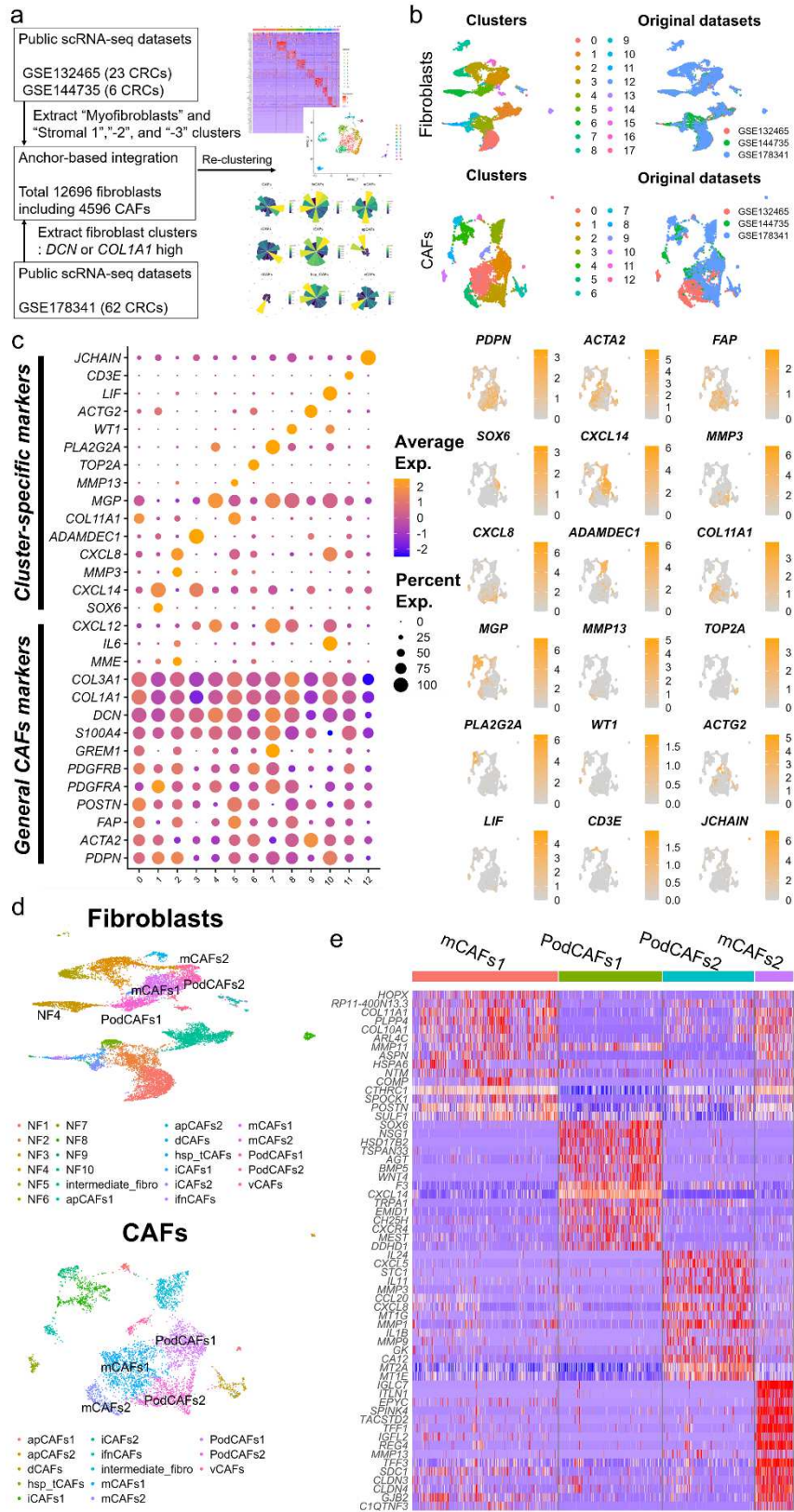


Figure 3. Integrative scRNA-seq analysis of colorectal cancer fibroblasts and cancer-associated fibroblasts.

a: Workflow of the integrative analysis of the GSE132465, GSE144735, and GSE178341 colorectal cancer (CRC) scRNA-seq datasets. b: UMAP plots of all fibroblasts (upper) and CAFs (lower), colored by cluster assignment (left) and original dataset (right). c: Dot plot of general CAF markers and cluster-specific marker genes (left), and UMAP plots showing the expression of *PDPN*, *ACTA2*, and representative differentially expressed genes for each CAF cluster (right). d: UMAP plots of fibroblasts (upper) and CAFs (lower), colored according to annotations based on *PDPN* expression and AUCell classification. e: Heatmap showing scaled expression of the top 10 significantly upregulated genes in mCAFs1, PodCAFs1, PodCAFs2, and mCAFs2 clusters.

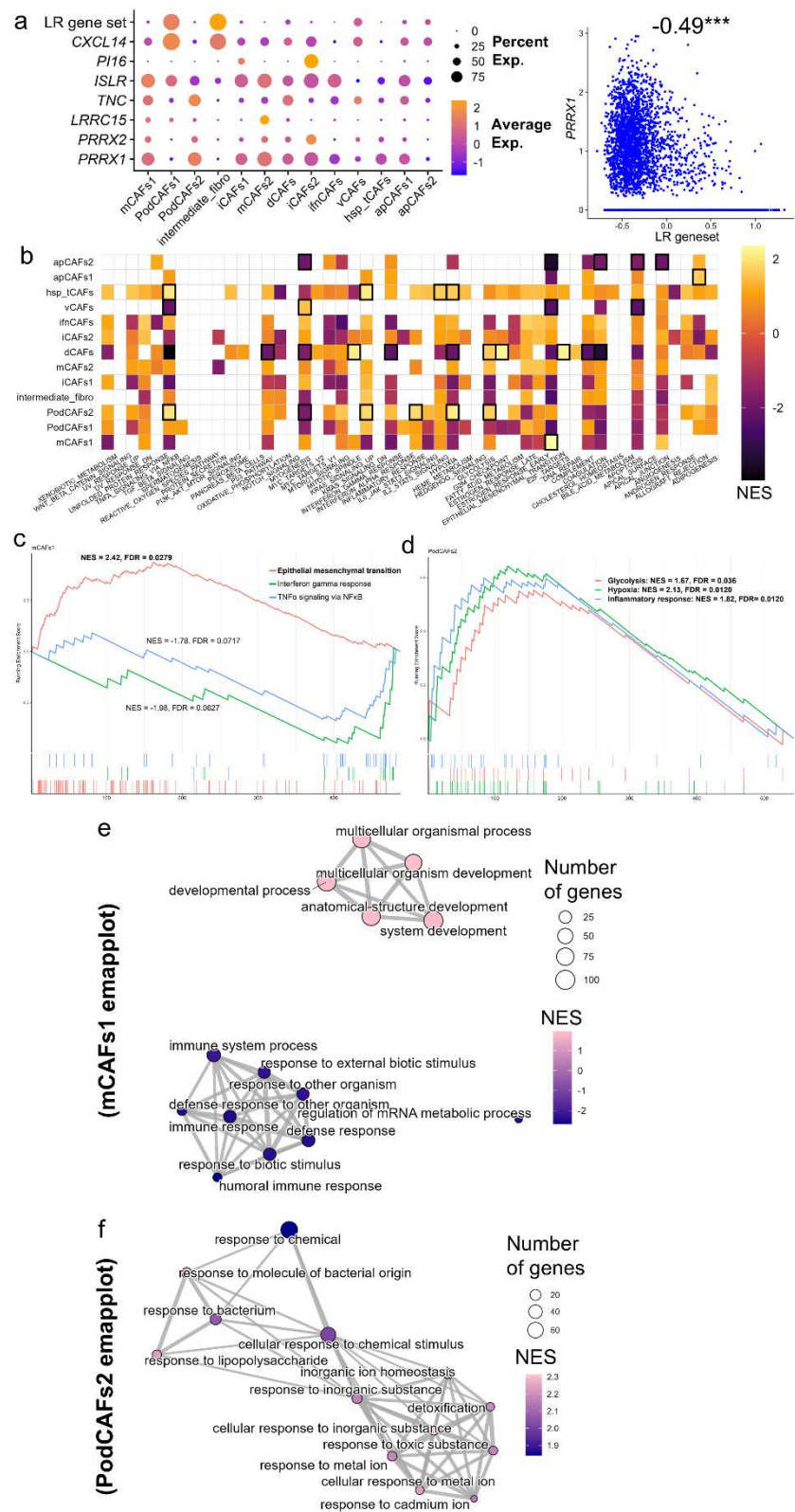


Figure 4. Gene-expression and functional enrichment profiles of mCAFs1, PodCAFs1, and PodCAFs2.

a: Dot plot showing the expression of genes associated with favorable or poor prognosis across CAF clusters (left). The low-risk-associated geneset (LR geneset) represents a module score derived from genes associated with favorable prognosis, as proposed by Ku and colleagues. The feature scatter plot shows an inverse correlation between the LR geneset score and *PRRX1* expression across individual cells (right). The number above the plot indicates the Pearson correlation coefficient. b: Heatmap showing normalised enrichment scores from geneset enrichment analysis (GSEA) of the Hallmark gene sets across CAF clusters. Black outlines indicate significantly enriched gene sets with a false discovery rate (FDR) of less than 0.05. c, d: GSEA enrichment plots of representative Hallmark genesets in mCAFs1 (c) and PodCAFs2 (d). e, f: Enrichment maps of Gene Ontology (GO) biological process terms identified by GSEA for mCAFs1 (e) and PodCAFs2 (f). Each node represents an enriched term, node size indicates the number of genes, and edges connect terms with overlapping genes. Node color indicates the normalised enrichment score. NES, normalised enrichment score. ***: $p < 0.001$

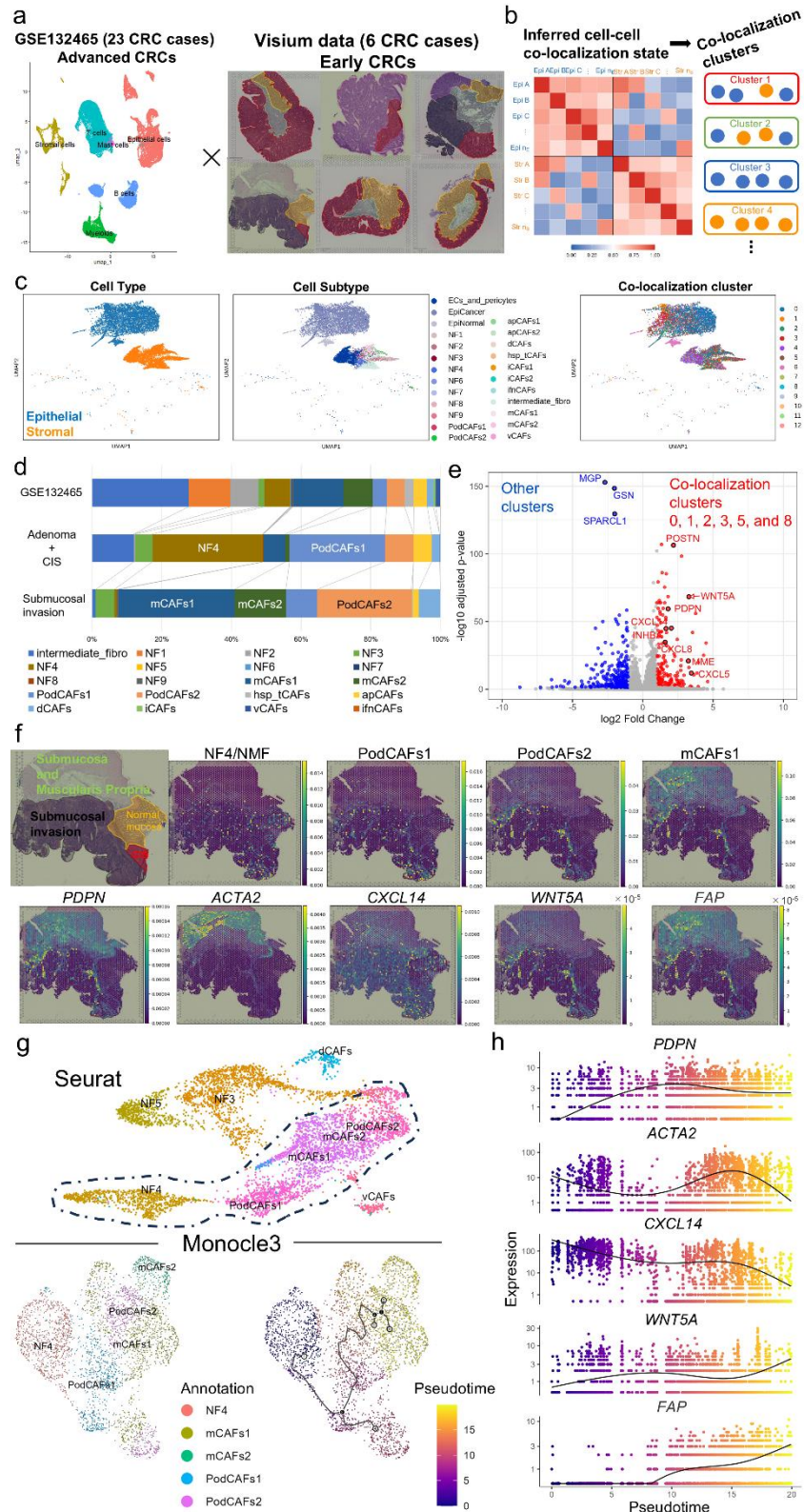


Figure 5. Inferred spatial distribution of CAF subtypes in eCRC.

a: Workflow for the integrated analysis of the GSE132465 scRNA-seq dataset and Visium spatial

transcriptomic data from six eCRCs. b: Schematic diagram of epithelial and stromal cell co-localization analysis using DeepCOLOR package. Cell are clustered based on similarity of their inferred co-localization profiles. Clusters represent shared co-localization states. c: DeepCOLOR-based co-localization analysis of epithelial and stromal cells. UMAP plots show major cell types (left), annotated cell subtypes (middle), and co-localization clusters (right). d: Stacked bar plots showing the distributions of fibroblast clusters in the original GSE132465 dataset (upper), among cells in the DeepCOLOR-defined “top10% cells” within adenoma and CIS regions (middle), and among those within submucosally invasive carcinoma regions (lower). e: Volcano plot showing differential gene expression in stromal cells predicted to be spatially adjacent to tumor cells. Red and blue indicate upregulated and downregulated genes, respectively, and grey indicates genes that did not meet the predefined significance criteria. f: Spatial feature plots showing enrichment scores for NF4/NMF, PodCAFs1, PodCAFs2, and mCAFs1 (upper row), and the expression of the indicated marker genes (middle and lower rows), in a representative Visium tissue section. g, h: Monocle 3 pseudotime analysis of scRNA-seq data. (g) UMAP plots showing the selected NF4/NMF, PodCAFs1, PodCAFs2, mCAFs1, and mCAFs2 clusters and the inferred trajectory, with NF4 specified as the root state. (h) Expression of the indicated marker genes across pseudotime. Each point represents an individual cell, colors indicate pseudotime, and black lines show fitted expression trends.

a: UMAP plots of fibroblasts from the scATAC-seq dataset, colored by sample condition (left), fibroblast annotation transferred from the integrated scRNA-seq data (middle), and inferred pseudotime from NF4/NMF to mCAFs1 (right, cropped). Sample conditions comprised normal mucosa from individuals without CRC, histologically unaffected mucosa from patients with familial adenomatous polyposis, adenomatous polyp, and invasive carcinoma. b: UMAP feature plots showing imputed expression of *PDPN*, *ACTA2*, *CXCL14*, *MMP3*, *CXCL8*, *WNT5A*, *POSTN*, *FAP*, and *COL11A1* in the scATAC-seq dataset. c: Heatmap showing chromVAR motif deviation scores across the inferred pseudotime trajectory from NF4/NMF to mCAFs1. Selected transcription factor families are indicated on the right. d: Correlation between transcription factor motif accessibility and imputed gene expression (x-axis) plotted against the maximum change in motif accessibility across the inferred pseudotime trajectory (y-axis). Transcription factors with a maximum accessibility change in the upper quartile and a correlation coefficient greater than 0.5 are highlighted in red. e: Coverage plots showing chromatin accessibility and peak-to-gene links at the *FAP* locus across NF4/NMF, PodCAFs1, PodCAFs2, and mCAFs1. Orange shading highlights genomic regions with marked accessibility differences between NF4/NMF and PodCAFs1 or between PodCAFs1 and mCAFs1. Selected transcription factor binding to regions chr2:162,244,448–162,244,948 and chr2:162244974–162245474) was predicted using the JASPAR database. f: Transcription factor footprinting profiles for FOXF1, FOXF2, MSC, FOSL1, FOSL2, BACH1, JUND, CEBPA, RUNX1, and RUNX2 across NF4, PodCAFs1, PodCAFs2, and mCAFs1. Lines show the mean bias-corrected Tn5 insertion signal, and shaded areas indicate variability around the mean.

a: Low-magnification images of haematoxylin and eosin staining and immunohistochemistry for PDPN, α SMA, POSTN, and FAP in serial sections from the same eCRC specimen. The annotated regions indicate normal mucosa, carcinoma in situ (CIS), and submucosal invasion (SM inv). Differences in the boundaries of these regions among the serial sections reflect differences in tissue sectioning planes. Scale bars, 1 mm.

b: Heatmap summarising PDPN, α SMA, POSTN, and FAP immunoreactivity in normal mucosa (N), CIS stroma (C), and SM inv stroma (S) for individual eCRCs, grouped according to macroscopic classification. Red, high expression; Blue, low expression; Grey, not applicable (NA). Ped, Pedunculated; Ped or Gran, Pedunculated or Granular.

c: High-magnification images of FAP immunohistochemistry in normal mucosa, CIS, shallow SM inv, and deep SM inv regions. In the CIS image, FAP-positive pericryptal stroma is present on the left, whereas the stroma on the right shows little or no FAP expression. Scale bars, 250 μ m.

d: Length-based method for quantifying the target regions. Because each target region could be distributed discontinuously throughout a tissue section, the horizontal lengths of the individual segments, $L_i(x)$, were measured and summed to obtain the total horizontal length of target region x : $L(x) = \sum_i L_i(x)$.

e: Scatter plot showing the relationship between the horizontal extent of FAP-positive stroma relative to the total carcinoma length, $L(\text{FAP})/L(\text{carcinoma})$, and the horizontal extent of SM inv relative to the total carcinoma length, $L(\text{SM inv})/L(\text{carcinoma})$. The dotted line represents the fitted linear regression line. Pearson correlation coefficient, $r = 0.7880$.

f: Paired comparison of the proportions of the total carcinoma length occupied by FAP-positive stroma and SM inv. Lines connect paired measurements from the same tumor. Statistical significance was assessed using a two-sided Wilcoxon signed-rank test.

g: Box plots showing the proportion of the total carcinoma length associated with FAP-positive CIS stroma, $L(\text{FAPposCIS})/L(\text{carcinoma})$, according to the four macroscopic classification schemes. Boxes indicate the interquartile range, center lines indicate the median, individual tumors are shown as dots, and error bars indicate mean $\pm$ SD. Between-group differences were assessed using two-sided Wilcoxon rank-sum tests. NS, not significant; * $p < 0.05$; *** $p < 0.001$.

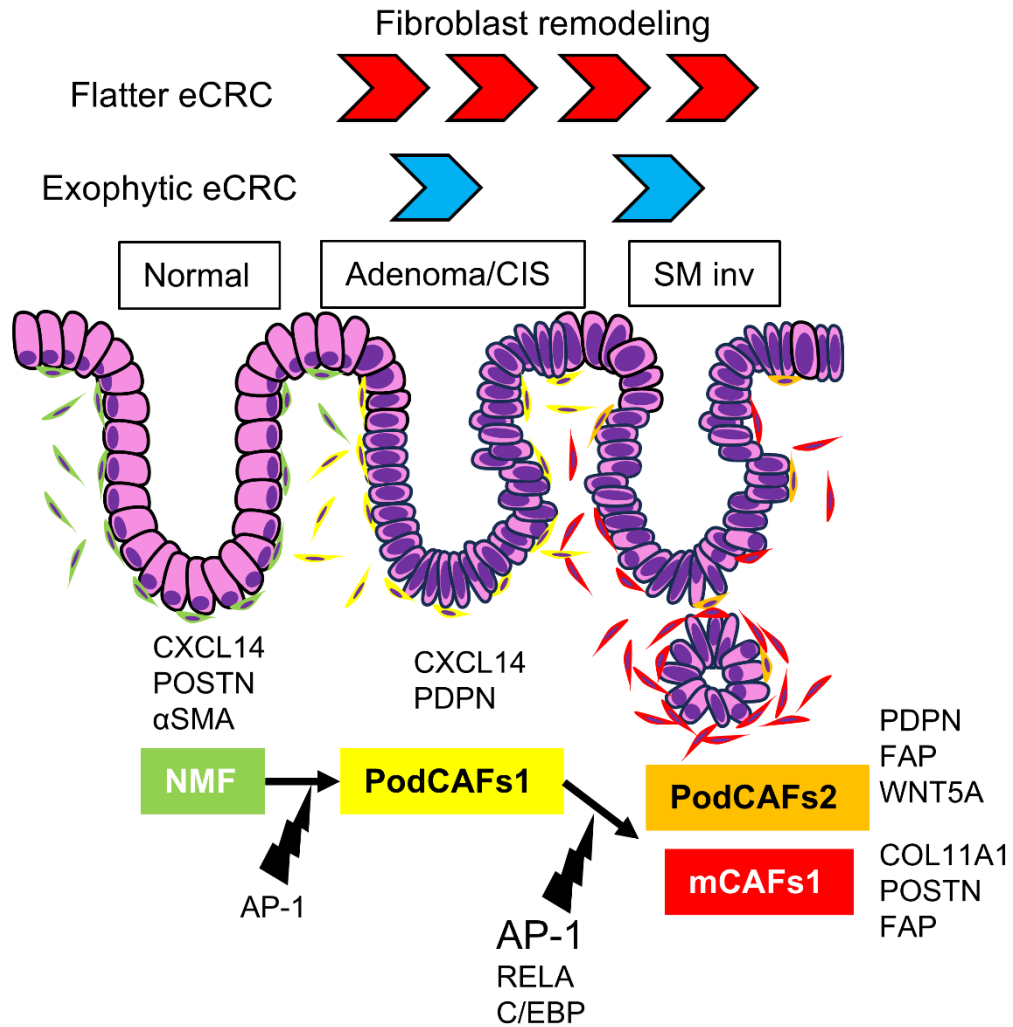


Figure 8. Schematic summary of this study.

Schematic model showing the inferred remodeling of fibroblast states during the progression from normal mucosa through adenoma or carcinoma in situ (CIS) to submucosal invasion (SM inv). Fibroblast-state remodeling in flatter eCRCs may progress further at smaller tumor sizes than in exophytic eCRCs, potentially contributing to their greater propensity for SM inv. AP-1-associated regulatory programs emerge during the transition from NMFs to PodCAFs1 and become more prominent in PodCAFs2 and mCAFs1, together with programs associated with other candidate transcription factors, including RELA and C/EBP family members. Selected marker genes associated with each fibroblast state are shown.

Table 1: Clinicopathological characteristics of 349 early colorectal cancers

Characteristic	Category	Number (%)
Age	< 65	108 (30.9%)
	≥ 65	241 (69.1%)
Sex	Male	234 (67.0%)
	Female	115 (33.0%)
Tumor location	Cecum	23 (6.6%)
	Ascendinding colon	47 (13.5%)
	Hepatic flexure	2 (0.6%)
	Transverse colon	58 (16.6%)
	Splenic flexure	6 (1.7%)
	Descending colon	19 (5.4%)
	Sigmoid colon	122 (35.0%)
	Rectosigmoid colon	17 (4.9%)
	Upper rectum	17 (4.9%)
	Lower rectum	31 (8.9%)
	Rectum, not further specified	4 (1.1%)
	Anal canal	1 (0.3%)
	Colorectum, not further specified	2 (0.6%)
Submucosal invasion	Absent (Carcinoma in situ)	231 (66.2%)
	Present	118 (33.8%)
Pedunculated vs Sessile or Flat	Pedunculated	123 (35.2%)
	Sessile or Flat	226 (64.8%)
Purely Elevated vs Flat	Purely Elevated	287 (82.2%)
	Flat	62 (17.8%)
Pedunculated or Granular vs Single-based Sessile	Pedunculated or Granular	169 (48.4%)
	Single-based Sessile	180 (51.6%)
Polypoid vs Non-Polypoid	Polypoid	208 (59.6%)
	Non-Polypoid	141 (40.4%)
Resection methods	Traditional polypectomy	56 (16.0%)
	Cold snare polypectomy	6 (1.7%)
	Endoscopic mucosal resection	89 (25.5%)
	Endoscopic submucosal dissection	110 (31.5%)
	Surgery	88 (25.2%)

Data are n (%). Lesions in the transverse colon or of unspecified location were classified separately because they could not be categorised as left-sided or right-sided. The unit of analysis was the tumor; multiple lesions from the same participant were treated as separate eCRCs.

Table 2. Associations between macroscopic classification and submucosal invasion in the present and previously reported cohorts

Study	Macroscopic classification	No SM inv (CIS)	SM inv	p-value
Shimoda et al. (1989) ¹⁵	Polypoid growth (PG)	114	32	< 0.0001
	Non-polypoid growth (NPG)	10	22	
Medawar et al. (2025) ¹⁶	Pedunculated	163	24	< 0.0001
	Sessile or Flat	166	102	
Present study	Pedunculated	89	34	0.09
	Sessile or Flat	142	84	
	Purely Elevated	208	79	< 0.0001
	Flat	23	39	
	Pedunculated or Granular	127	42	< 0.001
	Single-based Sessile	104	76	
	Polypoid	152	56	< 0.01
	Non-polypoid	79	62	

Data are numbers of early-stage colorectal carcinomas. For the previously reported cohorts, p-values are reproduced from the original publications. For the present study, p-values were calculated using two-sided Pearson χ^2 tests. The polypoid growth versus non-polypoid growth classification reported by Shimoda and colleagues differs from the Polypoid versus Non-Polypoid classification used in the present study; see reference 15 and 16 for the original definitions. CIS, carcinoma in situ; SM inv, submucosal invasion.

Supplementary Files

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

- [Supplemtarytables.xlsx](#)
- [SupplementaryInformation.pdf](#)