nature medicine

Article

https://doi.org/10.1038/s41591-026-04721-y

Insights into early-onset colorectal cancer from a single-nucleus RNA-seq atlas of colorectal cancer across the age continuum

Received: 15 December 2025

Accepted: 23 September 2026

Published online: 29 October 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

The incidence of early-onset colorectal cancer (EO-CRC; diagnosed before age 50 years) is increasing worldwide. Although current data indicate that EO-CRC may be biologically distinct from average-onset CRC (AO-CRC; diagnosed at age ≥50 years), molecular differences remain poorly defined. To address this knowledge gap, we generated a single-nucleus RNA sequencing (snRNA-seq) atlas from 189 primary CRC tumors (54 EO-CRC, 135 AO-CRC) spanning eight decades of life, with at least 20 tumors in most age decades. Our analyses stratified tumors by microsatellite status and accounted for tumor sidedness, stage, patient sex and neoadjuvant treatment history, modeling age continuously or categorically (EO-CRC versus AO-CRC). Among microsatellite-stable (MSS) tumors, those from younger patients were characterized by increased stromal and decreased epithelial proportions, which we confirmed using multispectral immunofluorescence and machine learning in an external cohort of 712 MSS CRC tumors (18 EO-CRC, 694 AO-CRC). Tumor epithelial cells from young patients exhibited increased expression of long noncoding RNAs and *FN1*, with a trend toward increased heterogeneity in copy number alterations. MSS left-sided EO-CRC tumors had more cell–cell interactions, with epithelial-to-mesenchymal transition-related signaling pathways (FN1/GDF/ACTIVIN) showing increased, cell type-specific interactions. Our large high-resolution snRNA-seq atlas uncovers age-associated features of CRC and provides a resource for the community to understand EO-CRC biology.

Although the overall incidence and mortality of colorectal cancer (CRC) have been decreasing in the USA, the incidence of early-onset CRC (EO-CRC), defined as CRC diagnosed before age 50 years, has been increasing at an alarming rate of 1–3% per year since the mid-1990s[1,2], and it is now the leading cause of cancer death in adults aged 18–49 years[3]. As a result, the American Cancer Society[4] and the US Preventive Services Task Force[5] lowered the recommended CRC screening age from 50 to 45 years for average-risk individuals.

Despite more aggressive treatment, EO-CRC survival rates are not significantly better than those of average-onset CRC (AO-CRC), defined as CRC diagnosed at age ≥50 years[6–8], suggesting that patients with EO-CRC, particularly the youngest[9,10], may have biologically distinct and/or less responsive disease. While these statistics may be partially explained by young patients being 58% more likely than older patients to be diagnosed with advanced-stage disease[9,11], emerging data suggest that EO-CRC has distinct clinicopathologic features, including a higher frequency of distal colon and rectal primary tumors, as well as poorly differentiated and signet ring cell histologies[9,12]. Although some studies report differences in somatic genomic profiles in EO-CRC, including fewer *APC*, *KRAS* and *BRAF* p.V600E mutations and more *TP53* and

e-mail: rxavier@broadinstitute.org; acanderson@bwh.harvard.edu; Kimmie_Ng@DFCI.HARVARD.EDU

CTNNB1 mutations compared to AO-CRC[12,13], other studies found no significant differences after adjustment for microsatellite status and primary tumor sidedness[9]. Many EO-CRC studies are limited by a lack of information about prognostic factors, small EO-CRC cohorts[14–18], variable age cutoffs, restricted next-generation sequencing panels[9,12,13] or bulk sequencing approaches[15–17,19], and/or lack of adjustment for clinical characteristics[12,13,18].

To explore the transcriptional features underlying EO-CRC, we leveraged banked CRC specimens and generated a single-nucleus RNA sequencing (snRNA-seq) atlas of 189 primary CRC tumors spanning eight decades of life.

Results

snRNA-seq atlas of CRC across the age continuum

We profiled 54 EO-CRC and 135 AO-CRC primary tumor samples from banked frozen resection specimens using 3′ snRNA-seq (Fig. 1a), with >20 samples in most age decades (Fig. 1b). Tumors were annotated for patient sex, primary tumor sidedness, microsatellite status, neoadjuvant treatment history, stage and Lynch syndrome status (Fig. 1c and Supplementary Table 1). Consistent with prior reports[9,12], tumor sidedness and stage were significantly associated with age in our cohort (chi-square test, two-sided; sidedness: $\chi^2 = 19.3$, degrees of freedom = 2, false discovery rate (FDR) = 0.00032; stage: $\chi^2 = 13.0$, degrees of freedom = 3, FDR = 0.016), with more left-sided tumors (Fisher's exact test, two-sided; odds ratio (OR) 3.61, 95% confidence interval (CI) 1.85–7.04, $P = 0.0002$) and stage IV tumors (OR 2.55, 95% CI 1.23–5.27, $P = 0.017$) in EO-CRC compared to AO-CRC. There was no significant association between age and sex, microsatellite status or neoadjuvant treatment history in our cohort.

From the 189 patient tumor samples, we assembled an snRNA-seq CRC atlas of 850,222 high-quality nuclei (median per sample: 1,582 genes, 2,408 unique molecular identifiers (UMIs), 4,483 nuclei, 0.25% mitochondrial UMIs; Extended Data Fig. 1a and Supplementary Table 1). We identified seven cell lineages: epithelial (84.9%), stromal (5.6%), myeloid (3.1%), endothelial (2.7%), T (2.5%), B (1.1%) and glial/neuronal (0.1%) (Fig. 1d and Extended Data Fig. 1b). The high proportion of epithelial cells is expected for snRNA-seq[20]. All epithelial cells scored highly for a tumor gene signature[21] (Methods and Extended Data Fig. 1c) and were therefore assigned as malignant. Nuclei clustered primarily by cell lineage, microsatellite status and sidedness but not by age cohort, suggesting subtle age-related transcriptomic differences (Extended Data Fig. 1d).

We examined the consensus molecular subtypes (CMSs) in our cohort[22] using CMScaller[23] on pseudobulked data from our atlas (Methods). Aligned with established CMS features, CMS1 (immune) tumors were primarily microsatellite instability-high (MSI-H) and right-sided with higher immune cell proportions (Fig. 1e); CMS4 (mesenchymal) tumors had the highest stromal proportion; and CMS2 (canonical) and CMS3 (metabolic) tumors had the highest epithelial proportions[24,25]. EO-CRC tumors were more likely to be CMS2 than AO-CRC tumors (OR 2.58, 95% CI 1.28–5.21, FDR = 0.019) and less likely to be CMS1 (OR 0.21, 95% CI 0.06–0.72, FDR = 0.019) by Fisher's exact test. However, a multivariable logistic regression analysis modeling age continuously (scaled to unit variance) (Methods) with microsatellite status, sidedness, neoadjuvant treatment history, sex and stage as covariates showed that sidedness attenuated the age–CMS2 association significance (adjusted (adj.) OR 0.68, 95% CI 0.44–1.03, $P = 0.072$), likely influenced by the known age–sidedness association in EO-CRC. CMS1 remained significantly less likely in tumors from young patients after covariate adjustment (adj. OR 0.36, 95% CI 0.18–0.76, $P = 0.007$).

To identify granular malignant cell types, we subclustered the epithelial cells after patient batch correction (Methods) and annotated the subsets using canonical markers, colon-specific gene signatures, cluster marker genes and automated methods (Extended Data Fig. 2a,b and Supplementary Table 2). Among the 722,091 epithelial cells, we identified five subsets—stem cell-like (*LGR5*$^+$), colonocyte-like (*CEACAM1*$^+$), transit-amplifying (TA)-like/proliferating, goblet-like (*MUC2*$^+$) and enteroendocrine-like (*CHGA*$^+$) (Extended Data Fig. 2a and Fig. 1f)—as well as a mixed-marker cluster and a patient-specific cluster (MSI-H) that did not integrate during batch correction. Our epithelial subsets were highly concordant with established CRC cell states (Extended Data Fig. 2c–e)[26–28]. Visualizing the subset annotations on the non-batch-corrected Uniform Manifold Approximation and Projection (UMAP) revealed continuous annotation gradients across tumor-specific clusters and annotations that grouped together in the blended tumor clusters, confirming that batch correction preserved common biological cell states (Fig. 1f). Cell subset identity and patient-specific variation dominated epithelial transcriptional structure, with some clustering by microsatellite status and sidedness but not by age cohort (Fig. 1f–h and Extended Data Fig. 2f).

Epithelial-specific intrinsic CMS (iCMS) subtyping showed areas of high iCMS2 and iCMS3 scores (Extended Data Fig. 2g)[29]. Multivariable linear regression showed a nonsignificant trend toward higher iCMS2 scores in tumors from younger patients when sidedness was excluded as a covariate ($\beta = -0.01$, 95% CI −0.022 to 0.001, $P = 0.076$), but no association was found after adjusting for sidedness, consistent with the established age–sidedness association. However, tumors from younger patients had significantly lower iCMS3 scores ($\beta = 0.01$, 95% CI 0.004–0.022, $P = 0.004$).

Although nonepithelial cell types showed lower nuclei recovery per tumor (Extended Data Fig. 3a), our atlas contains a larger stromal compartment (47,712 nuclei) than that reported in previous CRC studies[26]. Following the epithelial compartment annotation process, we identified 13 stromal cell subsets: pericytes, adipocytes, six fibroblast subsets, myofibroblasts/smooth muscle cells, heat shock protein (HSP)-high stromal cells, mitochondrial-ribosomal-high stromal cells, cycling stromal cells and a mixed-marker subset (Extended Data Fig. 3b and Supplementary Table 2). Other nonepithelial compartments included myeloid (8 subsets; 26,248 nuclei), endothelial (5 subsets; 22,705 nuclei), T (10 subsets; 21,497 nuclei), B (6 subsets; 8,990 nuclei) and glial/neuronal (3 subsets; 979 nuclei) (Extended Data Fig. 3c–g).

Given the epithelial-dominant recovery in our atlas, we focused our subsequent analyses on epithelial cells. As microsatellite-stable (MSS) and MSI-H tumors are biologically and clinically distinct, we stratified analyses by microsatellite status (grouping MSI-low with MSS tumors), controlled for sidedness, neoadjuvant treatment history, sex and stage, and modeled age scaled to unit variance as a continuous variable when possible to capture gradual transcriptomic changes across the age continuum and avoid a rigid age cutoff (Methods). For analyses requiring categorical groups, we prioritized left-sided MSS tumors—the subgroup driving the increased EO-CRC incidence[30]—and compared EO-CRC to AO-CRC[31].

EO-CRC tumors exhibit a higher stromal than epithelial composition

To identify age-associated compositional shifts in MSS tumors (45 EO-CRC, 89 AO-CRC), we performed Dirichlet regression, modeling age as a continuous variable and adjusting for clinical covariates (Methods). The epithelial cell proportion significantly increased with age, indicating fewer epithelial cells in tumors from younger patients (Fig. 2a and Supplementary Table 3). Using epithelial cells as the reference, the proportions of the remaining cell lineages significantly decreased with age (that is, more cells from these lineages were observed in tumors from younger patients). Using granular cell subset annotations from all cell lineages, we observed that the proportions of the four largest epithelial cell types (stem-, TA-, colonocyte- and goblet-like) significantly increased with age (Fig. 2b and Supplementary Table 3); however, within the epithelial compartment alone, there were no significant compositional shifts. We

https://doi.org/10.1038/s41591-026-04721-y

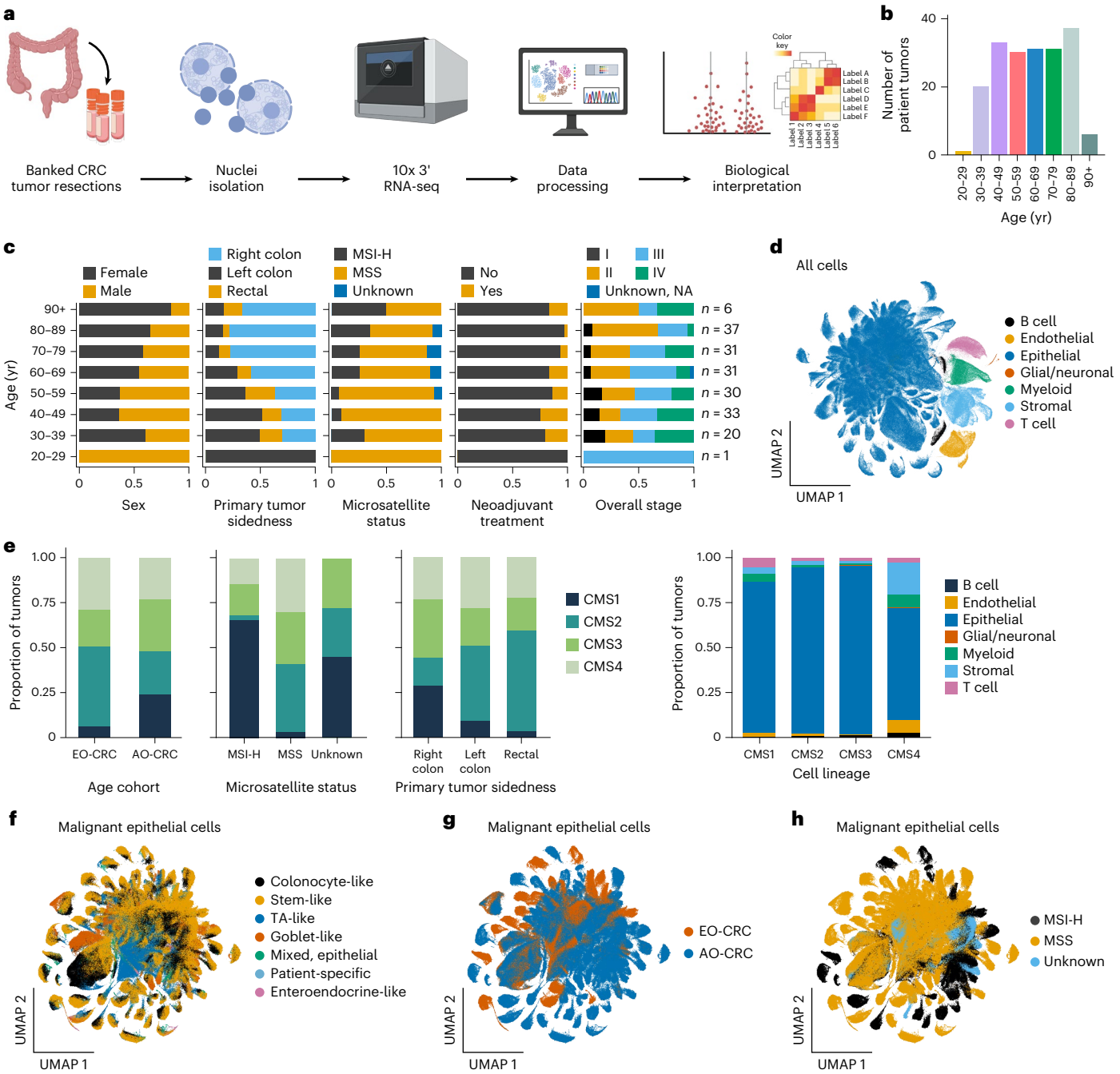
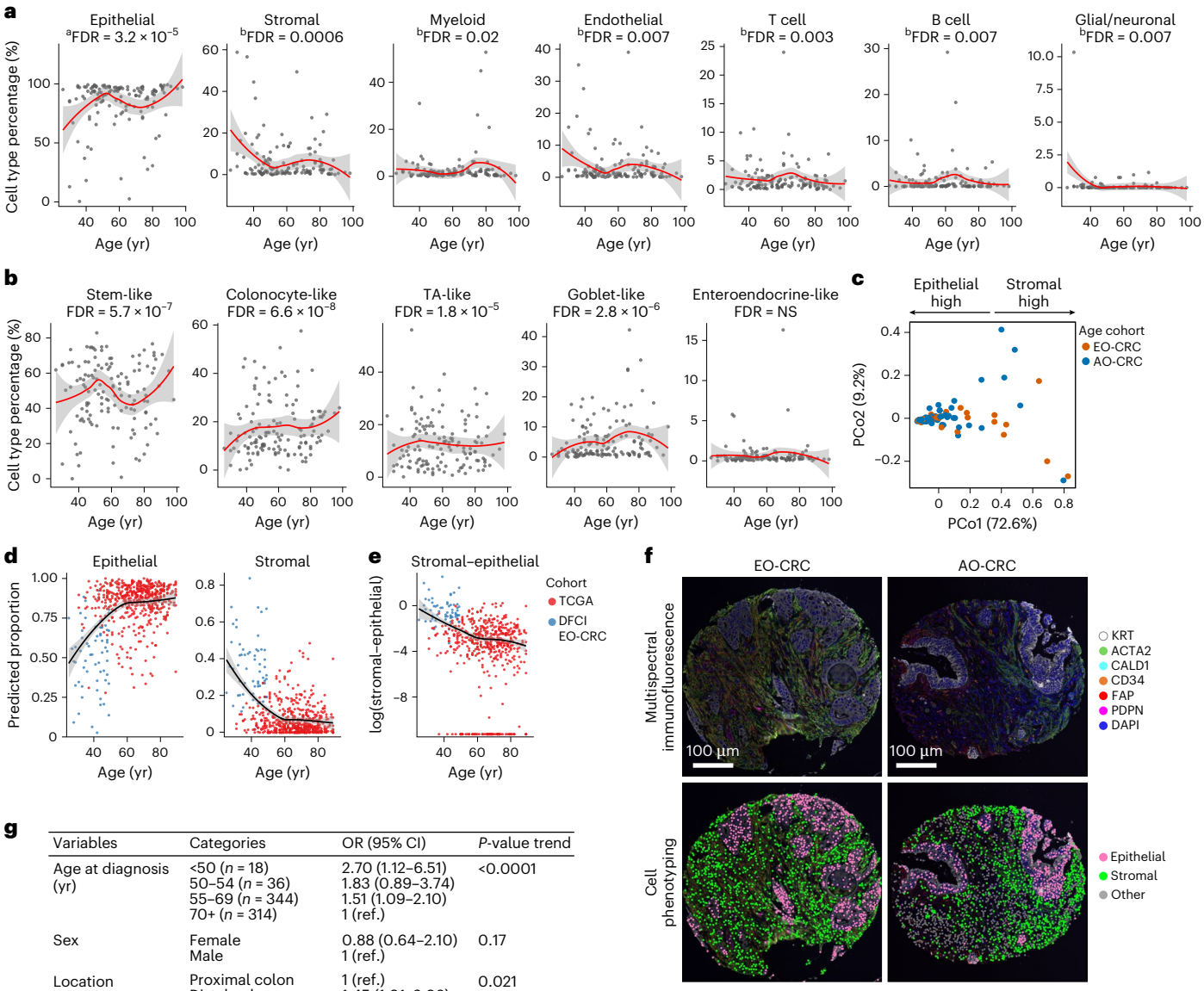


Fig. 1 | snRNA-seq EO-CRC atlas. a, Schematic illustrating nuclei isolation from 189 banked CRC tumor resection specimens and subsequent snRNA-seq profiling. High-quality transcriptomic data were retained and analyzed to explore differences with age. **b**,**c**, Bar plots showing the number of samples in each age decade (**b**) and the breakdown of key clinical metadata by age decade (**c**): patient sex; tumor sidedness (right colon, left colon, rectal); microsatellite status (MSI-H, MSS/MSI-low, unknown (NA, not available)); neoadjuvant treatment history (yes or no); and stage (*n* per age decade: 20–29 years, *n* = 1; 30–39 years, *n* = 20; 40–49 years, *n* = 33; 50–59 years, *n* = 30; 60–69 years, *n* = 31; 70–79 years, *n* = 31; 80–89 years, *n* = 37; 90+ years, *n* = 6; total: *n* = 189). **d**, UMAP embedding of all cells in the atlas (no batch correction; *n* = 189 tumors, 850,222 nuclei), colored by lineage-level cell type annotation. Each dot represents one nucleus. **e**, Bar plots showing the predicted CMS (Methods) group for each tumor by age cohort, microsatellite status and tumor sidedness (left), as well as cell type composition (right) (tumors with predicted CMS: *n* = 49 EO-CRC, 121 AO-CRC). **f**–**h**, UMAP embeddings of epithelial cells (no batch correction, *n* = 189 tumors, 722,091 nuclei) colored by granular cell subset annotation (**f**), patient age cohort (EO-CRC or AO-CRC) at the time of resection (**g**) and microsatellite status of the tumor (**h**). Each dot represents one nucleus. Schematic in **a** created in BioRender; Thalappillil, J. https://biorender.com/mqvba59 (2026).

confirmed these findings using scCODA[32], which showed that the proportion of epithelial cells significantly increased with age, with each nonepithelial cell type iteratively set as the reference (final parameter = 1.3–1.35; posterior inclusion probability > 0.95). Finally, a Bray–Curtis dissimilarity analysis of the cell lineages within the MSS cohort showed that differences in epithelial and stromal proportions drove the greatest source of variation (principal coordinate 1 (PCo1): 72.6%; Spearman ρ = −0.99 and 0.96, respectively), with a nonsignificant EO-CRC enrichment in the high-stroma region (Fig. 2c).

Repeating these analyses in the 9 EO-CRC and 33 AO-CRC MSI-H tumors did not reveal any lineage-level cell type that significantly changed with age (Dirichlet regression, FDR > 0.05) using the full model

https://doi.org/10.1038/s41591-026-04721-y



g

Variables	Categories	OR (95% CI)	*P*-value trend
Age at diagnosis (yr)	<50 (*n* = 18) 50–54 (*n* = 36) 55–69 (*n* = 344) 70+ (*n* = 314)	2.70 (1.12–6.51) 1.83 (0.89–3.74) 1.51 (1.09–2.10) 1 (ref.)	<0.0001
Sex	Female Male	0.88 (0.64–2.10) 1 (ref.)	0.17
Location	Proximal colon Distal colon Rectum	1 (ref.) 1.45 (1.01–2.08) 0.79 (0.53–1.19)	0.021
Stage	I, II III, IV	1 (ref.) 1.09 (0.79–1.49)	0.75

Fig. 2 | EO-CRC MSS tumors exhibit a higher stromal composition. **a**,**b**, Scatter plots showing the cell type percentage per MSS tumor ($n = 134$ MSS tumors: 45 EO-CRC, 89 AO-CRC) for all seven cell type lineages (**a**) and granular cell types (**b**). The red line indicates the fitted values from locally estimated scatterplot smoothing (LOESS), with shaded regions showing the 95% CI; FDRs were from a Dirichlet regression analysis modeling age as a continuous variable with sidedness, neoadjuvant treatment history, sex and stage as covariates ([a]full model; [b]model with epithelial cells as the reference). NS, no significance. **c**, Principal coordinates analysis of lineage-level compositional profiles of MSS tumor samples, colored by age cohort (Bray–Curtis) (Methods). Arrows indicate cell lineages with the greatest Spearman correlation with PCo1. **d**, Scatter plots showing MuSiC[34]-predicted epithelial (left) and stromal (right) proportions within the bulk RNA-seq validation cohort ($n = 564$ MSS tumors: 126 EO-CRC, 438 AO-CRC; blue: Dana-Farber Cancer Institute (DFCI) EO-CRC internal cohort; red: TCGA-COAD and TCGA-READ). The black line indicates the fitted values from LOESS, with shaded regions showing the 95% CI. **e**, Scatter plot showing the natural logarithm-transformed stromal-to-epithelial ratio from compositions represented in **d**; the black line and shaded region are as in **d**. **f**, Histopathology images colored by key cell type markers (top) and final cell type (bottom) from a representative EO-CRC and AO-CRC sample belonging to the Nurses' Health Study (NHS)/Health Professionals Follow-Up Study (HPFS) digital histopathology validation cohort. **g**, Inverse probability weighted multivariable logistic regression evaluating the associations of clinical features with quartiles of stromal-to-epithelial ratio from 712 MSS tumors. *P* values were two-sided, and adjustments were not made for multiple comparisons.

(Extended Data Fig. 4a and Supplementary Table 3). Among the granular subsets, the stem-like, TA-like, colonocyte-like, goblet-like and macrophage–monocyte subsets significantly increased in proportion with age, indicating lower proportions in MSI-H tumors from the youngest patients (Extended Data Fig. 4b and Supplementary Table 3). As in the MSS tumors, Bray–Curtis dissimilarity analysis did not identify age-associated differences, but PCo1 split samples by epithelial and stromal cell proportions (PCo1: 76.2%; Spearman $\rho = -0.99$ and 0.79, respectively) (Extended Data Fig. 4c).

Given the prevalence of MSS tumors in EO-CRC and the smaller sample size of MSI-H EO-CRC tumors in our cohort, we validated the observed compositional changes in the epithelial and stromal cells of MSS tumors. We leveraged an independent bulk RNA-seq validation cohort containing 497 MSS tumors from The Cancer Genome Atlas

https://doi.org/10.1038/s41591-026-04721-y

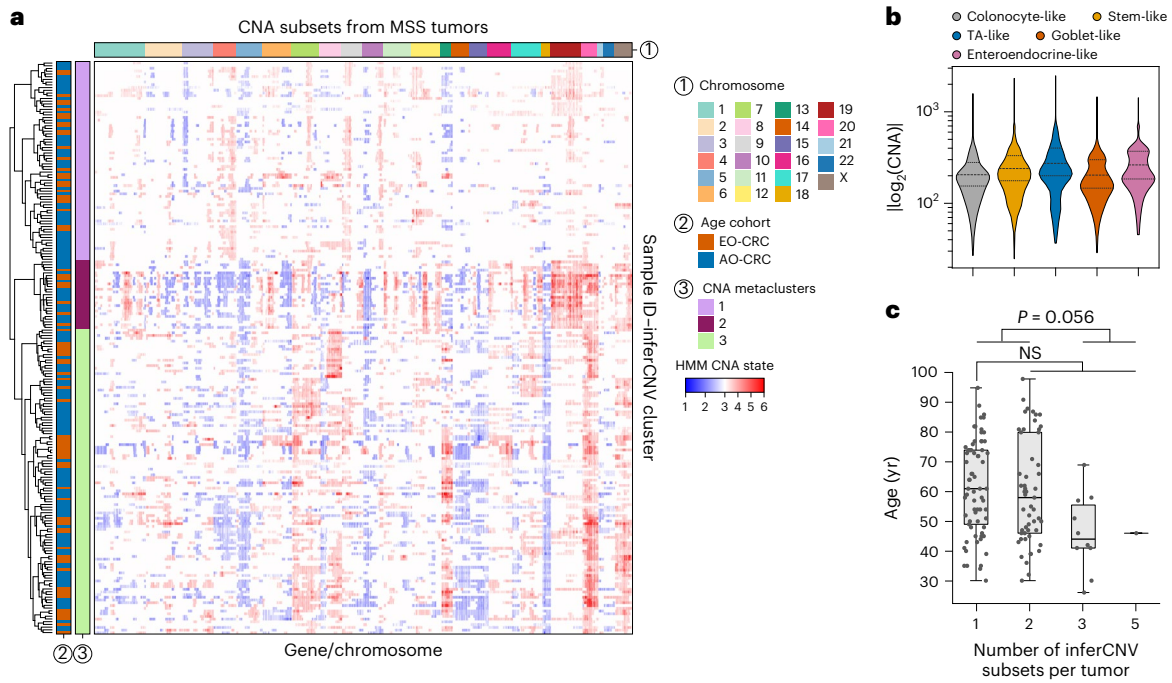


Fig. 3 | Age associations of CNA profiles in MSS tumors. CNA analysis was performed on MSS samples for which CNAs could be predicted by inferCNV (n = 44 EO-CRC tumors with 79 CNA subsets; 89 AO-CRC tumors with 131 CNA subsets). **a**, Heatmap of epithelial CNAs. Each row corresponds to a tumor-level inferCNV subset (one to five subsets per tumor) (Methods), and each column corresponds to a gene, ordered by chromosomal locus. Colors correspond to hidden Markov model (HMM)-derived CNA states (red, duplications, values 4–6; blue, deletions, values 1–2; white, neutral/no call, value 3). Top annotation bar (1): chromosome; left annotation bar (2): age cohort (orange, EO-CRC; blue, AO-CRC); right annotation bar (3): CNA metacluster. Chr. Y and chr. M are excluded from the visualization. **b**, Violin plot showing the distribution of CNA magnitude $|\log_2(\text{CNA})|$ by epithelial cell type. Lines indicate the median and the first and third quartiles. **c**, Box plot showing the distribution of the number of CNA subsets per tumor by age. Box limits indicate the interquartile range; the line indicates the median; and whiskers represent 1.5× the interquartile range. Significance was determined using two-sided multivariable logistic regression modeling the association of age with two or more (no significance) or three or more (P = 0.056) CNA subsets, adjusting for sidedness, neoadjuvant treatment history, sex, stage, mean UMI count per nucleus and epithelial nuclei count per tumor; no adjustment for multiple comparisons was made. Each dot represents one MSS tumor.

(TCGA) colon adenocarcinoma (COAD) and rectal adenocarcinoma (READ) datasets[33] (59 EO-CRC and 438 AO-CRC), supplemented with internally generated bulk RNA-seq data from 67 MSS EO-CRC tumors (not overlapping with our snRNA-seq cohort) (Supplementary Table 4). Cell type deconvolution[34] using our snRNA-seq atlas as a reference (Methods) showed an increase in epithelial cells and a decrease in stromal cells with age (Fig. 2d). A multivariable linear regression analysis modeling age continuously (including dataset batch and clinical covariates) revealed a significant decrease in the stromal-to-epithelial cell ratio with age (Fig. 2e; n = 493 finite samples, β = −0.19, 95% CI −0.34 to −0.05, P = 0.01), thus validating the compositional changes observed by snRNA-seq.

We further validated these compositional findings using multispectral immunofluorescence combined with machine learning (Methods) in 712 MSS tumors from the Nurses' Health Study (NHS)[35] and Health Professionals Follow-up Study (HPFS)[36] (see Supplementary Table 5 for baseline characteristics). The stromal-to-epithelial cell ratio was inversely correlated with age at diagnosis (Spearman's ρ = −0.11, P = 0.003). A multivariable logistic regression analysis (inverse probability weighting (IPW)[37]) controlling for sex, tumor location/sidedness and stage confirmed the higher stromal-to-epithelial cell ratio in younger age. Compared to patients aged ≥70 years, the multivariable-adjusted ORs (with 95% CIs) were 2.70 (1.12–6.51) for age <50 years, 1.83 (0.89–3.74) for age 50–54 years and 1.51 (1.09–2.10) for age 55–69 years (P_{trend} < 0.0001) (Fig. 2f,g and Supplementary Table 5). Collectively, our data indicate an altered cellular composition in MSS tumors from young patients, characterized by a higher proportion of stromal than epithelial cells.

EO-CRC and AO-CRC tumors exhibit similar CNA patterns

We used inferCNV[38] to infer copy number alterations (CNAs) in the epithelial cells within each tumor. To identify cohort-level CNA patterns, we merged the tumor-specific CNA subsets from 133 MSS and 42 MSI-H tumors (Methods). As expected, MSS tumors exhibited more CNAs than MSI-H tumors[33] (Fig. 3a and Extended Data Fig. 5a)[39]. Validating the inferred copy number losses and gains from our MSS cohort against 468 cBioPortal[40–42] MSS CRC tumors (MANTIS score < 0.4) and the published literature, we observed canonical gains (7p/q, 8q, 13q, 20q) and losses (1p, 8p, 17p, 18q), alongside iCMS2-specific gains on 20p and losses on 4p/q, 14q, 15q and 18p[43] (Fig. 3a and Extended Data Fig. 5b).

Within our MSS cohort, the observed metaclusters corresponded to high (green), moderate (maroon) and low (light purple) CNAs (Fig. 3a). Multivariable mixed-effects logistic regression adjusting for patient-level random effects and clinical covariates showed no significant association between age and CNA metaclusters. We confirmed that there were no CNA subclusters in normal epithelium by running inferCNV on MSS tumor samples and adjacent normal samples from Pelka et al.[26] and by performing unsupervised clustering with the CNA subsets from our atlas. The CNA subsets from our tumors separated from the adjacent normal subsets from Pelka et al.[26] (Extended Data Fig. 5c and Methods). Even our lowest-CNA metacluster had significantly more CNAs than the adjacent normal metacluster (Mann–Whitney U = 2,435.5, P = 0.0013; Extended Data Fig. 5d).

To evaluate age-associated changes in the CNA burden, we calculated the mean $|\log_2(\text{CNA})|$ per tumor. CNA burden varied across epithelial subsets in both the MSS and MSI-H cohorts, with the highest burden observed in TA-like cells (Fig. 3b and Extended Data Fig. 5e).

https://doi.org/10.1038/s41591-026-04721-y

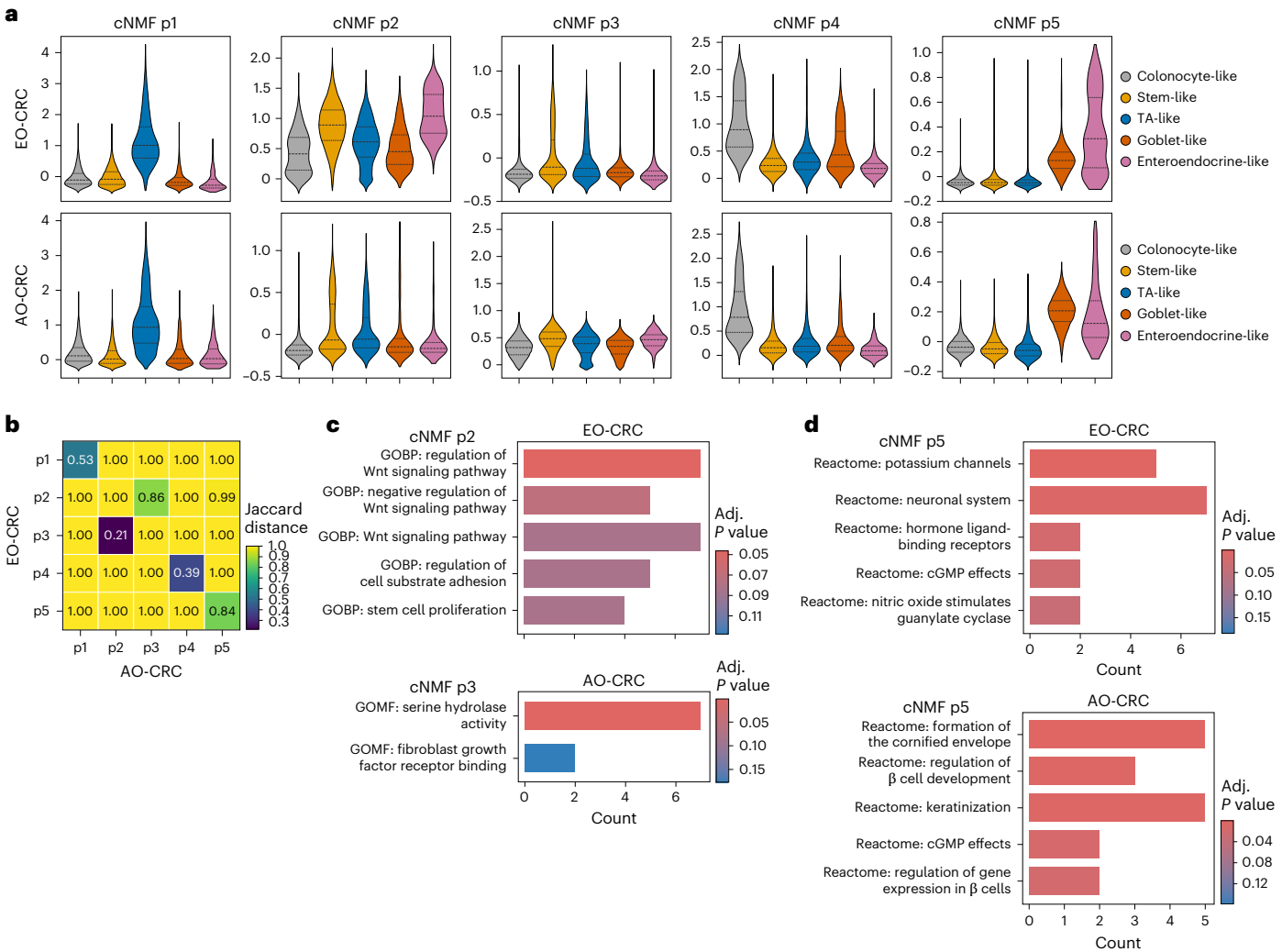


Fig. 4 | Cluster-agnostic metagene programs identified by cNMF for MSS left-sided tumors. a, Violin plots showing the distribution of cNMF metaprogram scores (Methods) in epithelial cells from MSS left-sided tumors, grouped by granular cell subset annotation, in EO-CRC (top) and AO-CRC (bottom) tumors (n = 24 EO-CRC and 25 AO-CRC). Lines indicate the median and the first and third quartiles. **b**, Heatmap showing the Jaccard distance between the top 50 genes in EO-CRC metaprograms p1–p5 and the AO-CRC metaprograms p1–p5. A Jaccard distance of 1.00 (yellow) indicates that the programs are maximally dissimilar between EO-CRC and AO-CRC, while a Jaccard distance of 0 (dark purple) indicates that the genes are identical. **c**, Top five enriched pathways from GO databases for the gene program pair that is most dissimilar (but shares some overlapping genes) between EO-CRC and AO-CRC (ranked by adj. $P < 0.2$; one-sided Fisher's exact test with Benjamini–Hochberg adjustment) (Methods). **d**, Top five enriched pathways from the Reactome database for the gene program that is the second-most dissimilar between EO-CRC and AO-CRC (ranked by adj. $P < 0.2$; one-sided Fisher's exact test with Benjamini–Hochberg adjustment) (Methods).

A multivariable linear regression analysis showed no significant association between CNA burden and age in MSS tumors (Methods). When sidedness was excluded as a covariate, there was a nonsignificant trend toward a higher CNA burden in MSS tumors from the youngest patients (β = −12.55, 95% CI −35.41 to 10.31, P = 0.28), consistent with the established age–sidedness association. In MSI-H tumors, the youngest patients had a significantly lower CNA burden (β = 34.0, 95% CI 7.1–60.9, P = 0.015).

We identified one to five CNA subsets per MSS tumor, reflecting intratumoral heterogeneity. A multivariable logistic regression analysis modeling age continuously and adjusting for clinical covariates, mean UMI count per nucleus and total epithelial cell count per sample revealed a borderline significant trend for MSS tumors from younger patients having three or more subsets per tumor (adj. OR 0.38, 95% CI 0.14–1.03, P = 0.056; Fig. 3c). The opposite, nonsignificant trend was observed in MSI-H tumors, with younger patients having fewer CNA subsets (Extended Data Fig. 5f). These results suggest that EO-CRC MSS tumors may have increased CNA heterogeneity.

cNMF reveals conserved gene programs in MSS left-sided CRC with subtle age-related differences

To capture cluster-agnostic gene programs and assess age-associated differences in tumor cell states, we leveraged consensus non-negative matrix factorization (cNMF)[44] and an approach used to identify pan-cancer gene programs[45,46] to define EO-CRC- and AO-CRC-specific epithelial cell metaprograms. Because cNMF cannot model continuous variables or include covariates, we focused our analyses on the 24 EO-CRC and 25 AO-CRC left-sided MSS tumors with >100 epithelial cells. Within the MSS left-sided tumor subset, age cohort was not significantly associated with stage, neoadjuvant treatment history or sex (two-sided chi-square test; FDR = 0.18, 0.68 and 1.0, respectively), nor with CMS classification (P = 0.76). We first identified ten gene programs per tumor based on stability and usage scores (Extended Data Fig. 6a,b) and then used unsupervised clustering of the gene program scores to define five robust, reproducible consensus metaprograms (p1–p5) per age cohort (Extended Data Fig. 6c,d and Methods). To examine the stability and similarity of the metaprograms with age, we calculated the

https://doi.org/10.1038/s41591-026-04721-y

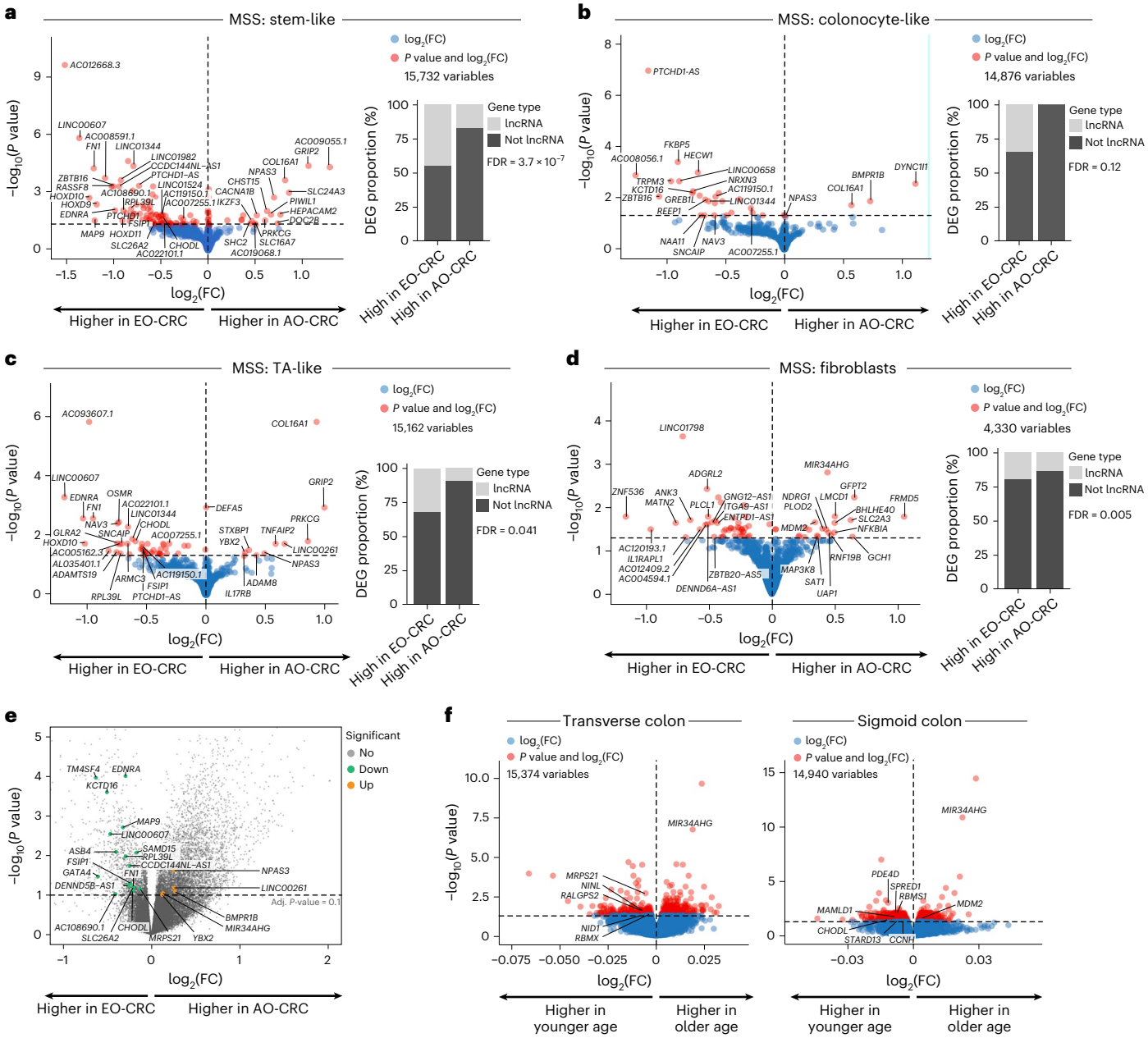


Fig. 5 | DEGs across the age continuum in MSS tumors. a–**d**, Volcano plots showing DEGs across the age continuum. Dots represent $\log_2$(fold change (FC)) per s.d. of age and adjusted *P* value (Benjamini–Hochberg) for each tested gene. Negative $\log_2$(FC) indicates decreasing expression as age increases, whereas positive $\log_2$(FC) indicates increasing expression as age increases (red, FDR < 0.05; blue, FDR ≥ 0.05). Bar plots indicate the fraction of significant DEGs (FDR < 0.05) that are lncRNAs or not lncRNAs; enrichment was determined by Fisher's exact test (Benjamini–Hochberg *P*-value correction) (Methods). **a**, Stem-like subset (n = 44 EO-CRC and 89 AO-CRC tumors). **b**, Colonocyte-like subset (n = 44 EO-CRC and 88 AO-CRC tumors). **c**, TA-like subset (n = 44 EO-CRC and 89 AO-CRC tumors). **d**, Fibroblasts (n = 27 EO-CRC and 46 AO-CRC tumors). Differential expression was calculated using DESeq2 based on expression data pseudobulked per patient within each granular subset, with age modeled as a continuous variable and sidedness, neoadjuvant treatment history, sex and stage included as covariates. Significance was determined using a two-sided Wald test (Methods). **e**, Volcano plot showing DEGs across the age continuum from the MSS tumor bulk RNA-seq validation cohort. Differential expression was determined using DESeq2, with age modeled continuously and sidedness, neoadjuvant treatment history, sex, stage and the bulk RNA-seq dataset as covariates (n = 564 MSS tumors: 126 EO-CRC, 438 AO-CRC). Significant genes (FDR < 0.1, two-sided Wald test with Benjamini–Hochberg correction) that overlap with the snRNA-seq DEGs are labeled (orange, higher in AO-CRC; green, higher in EO-CRC; gray, not significant; $\log_2$(FC) is per s.d. of age). **f**, Volcano plots showing genes found to change with age in GTEx transverse colon (left) and sigmoid colon (right) bulk RNA-seq data. Labeled genes overlap with the snRNA-seq DEGs (red dots, FDR < 0.05; blue dots, FDR ≥ 0.05). Differential expression was determined according to the method of García-Pérez et al.[70].

Jaccard distance between the top 50 genes from each of the EO-CRC and AO-CRC metaprograms (Fig. 4a,b and Supplementary Table 6).

The five EO-CRC and AO-CRC metaprograms exhibited varying levels of transcriptional overlap. EO-p2 and AO-p3 had the lowest gene overlap (Jaccard distance = 0.86); both showed high activity in stem-like and enteroendocrine-like subsets, although EO-p2 exhibited broader activity across all cell subsets, while AO-p3 activity was more restricted. EO-p2 genes were enriched for Wnt signaling (Gene Ontology Biological Process (GOBP): adj. P = 0.047), which is known to be dysregulated in CRC[47], whereas AO-p3 genes were enriched for serine

hydrolase activity (GO Molecular Function (GOMF): adj. $P = 0.0017$), which has also been implicated in CRC[48] (Fig. 4c).

EO-p5 and AO-p5 formed the second-most dissimilar program pair (Jaccard distance = 0.84), and these programs scored highest in the enteroendocrine-like and goblet-like subsets. EO-p5 was enriched for potassium ion channels and the neuronal system (Fig. 4d; Reactome: adj. $P = 0.0014$ and 0.00093, respectively), consistent with its high activity in the enteroendocrine-like subset. AO-CRC was enriched for keratinization (Reactome: adj. $P = 0.005$), including *KRT6A*, which has been implicated in promoting CRC invasion[49].

EO-p1 and AO-p1 had a Jaccard distance of 0.53, indicating that about half of the top genes overlapped between the age cohorts. These programs scored highest in the TA-like cells. Genes from both programs were enriched for ribosomal genes (Kyoto Encyclopedia of Genes and Genomes (KEGG): adj. $P = 5 \times 10^{-49}$ EO-CRC, 4.5×10^{-40} AO-CRC) (Methods), oxidative phosphorylation genes (KEGG: adj. $P = 5.2 \times 10^{-7}$ AO-CRC, 3.1×10^{-7} EO-CRC) and protein translation genes (Reactome: adj. $P = 1.4 \times 10^{-47}$ AO-CRC, 5.7×10^{-63} EO-CRC), indicative of a highly metabolically active epithelial subset.

The most conserved metaprogram pairs across the age cohorts were EO-p3/AO-p2 and EO-p4/AO-p4. EO-p3 and AO-p2 had the highest activity in the stem-like and TA-like cells, with only six and seven genes unique to AO-CRC and EO-CRC, respectively (Jaccard distance = 0.21). These programs were enriched for cell cycle and division (Reactome: adj. $P = 8.5 \times 10^{-11}$ AO-CRC, 1.6×10^{-11} EO-CRC), indicating shared proliferative features across age cohorts. EO-p4 and AO-p4 (Jaccard distance = 0.39) scored highest in the colonocyte-like cells in both age cohorts and additionally scored high in the goblet-like cells in EO-CRC. These programs were enriched for ERBB signaling (GOBP: adj. $P = 0.002$ AO-CRC, 0.03 EO-CRC) (Methods and Supplementary Table 6), which is known to be implicated in CRC[50,51].

Long noncoding RNA expression is enriched in CRC tumors from younger patients

Because our cohort included tumor samples spanning eight decades of life, with at least 20 tumors in most age decades, our snRNA-seq atlas was well powered to model age-associated gene expression changes. We performed this analysis on the 134 MSS tumors, modeling age continuously with clinical covariates (Methods and Supplementary Table 7).

Across the epithelial cell subsets, we found 215 genes with higher expression at a young age (that is, expression significantly decreased with increasing age) and 44 genes with higher expression at an older age (that is, expression significantly increased with increasing age). The greatest number of differentially expressed genes (DEGs) was found in the stem-like (Fig. 5a and Extended Data Fig. 7a), colonocyte-like (Fig. 5b and Extended Data Fig. 7b) and TA-like (Fig. 5c and Extended Data Fig. 7c) subsets (Supplementary Table 7).

Forty-nine DEGs were shared across two or more epithelial subsets (38 with higher expression in tumors from younger patients and 11 with higher expression in tumors from older patients), indicating recurring age-associated features. Among the DEGs with higher expression in young age, eight were shared across the stem-, TA- and colonocyte-like subsets: *SNCAIP*, a candidate oncogene in MSS CRC[52]; *FN1*, implicated in increased epithelial-to-mesenchymal transition (EMT)[53]; *CCDC88B*, implicated in colonic inflammation but mostly studied in immune cells[54]; and multiple long noncoding RNA (lncRNA) genes (*AC007255.1*, *AC022101.1*, *AC119150.1*, *LINC01344*, *PTCHD1-AS*) that are poorly studied in CRC.

LncRNAs comprised approximately 30–45% of the DEGs increased at a young age in the stem-, TA- and colonocyte-like epithelial subsets, with the highest proportion and most significant enrichment in EO-CRC in the stem-like subset (Fisher's exact test, two-sided; OR 3.4, FDR = 3.7×10^{-7}) (Fig. 5a). Approximately half of these lncRNAs were poorly characterized 'dark lncRNAs'. Others have established associations, including *LINC01524* (part of an inflammation-associated 13-lncRNA signature with prognostic value in CRC[55]) and *CCDC144NL-AS1* (shown to act as a sponge for miR-363-3p, resulting in increased *GALNT7* expression and promotion of CRC proliferation[56]). LncRNAs were also significantly enriched among the DEGs with higher expression at younger ages in the TA-like subset (OR 2.41, FDR = 0.041; Fig. 5c) and the fibroblast subset (OR 4.36, FDR = 0.005; Fig. 5d) and were nonsignificantly overrepresented in the colonocyte-like subset (OR 2.55, FDR = 0.12; Fig. 5b). Overall, these data suggest high activity of lncRNAs in the epithelial and stromal compartments of MSS tumors from young patients.

Among the 11 shared DEGs that exhibited increased expression with older age, only two genes were shared across the three largest epithelial subsets (stem-, TA- and colonocyte-like): *COL16A1*, encoding collagen type 16—which maintains the integrity of the extracellular matrix and has been investigated as a potential biomarker for CRC[57]—and *NPAS3*, encoding a tumor suppressor that has been studied in astrocytomas[58]. Another notable feature of the shared DEGs with higher expression in older age was that 3 of the 11 genes were shared with one or more fibroblast subsets (Fig. 5d and Extended Data Fig. 7d).

Validating our snRNA-seq DEGs against our independent bulk RNA-seq cohort revealed 19 overlapping DEGs: 14 genes with higher expression in tumors from younger patients and 5 genes with higher expression in tumors from older patients (Fig. 5e, Extended Data Fig. 7e and Supplementary Table 8), despite the differences in modality and resolution. Ten of the overlapping genes were shared across two or more epithelial subsets in our snRNA-seq data: *FN1*, *LINC00607*, *RPL39L*, *FSIP1*, *SLC26A2*, *EDNRA*, *KCTD16*, *NPAS3*, *LINC00261* and *CHODL*. Notably, among the DEGs specific to the stem-like subset and with higher expression in tumors from young patients, three lncRNAs—*CCDC144NL-AS1*, *DENND5B-AS1* and *AC108690.1*—were also DEGs in the bulk cohort.

To evaluate whether these DEGs represent tumor-associated age differences rather than normal colon aging, we examined their overlap with normal age-related genes in the Genotype–Tissue Expression (GTEx) Project[59] bulk RNA-seq data from the sigmoid and transverse colon (558 DEGs with expression increasing with age and 500 DEGs with expression decreasing with age). Only five of the epithelial snRNA-seq DEGs (*CHODL*, *MAMLD1*, *NINL*, *RBMS1*, *MRPS21*) overlapped (Fig. 5f); *CHODL* was also a bulk RNA-seq DEG. We further identified DEGs in adjacent normal tissue from the TCGA-COAD and TCGA-READ datasets and found no overlap with our snRNA-seq DEGs (Extended Data Fig. 7f and Supplementary Table 9). Together, these data provide confidence that our identified DEGs reflect tumor biology rather than normal colon aging.

Increased and unique cell–cell interactions in left-sided MSS EO-CRC tumors suggest a distinct tumor microenvironment

To evaluate changes in cell–cell interactions across the age cohorts, we used CellChat[60] at an intermediate annotation level that merged our granular epithelial annotations with our lineage-level nonepithelial cell annotations. As with cNMF, we compared EO-CRC and AO-CRC in left-sided MSS tumors because CellChat does not accommodate continuous variables or covariates (Supplementary Table 10). As noted above, age was not significantly associated with stage, neoadjuvant treatment history, sex or CMS classification within the MSS left-sided subset.

EO-CRC tumors had a higher number of interactions overall than AO-CRC tumors (Fig. 6a and Methods), driven by more interactions in stromal, myeloid and enteroendocrine-like cells, followed by colonocyte- and TA-like cells (Fig. 6b). A larger number of signaling pathways showed significantly increased interaction strength in EO-CRC compared to AO-CRC (29 versus 7), with 9 pathways unique to EO-CRC (for example, ACTIVIN, NECTIN, SPP1) (Fig. 6c) and none unique to AO-CRC. Pathways with higher or unique activity in EO-CRC indicated increased promotion of EMT through SPP1, FN1, GDF and ACTIVIN; promotion of angiogenesis through ANGPT, TENASCIN, SEMA3 and SEMA6; and involvement of neuronal signaling pathways

https://doi.org/10.1038/s41591-026-04721-y

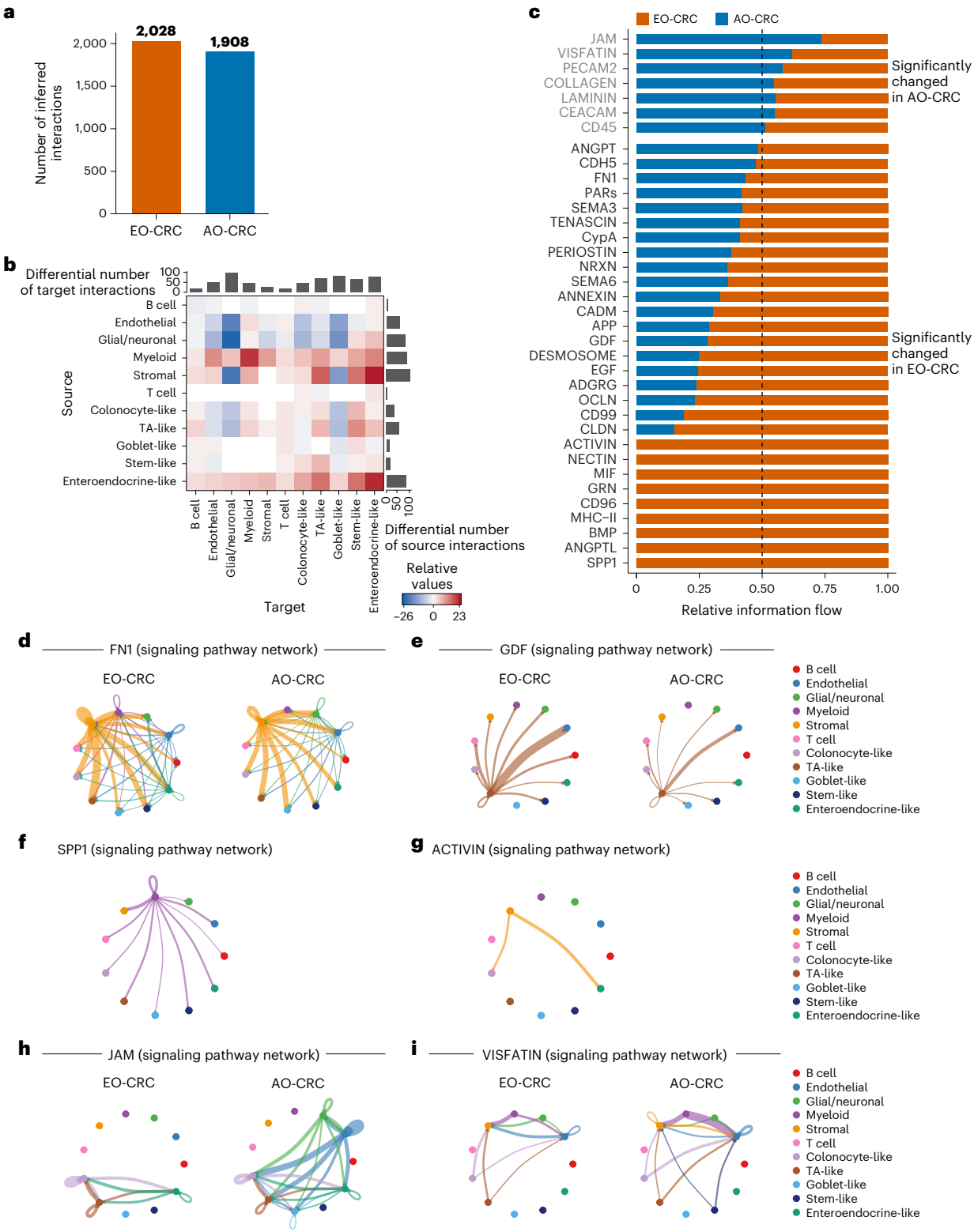


Fig. 6 | Cell–cell interactions. a, Bar plot showing the number of interactions as determined by CellChat 2.0 (ref. 60; Methods) in EO-CRC (orange) and AO-CRC (blue) MSS left-sided tumor samples ($n = 25$ EO-CRC and 26 AO-CRC). **b**, Heatmap showing the differential number of interactions between EO-CRC and AO-CRC samples, organized by cell subsets as sources (*y* axis) and targets (*x* axis) (red, more interactions in EO-CRC samples; blue, more interactions in AO-CRC samples). Top and right bars show the overall number of interactions for the indicated cell type and interaction direction. **c**, Bar plot showing the relative information flow for signaling pathways within each age cohort that are significantly changed (CellChat: two-sided Wilcoxon, $P < 0.05$) between EO-CRC (orange) and AO-CRC (blue). **d**–**i**, Circle plots showing interactions between cell subsets—split by age cohort—for select pathways with significant changes in information flow as indicated in **c**: FN1 (**d**), GDF (**e**), SPP1 (**f**), ACTIVIN (**g**), JAM (**h**) and VISFATIN (**i**). Plots for SPP1 (**f**) and ACTIVIN (**g**) show only EO-CRC interactions, as no significant AO-CRC interactions were found.

(NRXN, ADGRG). The major histocompatibility complex class II (MHC-II) pathway showed interactions of HLA-DR, HLA-DP and HLA-DQ with CD4 only in EO-CRC, involving exclusively myeloid–myeloid interactions. Pathways with higher activity in AO-CRC included JAM, which maintains endothelial and epithelial tight junctions; the adipokine VISFATIN; and pathways involved in tissue remodeling, such as COLLAGEN and LAMININ.

Looking at the circle plots, we found unique involvement of specific cell types in several EO-CRC-enriched pathways (Fig. 6d–i). In the FN1 pathway, myeloid and enteroendocrine-like cells were senders (cells expressing ligands that initiate signaling) in EO-CRC, but not in AO-CRC, and exhibited stronger interactions originating from stromal cells (Fig. 6d). In the GDF pathway, interactions targeting myeloid cells and B cells were identified only in EO-CRC, with stronger interactions involving other cell types, particularly endothelial cells (Fig. 6e). SPP1 was unique to EO-CRC and showed interactions between myeloid cells and all other cell types (Fig. 6f). In contrast, ACTIVIN, also unique to EO-CRC, was active only between stromal cells and enteroendocrine- and colonocyte-like cells (Fig. 6g). JAM pathway activity was increased in AO-CRC, with enhanced interactions between epithelial cell subsets, including goblet-like cells, and unique interactions involving glial/neuronal and endothelial cells (Fig. 6h). The VISFATIN pathway, also showing increased activity in AO-CRC, displayed unique stromal–stromal, stem-like–stromal and stem-like–endothelial cell interactions, as well as strong myeloid–endothelial cell interactions (Fig. 6i). Collectively, these data indicate that EO-CRC may have a tumor microenvironment characterized by EMT and angiogenesis, with unique myeloid cell involvement.

Discussion

Our snRNA-seq atlas of primary CRC tumors advances prior EO-CRC studies by combining a large sample size spanning the age continuum, single-nucleus resolution and rigorous covariate adjustment. Investigating CRC across the age continuum is particularly important given the known birth cohort effects that characterize rising EO-CRC incidence[61,62] and the clinical and prognostic differences affecting the youngest patients. We identified shared and distinct features across the age continuum in CRC tumors. Broad features, such as epithelial cell subsets, CNA patterns and core epithelial gene programs, were largely conserved across age, while tumor composition, epithelial and fibroblast gene expression, and cell–cell interactions in tumors from younger patients showed features indicative of a stromal-enriched tumor microenvironment and EMT-associated signaling.

The increased stromal relative to epithelial cell content was a notable feature of CRC tumors from young patients in our atlas, and has been associated with resistance to chemotherapy in several solid tumors[63], possibly contributing to poorer survival in patients with EO-CRC despite increased treatment dose intensity[10]. The higher stromal composition of colorectal tumors in young patients is consistent with the increased activity of EMT pathways in EO-CRC identified by our CellChat analysis, corroborating smaller published studies[14,64,65] that implicate stromal remodeling and EMT in EO-CRC.

Patients with EO-CRC were less likely to have CMS1 tumors in both unadjusted and multivariable-adjusted analyses, consistent with prior studies[9,66] and with the lower likelihood of right-sided and sporadic MSI-H tumors (which are more often CMS1) in young patients[67,68]. Although Willauer et al.[12] found more CMS1 tumors in patients <40 years old and those >70 years old, they used a different age cutoff (40 years) than that used in our study (50 years). They also did not perform multivariable adjustment and used a different data modality (bulk RNA-seq in ref. 12 versus snRNA-seq in our study).

We did not observe an EO-CRC-specific CNA profile or a change in CNA burden with age in MSS tumors. While a prior study reported increased CNA burden in CRC tumors from patients <40 years old[18], it included only 17 patients with EO-CRC, did not adjust for sidedness and used a different metric for calculating the CNA burden. In our study including 45 EO-CRC MSS tumors, an initial nonsignificant trend toward a higher CNA burden in the youngest patients was abolished when sidedness was included as a covariate. Despite the lack of significance in CNA burden, we observed a trend toward greater intratumor CNA heterogeneity (three or more CNA subsets per tumor) in EO-CRC, suggesting a potentially more complex evolutionary trajectory that warrants further study.

Our analysis of gene expression changes across the age continuum revealed high expression of lncRNAs in EO-CRC. Because most lncRNAs (~60–70%) are nuclear[69], our snRNA-seq atlas was well suited to detect lncRNA differences. This may explain their lack of detection in prior CRC studies and underscores the need to validate their potential role in EO-CRC.

We acknowledge several limitations of this atlas-based study. First, many of the reported findings should be considered hypothesis-generating and warrant validation in future studies. Given the observational nature of these results, we cannot comment on the precise mechanisms underlying EO-CRC biology, which are currently being investigated in ongoing functional studies. Second, the established correlation between EO-CRC and left-sided tumors makes it difficult to determine whether multivariable analyses that include sidedness as a covariate limit our ability to observe genuine EO-CRC biology or appropriately remove the sidedness signal (for example, as seen in our CMS2, iCMS2 and CNA burden analyses). Third, two of our analyses—the cNMF gene program and CellChat interaction analyses—could not account for potential confounding clinical covariates and had to be performed on smaller subsets of each age cohort. Finally, this study does not include matched normal tissue, necessitating the use of external data cohorts (GTEx, TCGA) as proxies for normal tissue and for evaluating observed age-associated differences in gene expression in malignant epithelial cells.

In conclusion, we constructed a large snRNA-seq atlas of primary colorectal tumors across the age continuum that identified both shared and distinct biology between EO-CRC and AO-CRC. Our observations of increased stromal composition, lncRNA activity and EMT in EO-CRC provide hypotheses for possible mechanistic drivers of the advanced stage and poorer survival that characterize patients with EO-CRC. These findings require further investigation aimed at advancing prevention and treatment strategies for patients with EO-CRC.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-026-04721-y.

References

1. Giannakis, M. & Ng, K. A common cancer at an uncommon age. *Science* **379**, 1088–1090 (2023).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. *CA Cancer J. Clin.* **67**, 7–30 (2017).
3. Siegel, R. L., Wagle, N. S. & Jemal, A. Leading cancer deaths in people younger than 50 years. *JAMA* **335**, 632–634 (2026).
4. Wolf, A. M. D. et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J. Clin.* **68**, 250–281 (2018).
5. US Preventive Services Task Force. Screening for Colorectal Cancer. *JAMA* **325**, 1965–1977 (2021).
6. Kneuertz, P. J. et al. Overtreatment of young adults with colon cancer: more intense treatments with unmatched survival gains. *JAMA Surg.* **150**, 402–409 (2015).
7. Lipsyc-Sharf, M. et al. Survival in young-onset metastatic colorectal cancer: findings from Cancer and Leukemia Group B (Alliance)/SWOG 80405. *J. Natl Cancer Inst.* **114**, 427–435 (2022).

8. Fontana, E. et al. Early-onset colorectal adenocarcinoma in the IDEA database: treatment adherence, toxicities, and outcomes with 3 and 6 months of adjuvant fluoropyrimidine and oxaliplatin. *J. Clin. Oncol.* **39**, 4009–4019 (2021).
9. Cercek, A. et al. A comprehensive comparison of early-onset and average-onset colorectal cancers. *J. Natl Cancer Inst.* **113**, 1683–1692 (2021).
10. Lipsyc-Sharf, M. et al. Circulating tumor DNA and late recurrence in high-risk hormone receptor-positive, human epidermal growth factor receptor 2-negative breast cancer. *J. Clin. Oncol.* **40**, 2408–2419 (2022).
11. Abdelsattar, Z. M. et al. Colorectal cancer outcomes and treatment patterns in patients too young for average-risk screening. *Cancer* **122**, 929–934 (2016).
12. Willauer, A. N. et al. Clinical and molecular characterization of early-onset colorectal cancer. *Cancer* **125**, 2002–2010 (2019).
13. Lieu, C. H. et al. Comprehensive genomic landscapes in early and later onset colorectal cancer. *Clin. Cancer Res.* **25**, 5852–5858 (2019).
14. Hou, Y. et al. Integrated muti-omics data and machine learning reveal CD151 as a key biomarker inducing chemoresistance in metabolic syndrome-related early-onset left-sided colorectal cancer. *Funct. Integr. Genomics* **25**, 122 (2025).
15. Du, M. et al. Integrated multi-omics approach to distinct molecular characterization and classification of early-onset colorectal cancer. *Cell Rep. Med.* **4**, 100974 (2023).
16. Holowatyj, A. N. et al. Distinct molecular phenotype of sporadic colorectal cancers among young patients based on multiomics analysis. *Gastroenterology* **158**, 1155–1158 (2020).
17. Venugopal, A. et al. ALDH1A1 expression is enriched in early-onset colorectal cancers. *Gastroenterology* **163**, 1679–1681 (2022).
18. Peng, T. & Zhan, Q. Single-cell integration analysis reveals molecular signatures of the tumor microenvironment in early-onset colorectal cancer. *NPJ Precis. Oncol.* **9**, 360 (2025).
19. Mauri, G. et al. Early-onset colorectal cancers exhibit distinctive placental-like features. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-7193450/v1 (2025).
20. Slyper, M. et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
21. Dohmen, J. et al. Identifying tumor cells at the single-cell level using machine learning. *Genome Biol.* **23**, 123 (2022).
22. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
23. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci. Rep.* **7**, 16618 (2017).
24. Sugai, T. et al. Analysis of molecular alterations in left- and right-sided colorectal carcinomas reveals distinct pathways of carcinogenesis: proposal for new molecular profile of colorectal carcinomas. *J. Mol. Diagn.* **8**, 193–201 (2006).
25. Thanki, K. et al. Consensus molecular subtypes of colorectal cancer and their clinical implications. *Int. Biol. Biomed. J.* **3**, 105–111 (2017).
26. Pelka, K. et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734–4752 (2021).
27. Kong, L. et al. The landscape of immune dysregulation in Crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon. *Immunity* **56**, 444–458 (2023).
28. 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).
29. Identification of two intrinsic epithelial subtypes of colorectal cancer. *Nat. Genet.* **54**, 924–925 (2022).
30. Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A. & Jemal, A. Colorectal cancer statistics, 2023. *CA Cancer J. Clin.* **73**, 233–254 (2023).
31. Spaander, M. C. W. et al. Young-onset colorectal cancer. *Nat. Rev. Dis. Primers* **9**, 21 (2023).
32. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).
33. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
34. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
35. Liao, X. et al. Aspirin use, tumor *PIK3CA* mutation, and colorectal-cancer survival. *N. Engl. J. Med.* **367**, 1596–1606 (2012).
36. Nishihara, R. et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. *N. Engl. J. Med.* **369**, 1095–1105 (2013).
37. Liu, L. et al. Utility of inverse probability weighting in molecular pathological epidemiology. *Eur. J. Epidemiol.* **33**, 381–392 (2018).
38. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. inferCNV of the Trinity CTAT Project. *GitHub* https://github.com/broadinstitute/inferCNV (2019).
39. Xie, T. et al. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PLoS ONE* **7**, e42001 (2012).
40. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
41. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
42. de Bruijn, I. et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).
43. 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).
44. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, e43803 (2019).
45. Gavish, A. et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
46. Kinker, G. S. et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
47. Housini, M. et al. Colorectal cancer: genetic alterations, novel biomarkers, current therapeutic strategies and clinical trials. *Gene* **892**, 147857 (2024).
48. Montrose, D. C. et al. Exogenous and endogenous sources of serine contribute to colon cancer metabolism, growth, and resistance to 5-fluorouracil. *Cancer Res.* **81**, 2275–2288 (2021).
49. Harada-Kagitani, S. et al. Keratin 6A is expressed at the invasive front and enhances the progression of colorectal cancer. *Lab. Invest.* **104**, 102075 (2024).
50. Strickler, J. H., Yoshino, T., Graham, R. P., Siena, S. & Bekaii-Saab, T. Diagnosis and treatment of ERBB2-positive metastatic colorectal cancer: a review. *JAMA Oncol.* **8**, 760–769 (2022).
51. Fang, J. Y. & Richardson, B. C. The MAPK signalling pathways and colorectal cancer. *Lancet Oncol.* **6**, 322–327 (2005).

52. Gylfe, A. E. et al. Identification of candidate oncogenes in human colorectal cancers with microsatellite instability. *Gastroenterology* **145**, 540–543 (2013).
53. Kalluri, R. & Weinberg, R. A. The basics of epithelial–mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
54. Fodil, N. et al. CCDC88B is required for pathogenesis of inflammatory bowel disease. *Nat. Commun.* **8**, 932 (2017).
55. Zhang, Z. et al. An inflammation-related lncRNA signature for prognostic prediction in colorectal cancer. *Cancer Rep. (Hoboken)* **7**, e70043 (2024).
56. Zhang, Y. et al. Long non-coding RNA CCDC144NL-AS1 promotes cell proliferation by regulating the miR-363-3p/GALNT7 axis in colorectal cancer. *J. Cancer* **13**, 752–763 (2022).
57. Jensen, C. et al. Serum type XVI collagen is associated with colorectal cancer and ulcerative colitis indicating a pathological role in gastrointestinal disorders. *Cancer Med.* **7**, 4619–4626 (2018).
58. Moreira, F. et al. NPAS3 demonstrates features of a tumor suppressive role in driving the progression of astrocytomas. *Am. J. Pathol.* **179**, 462–476 (2011).
59. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
60. Jin, S., Plikus, M. V. & Nie, Q. CellChat for systematic analysis of cell–cell communication from single-cell transcriptomics. *Nat. Protoc.* **20**, 180–219 (2025).
61. Siegel, R. L. et al. Global patterns and trends in colorectal cancer incidence in young adults. *Gut* **68**, 2179–2185 (2019).
62. Sung, H. et al. Differences in cancer rates among adults born between 1920 and 1990 in the USA: an analysis of population-based cancer registry data. *Lancet Public Health* **9**, e583–e593 (2024).
63. Valkenburg, K. C., de Groot, A. E. & Pienta, K. J. Targeting the tumour stroma to improve cancer therapy. *Nat. Rev. Clin. Oncol.* **15**, 366–381 (2018).
64. Kim, S.-C. et al. Identification of a novel fusion gene, *FAM174A–WWC1*, in early-onset colorectal cancer: establishment and characterization of four human cancer cell lines from early-onset colorectal cancers. *Transl. Oncol.* **12**, 1185–1195 (2019).
65. Furuhashi, S. et al. Spatial profiling of cancer-associated fibroblasts of sporadic early onset colon cancer microenvironment. *NPJ Precis. Oncol.* **7**, 118 (2023).
66. Pretta, A. et al. Early-onset colorectal cancer patients exhibit a distinct molecular fingerprint: insights from a large-scale NGS study of 1209 patients. *ESMO Open* **10**, 105756 (2025).
67. Sinicrope, F. A. Increasing incidence of early-onset colorectal cancer. *N. Engl. J. Med.* **386**, 1547–1558 (2022).
68. Sinicrope, F. A. Lynch syndrome-associated colorectal cancer. *N. Engl. J. Med.* **379**, 764–773 (2018).
69. Cabili, M. N. et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).
70. García-Pérez, R. et al. The landscape of expression and alternative splicing variation across human traits. *Cell Genom.* **3**, 100244 (2022).

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

Caroline B. M. Porter[1], Shane P. Murphy[1], Jules Cazaubiel[2,3], Toni M. Delorey[1], Liat Amir-Zilberstein[1], Orr Ashenberg[1], Liubou Klindziuk[1], Elizabeth Andrews[2], William Tan[2], Kosuke Matsuda[4,5,6], Atsushi Kondo[4], Satoko Ugai[4,6,7], Tomotaka Ugai[4,6,7], Shuji Ogino[3,4,7,8,9], Emily Finan[1], Jacques Deguine[1], Ewa Sicinska[2], Lauren K. Brais[2], Jennifer S. Thalappillil[2], Marios Giannakis[2,3], Ramnik J. Xavier[1,10,11,13] ✉, Ana C. Anderson[11,12,13] ✉ & Kimmie Ng[2,13] ✉

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Program in Molecular Pathological Epidemiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [5]Department of Pathology, Keio University, Tokyo, Japan. [6]Division of Integrative Cancer Science, National Cancer Center Research Institute, Tokyo, Japan. [7]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Division of Nutrition, Harvard Medical School, Boston, MA, USA. [9]Institute of Science Tokyo, Tokyo, Japan. [10]Center for Computational and Integrative Biology, Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [11]Gene Lay Institute of Immunology and Inflammation of Brigham and Women's Hospital, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [12]Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [13]These authors jointly supervised this work: Ramnik J. Xavier, Ana C. Anderson, Kimmie Ng. ✉e-mail: rxavier@broadinstitute.org; acanderson@bwh.harvard.edu; Kimmie_Ng@DFCI.HARVARD.EDU

Methods

The research in this study complies with all relevant ethical regulations. The protocols used to generate our snRNA-seq and bulk RNA-seq data were approved by the institutional review board of the Dana-Farber/Harvard Cancer Center. Protocols related to the NHS/HPFS validation cohort were approved by the institutional review boards of Brigham and Women's Hospital and the Harvard T.H. Chan School of Public Health. Written informed consent was obtained from all participants, who were not compensated. For the snRNA-seq and bulk RNA-seq cohorts, the obtained consent allows samples and clinical information to be shared with internal and external collaborators for gastrointestinal cancer research. Patient sex was collected from clinical records. As all available tissue was processed, sex was not considered in the study design. Sex, age and other clinical metadata are reported in a disaggregated manner in Supplementary Table 1 (snRNA-seq) and Supplementary Table 4 (bulk RNA-seq). The ages of patients 90 years and older have been aggregated into a single '90+' category. For the NHS/HPFS multispectral immunofluorescence cohort, sex was self-reported and was only considered in the study design due to the nature of the cohort compositions (NHS is a cohort of women and HPFS is a cohort of men). Aggregated clinical metadata are reported in Supplementary Table 5. Instructions for accessing the aggregated clinical metadata are provided in the Data Availability statement.

Patient samples and clinical data

Primary colorectal tumor samples with a pathologic diagnosis of adenocarcinoma were collected from the operating room from patients who consented to participate in research under Dana-Farber/Harvard Cancer Center Protocols 03-189, 11-104, 14-091, 14-408, 17-000 and/or 20-000. Tumors were snap-frozen and stored at −80 °C in the DFCI Gastrointestinal Cancer Center Biobank. Demographic, clinical, pathologic and molecular characteristics, as well as treatment data, were manually abstracted from the medical records by a team of clinical research coordinators from the DFCI Gastrointestinal Cancer Center Biobank. The final cohort included 189 primary tumor samples collected from 189 unique patients.

Cohort statistical associations and enrichments

Associations between age cohort and clinical metadata were evaluated using chi-square tests (chi2_contingency) from SciPy (v1.14.1), and *P* values were adjusted for multiple comparisons using the multipletests function in statsmodels (v0.14.2) to calculate the Benjamini–Hochberg FDR. Patients with missing values for the association being tested were excluded from the analysis. To calculate enrichment within an age cohort of a specific metadata category, ORs with 95% CIs and associated *P* values were computed using Fisher's exact test (fisher_exact) and the Table2x2 class from statsmodels.

Nuclei isolation, library construction and sequencing

A biopsy-sized (~3 mm^2) piece of tissue was manually chopped in CST dissociation buffer for 10 min on ice[20]. The dissociated tissue was then filtered through a 40-µm strainer (pluriSelect, 43-10040-60) and centrifuged at 500*g* and 4 °C for 10 min. The nuclei were resuspended in PBS (Gibco, 10010031) with 1% BSA (Miltenyi Biotec, 130-091-376), filtered and counted after staining with 4′,6-diamidino-2-phenylindole (DAPI) (Sigma-Aldrich, D9542-1MG). Approximately 12,000 nuclei were loaded onto a 10x Chromium Controller and processed using the 10x 3′ v3.1 dual index assay (10x Genomics, 1000268). Following single-nuclei barcoding, reverse transcription and library construction were performed according to the manufacturer's guidelines (10x Genomics). Samples were sequenced on a NovaSeq 6000 using S2 flow cells (Illumina) with the following dual-indexed, paired-end read configuration: read 1, 28 nucleotides (nt); read 2, 90 nt; index read 1, 10 nt; index read 2, 10 nt.

Data preprocessing, quality control and atlas integration

Raw sequencing files were demultiplexed using bclconvert (v3.10.5)[71], run with the Terra Workflow Description Language (WDL). Gene expression count matrix was generated using Cell Ranger v7.0.1 (through the Cumulus v2.3.0 WDL[72,73] on Terra[74]). Ambient RNA counts and empty droplets were removed using CellBender remove-background (v0.2.0)[75,76]. The Cumulus WDL was used to further remove nuclei with fewer than 200 genes, fewer than 400 UMIs or more than 50% of counts mapped to mitochondrial genes[72,77]. Count matrices from individual samples were integrated using Scanpy (v1.10.2), after which Cumulus-predicted doublets were removed, as well as cells with ≥10,000 genes or ≥50,000 UMIs. Cells in the integrated atlas were normalized to 1×10^4 total UMIs and log-transformed. Highly variable genes were selected using scanpy.pp.highly_variable_genes (min_mean = 0.0125, max_mean = 3, min_disp = 0.5) from Scanpy. Principal component analysis (PCA) was performed using scanpy.tl.pca (svd_solver = 'arpack'); nearest neighbors were calculated using scanpy.pp.neighbors (n_neighbors = 10, n_pcs = 40); a UMAP embedding was generated using scanpy.tl.umap; and Leiden clustering was performed using scanpy.tl.leiden—all implemented in the Scanpy package.

Cell subset annotation

Cells were annotated at both the lineage and granular levels. Seven lineage-level annotations (epithelial, myeloid, B, T, endothelial, glial/neuronal, stromal) were assigned to each Leiden cluster (resolution = 0.1) using published[27] gene signatures for colon cells. Selected marker genes for each lineage are shown in Extended Data Fig. 2b. Granular annotations were assigned by subclustering raw counts within each lineage, followed by normalization, log transformation, calculation of variable genes and PCA, as described above. Within each lineage except glial/neuronal, we then performed patient batch correction using scanpy.pp.harmony_integrate from harmonypy (v0.0.10)[78], generated a UMAP and performed Leiden clustering at multiple resolutions. For the glial/neuronal subset, UMAP, clustering and annotation were performed without batch correction due to the smaller number of tumors. Non-batch-corrected UMAPs for all lineage subsets were also generated to facilitate the visualization of annotations and clinical metadata.

Clusters were annotated using a mix of published gene signatures[26,27], canonical cell type markers, calculated marker genes and celltypist.annotate from CellTypist (v1.6.3)[79,80] (model = Cells_Intestinal_Tract, model = Human_Colorectal_Cancer). Gene signatures were scored using scanpy.tl.score_genes, and marker genes were identified using scanpy.tl.rank_genes_groups from Scanpy (v1.10.2). The top 30 marker genes for the selected granular epithelial subsets are shown in Extended Data Fig. 3b. These subset markers, as well as the marker genes for the Leiden clusters from each lineage subset, are provided in Supplementary Table 2. In the epithelial subsets, three small patient-specific clusters remained after batch correction. Two of these clusters (6 and 7) were merged with the closest major epithelial subset based on transcriptional similarity; the third cluster was sufficiently distinct to be annotated as a separate patient-specific cluster.

CMS and iCMS classification

The CMS for each sample was classified using CMScaller[22,23] (v2.0.1) in R. Raw counts for all cells in the atlas were pseudobulked per tumor using the get_pseudobulk function (mode = 'sum', min_cells = 10, min_counts = 1000) from the decoupleR package (v1.8.0) in Python. Genes expressed in at least 10% of samples in either the AO-CRC or EO-CRC cohort were retained using filter_by_expr (min_prop = 0.1, min_count = 10) from decoupleR. Gene symbols were converted to Ensembl IDs using getBM from the biomaRt package (v2.58.2) in R, and pseudobulk counts for each sample were classified using CMScaller with the single-sample prediction (SSP) protocol. A two-sided Fisher's exact test was used to determine the enrichment of each CMS with age (one versus all). Associations between specific CMSs (one versus

all) and age were further evaluated using the logit function from the statsmodels package (v0.14.6) in Python to run a logistic regression analysis modeling age scaled to unit variance (*z*-scored) as a continuous variable with microsatellite status, sidedness, neoadjuvant treatment history, sex and stage as covariates. Significance was determined using a two-sided Wald test. Rectal and treated samples were removed from the CMS1 multivariable logistic regression test because no samples in these clinical groups were classified as CMS1.

iCMS2/3 scores were calculated using scanpy.tl.score_genes from Scanpy (v1.10.2) and the published gene sets from Joanito et al.[43]. To determine associations between mean iCMS2/3 scores and age as well as other clinical covariates, a multivariable linear regression was fit using ordinary least squares (OLS) in the statsmodels package (v0.14.6) in Python, modeling age scaled to unit variance (*z*-scored) as a continuous variable with microsatellite status, sidedness, neoadjuvant treatment history, sex and stage included as covariates. Tumors with unknown clinical metadata for any included covariate were excluded from the analysis. Significance was determined using a two-sided *t* test.

Compositional changes and principal coordinates analysis

Changes in cell subset composition across the age continuum were identified separately in the MSS and MSI cohorts using the DirichReg R package (v0.7.1), with age scaled to unit variance (*z*-scored) as a continuous variable and sidedness, neoadjuvant treatment history, sex and stage as covariates. At the lineage level, we ran Dirichlet regression in two ways: (1) using the 'common' model with no reference variable, where all cell types are modeled relative to each other and (2) using the 'alternative' model with epithelial cells as the reference variable. For the common model, we used counts ~ age_scaled + sideness + therapy + sex + stage as the design formula; for the alternative model, we used counts ~ age_scaled + side + therapy + sex + stage | age + side + therapy + sex + stage. Multiple hypothesis correction was performed across *P* values for all covariates by calculating the Benjamini–Hochberg FDR using p.adjust(pvals, method = 'fdr') in R. Dirichlet regression (common model and design formula indicated above) was also performed using all granular cell types. For the MSS tumors only, we further evaluated compositional changes in the granular subsets within the epithelial compartment (common model). Lineage-level compositional changes were confirmed using scCODA integrated into the pertpy package (v1.0.5) in Python, by applying the prepare, run_nuts and get_effect_df functions with default hyperparameters. In the scCODA analysis, each cell type was iteratively set as the reference cell type; min–max normalized age was modeled continuously with sidedness, neoadjuvant treatment history, sex and stage as covariates; and the final parameters were evaluated (posterior inclusion probability > 0.95 in all iterations). Principal coordinates of cell type composition were calculated from Bray–Curtis dissimilarities of lineage-level cell type proportions per sample, using the pco function from the labdsv R package (v2.1.0). Granular annotations with 'mixed-marker' assignments (that is, a mix of markers from the assigned lineage as well as other lineages) were excluded from all composition analyses[27].

inferCNV

CNAs were inferred using inferCNV (v1.19.1)[38], run with the Terra WDL infercnv_wf[81]. The WDL was run once per tumor sample, with epithelial cells as the test group and all nonepithelial cells as the reference. Cell subset annotations that showed marker gene expression from multiple lineages (containing 'mixed' in their name) were excluded from the analysis. The following arguments were specified: --analysis_mode = 'subclusters' --HMM --denoise --cutoff = 0.1 --leiden_function = 'modularity'. Two samples (COLFR0034_T1, COLFR0366_T1) were excluded from the downstream analysis because they had insufficient counts within one or more subsets.

The heatmaps in Fig. 3a and Extended Data Fig. 5a,c were generated by extracting the HMM cell state information from HMM_CNV_predictions.HMMi6.leiden.hmm_mode-subclusters.Pnorm_0.5.pred_cnv_genes.dat, resulting in a matrix of CNA subclusters by genes, with entries ranging from 1 to 6 (values of 1 and 2 indicate deletions; 3 indicates no CNA; 4–6 indicate duplications). Genes without a predicted cell state were assigned a default value of 3, and the genes were ordered according to their locus. The resulting cell states were clustered using the clustermap function from the Seaborn Python package (v0.13.2), with row distances calculated using pdist and linkage from SciPy (v1.14.1; method = 'ward' and metric = 'correlation'). Clustering was performed separately for the MSS and MSI samples in our cohort.

To assess the consistency of the predicted CNAs with known losses and gains in bulk RNA-seq cohorts, reference CNA profiles were accessed from cBioPortal on June 10, 2026; only samples with a MANTIS score of <0.4 (MSS) were used for comparison. For comparison to CNAs in adjacent normal tissue from the Pelka et al.[26] cohort, we ran inferCNV as described above on MSS tumor samples and the adjacent normal samples. The CNA subclusters from the Pelka et al.[26] samples were then clustered with the CNA subclusters from the MSS samples in our atlas. To evaluate age associations among the three CNA metaclusters formed across the MSS tumor samples in our cohort, we separately compared the high CNA and moderately high CNA metaclusters to the low CNA metacluster using a multivariable mixed-effects regression model with patient ID (that is, tumor sample) as a random effect, as there can be more than one CNA subset per tumor. We performed this analysis in R using glmer from the lme4 package (v1.1.38), again using age scaled to unit variance (*z*-scored) as a continuous variable: cluster ~ age_scaled + side + treat + stage + sex + (1 | patient_id) and family = binomial. Significance was determined using a two-sided Wald test.

CNA burden was calculated per cell by summing the absolute values of the $\log_2$-transformed CNA residual expression from the object: run.final.infercnv_obj. To evaluate the overall CNA burden per tumor across the age continuum, we averaged the per-cell absolute $\log_2$ values within each tumor and performed multivariable linear regression using OLS implemented in the statsmodels package (v0.14.6) in Python, modeling scaled age as a continuous variable with sidedness, neoadjuvant treatment history, sex and stage as covariates. Significance was determined using a two-sided *t* test.

To assess the change in the likelihood of having two or more and three or more CNA subsets per tumor with age, we used a multivariable logistic regression method (logit from statsmodels), again modeling scaled age continuously with sidedness, neoadjuvant treatment history, sex and stage as covariates. In this regression, we also added a covariate for the mean UMI count per tumor as a proxy for sequencing depth and a covariate for the total number of epithelial cells to capture bias due to detection power. Significance was determined using a two-sided Wald test.

Common gene programs using cNMF

To identify recurrent, shared malignant epithelial gene programs separately within the EO-CRC and AO-CRC cohorts, we used a two-stage workflow adapted from Kinker et al.[46], which first identified gene programs specific to individual tumors and then metagene programs at the age cohort level.

Patient-specific gene programs. We first ran cNMF[44] separately on the raw counts from the epithelial cells of each MSS left-sided tumor sample. Samples with fewer than 100 epithelial cells were excluded, and genes were filtered to retain protein-coding genes expressed in three or more cells (n = 24 EO-CRC and 25 AO-CRC tumor samples). We used the cNMF package (v1.6.0) in Python and tested a number of factorization ranks (k = 4–15). Each cNMF run was performed on the variable genes, calculated from normalized and log-transformed counts using scanpy.pp.highly_variable_genes (min_mean = 0.0125, max_mean = 5, min_disp = 0.5, flavor = 'seurat') from Scanpy (v1.10.2), with 100 iterations, random_seed = 14 and parallelization across four workers.

We ultimately selected $k = 10$ for each tumor based on visualization of the stability and error metrics from the k_selection_plot function, displayed across all tumors (Extended Data Fig. 6a). We also visualized the program usage scores on individual tumor UMAPs at various values of k to assess the activity of each program (see Extended Data Fig. 6b for a summarized visualization of this process). Our final selection of $k = 10$ provided a balance of stability and interpretability across tumors from each age cohort. Consensus gene programs were computed using the consensus command from cNMF for the selected k and a density threshold of 0.2.

Within each tumor sample, the cNMF usage values were normalized to sum to one within each cell and stored as observation-level metadata in each tumor's AnnData object. Additionally, cNMF z-scores for each gene were added to the AnnData object for each tumor as gene-level metadata, representing gene-level contribution weights to each tumor-specific gene expression program (GEP). Separately within the EO-CRC and AO-CRC age cohorts, the top 50 genes were identified for each tumor-specific GEP z-score across all tumors in the age cohort based on sorted z-scores.

Age cohort metagene programs. GEP scores for each tumor-specific program were calculated across all cells in each age cohort using the scanpy.tl.score_genes function in Scanpy (v1.10.2). The mean tumor-specific program scores were visualized using the clustermap function from the Seaborn Python package (v0.13.2). The rows and columns were clustered using the pdist and linkage functions from SciPy (v1.14.1; method = 'ward' and metric = 'correlation'), with rows representing the tumor-specific gene programs (10 programs × 24 tumors = 240 programs in EO-CRC and 10 programs × 25 tumors = 250 programs in AO-CRC) and columns representing the individual samples.

Metagene programs that were active across the EO-CRC and AO-CRC tumor cohorts were identified using unsupervised row clustering. These metaclusters represent robust and recurrent gene programs within each age cohort. We chose the final number of metaclusters per cohort using four criteria: (1) visualization of the resulting metaprograms on individual tumor UMAPs; (2) the overall activity level and patient specificity of the resulting metaprograms scored across all cells within an age cohort; (3) visualization of the mean program scores on the heatmap; and (4) the mean silhouette score. We ultimately selected five metaclusters per age cohort because this provided the best balance among the four metrics mentioned above. The top 50 genes for each EO-CRC and AO-CRC common gene program were determined by sorting the mean z-scores from the common gene program cluster tumor-specific GEPs.

Metagene program evaluation. Similarity between age cohort-specific gene programs was determined by calculating the Jaccard distance between all pairs of EO-CRC and AO-CRC programs. To determine gene program alignment with epithelial granular subsets, the top 50 genes for common programs within each age cohort were scored using scanpy.tl.score_genes, and the scores for each program were visualized with violin plots grouped by cell subset. Enriched pathways were calculated for the top 50 genes within each EO-CRC and AO-CRC cNMF common (as well as age cohort-specific) gene program. Enrichment was performed using the enricher function from the clusterProfiler package (v4.10.1; pAdjustMethod = 'BH'), using collections from MSigDB (msigdbr v25.1.0) and the default background (all genes appearing in the MSigDB collection). Significance was determined using a one-sided Fisher's exact test.

Differential gene expression

Within each granular subset, single-cell data were pseudobulked by summing raw UMI counts across cells separately for each tumor, using the get_pseudobulk function (mode = 'sum', min_cells = 10, min_counts = 1000) from the decoupleR package (v1.8.0) in Python. Genes were filtered to require expression in at least 10% of samples in either the AO-CRC or EO-CRC cohort using filter_by_expr (min_prop = 0.1, min_count = 10).

Differential expression was tested for the MSS tumors within each granular subset using the pseudobulked data. This analysis was performed using the DESeq2 package (v1.42.1) in R, with age scaled to unit variance (z-scored) modeled continuously and sidedness, neoadjuvant treatment history, sex and stage included as covariates (~sidedness + neoadjuvant_treatment + sex + stage + age_scaled). We applied an additional MSS-specific filter so that only genes expressed in at least 1% of EO-CRC or AO-CRC cells within a granular subset, as calculated at the single-cell level using the Seurat v5.1.0 framework in R, were tested for differential expression. Differential expression was not calculated for granular annotation subsets with fewer than three EO-CRC or AO-CRC samples. If there were fewer than three samples in a given covariate category (across both age cohorts) within a granular subset, that covariate was dropped from the design formula. All tested malignant epithelial subsets had sufficient samples to include all covariates.

Genes with FDR < 0.05 were considered differentially expressed, with a negative $\log_2$(FC) indicating decreased expression with increasing age (that is, highest expression in EO-CRC) and a positive $\log_2$(FC) indicating increased expression with increasing age (that is, highest expression in AO-CRC). DEGs along the age continuum were visualized using EnhancedVolcano (v1.20.0).

lncRNA enrichment

We annotated genes in our atlas as lncRNAs using the useEnsembl function from the biomaRt package (v2.58.2) in R. Additionally, we included uncharacterized genes beginning with A and ending with a decimal point in our set of lncRNAs. To determine the enrichment of lncRNA genes within DEG sets from each granular subcluster, we performed a two-sided Fisher's exact test (fisher.test) using the stats package (v4.3.3). P values from granular subsets with more than zero DEGs were adjusted for multiple comparisons using p.adjust (method = 'fdr').

GTEx aging signatures

The GTEx[82] aging signatures for the sigmoid and transverse colon were downloaded from the repository[70] at https://doi.org/10.5281/zenodo.6797627 and filtered for genes with an adjusted P value of <0.05. Genes with positive and negative FC values were labeled as having increased and decreased expression with age, respectively. These genes were intersected with DEGs and gene program lists from our snRNA-seq analyses to ensure that the identified genes were not primarily driven by patient age. DEGs along the age continuum from the GTEx cohort were visualized using EnhancedVolcano (v1.20.0).

Cell–cell interaction

CellChat 2.0 (v2.1.2) was run separately for the EO-CRC ($n = 25$ tumor samples) and AO-CRC ($n = 26$ tumor samples) MSS left-sided tumor cohorts to identify cell–cell interactions. Before running CellChat, cells belonging to granular subsets with 'mixed' annotations were removed. The CellChat analysis was performed at an intermediate annotation level, including granular epithelial cell annotations to facilitate subset-specific analysis and lineage-level annotations for the nonepithelial cells to maximize power. Raw counts were scaled (1×10^4) and log-normalized using NormalizeData from Seurat (v5.1.0). These normalized counts were used to create an age cohort-specific MSS left-sided CellChat object through createCellChat, grouped by the intermediate annotation level with datatype = 'RNA'.

CellChatDB.human was used as the interaction database, including only interactions annotated in the CellChat database and excluding non-protein signaling interactions. Genes were subsetted to those present in the database, and overexpressed genes and interactions were identified. Communication probability between intermediate-level cell types was calculated using computeCommunProb, with the

triMean method and nboot = 1000. Interactions involving fewer than ten cells were filtered using filterCommunication; communication probabilities were aggregated at the pathway level using computeCommunProbPathway; and a communication network was built using aggregateNet.

After identifying the cell–cell interactions separately for the EO-CRC and AO-CRC MSS left-sided tumor samples, the number of interactions for each age cohort was compared using compareInteractions (overall) and netVisual_heatmap (grouped by cell type and split by sender or receiver role). Differences in interaction strength at the pathway level were assessed using rankNet (mode = 'comparison', measure = 'weight', do.stat = TRUE); this function identifies pathways that are significantly different between the age cohorts using a two-sided Wilcoxon (Mann–Whitney *U*) test with a *P*-value cutoff of 0.05. Interaction differences between cell types at the pathway level were visualized using netVisual_aggregate.

Bulk RNA-seq validation data

TCGA-COAD and TCGA-READ bulk RNA-seq data. Clinical and RNA-seq data were obtained for the COAD-US and READ-US projects of TCGA on March 5, 2026. TCGAbiolinks (v2.37.1) was used in R (v3.4.0) to query the NCI Genomic Data Commons (GDC) portal as follows: project = c('TCGA-COAD', 'TCGA-READ'), data.category = 'Transcriptome Profiling', data.type = 'Gene Expression Quantification', experimental.strategy = 'RNA-Seq', workflow.type = 'STAR - Counts' and access = 'open'.

MSI status was obtained directly from the GDC portal's 'MSI status' filter in the Repository tab and mapped to TCGA-COAD and TCGA-READ samples using their submitter IDs. For the 16 samples with discordant status assignments, the status listed for the highest lexicographical barcode (prioritizing D analytes) was retrieved manually.

DFCI EO-CRC bulk RNA-seq data. Formalin-fixed paraffin-embedded samples were collected from the primary tumor resections of patients with EO-CRC who consented to participate in research under Dana-Farber/Harvard Cancer Center Protocols 03-189 and 14-408. RNA was quantified, normalized to 750 ng when sufficient material was available and processed using the KAPA RNA HyperPrep Kit with RiboErase (Human, Mouse, Rat; Roche). Libraries were pooled in equal volumes and sequenced on an Illumina MiSeq Nano flow cell to estimate each library's concentration based on the number of index reads per sample. The 83 sequenceable libraries (yield ≥1 nM) were normalized using the MiSeq-derived concentration, pooled, cleaned using Ampure beads and loaded onto two Illumina NovaSeq 6000 SP-200 flow cells. This RNA-seq procedure was performed by the Dana-Farber Center for Cancer Genomics.

FASTQ files were aligned following the published TCGA methodology to minimize batch effects. STAR (v2.7.5c) was run on Terra using the publicly available GDC.h38.d1.vd1 STAR2 index files (v36) and the DR32 command line parameters listed on the GDC website (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#rna-seq-alignment-command-line-parameters-dr32).

Bulk RNA-seq data processing. Duplicate samples from the same patient were removed from the TCGA cohort, retaining the sample with the highest lexicographical barcode. Within the DFCI EO-CRC cohort, nine samples from patients who were also profiled in the snRNA-seq cohort were excluded. The merged TCGA-COAD, TCGA-READ and DFCI EO-CRC bulk RNA-seq validation dataset was further filtered to include only primary MSS tumor samples with data on neoadjuvant treatment history, primary tumor sidedness, sex, stage and age. The final CRC cohort comprised 564 samples (n = 497 TCGA tumors, n = 67 DFCI EO-CRC tumors). The final adjacent normal cohort comprised 48 samples (n = 7 tumors from patients under 50 years of age and n = 41 tumors from patients aged 50 years and older).

Compositional analysis. Deconvolution of the bulk RNA-seq validation dataset was performed using the MuSiC[34] package (v1.0.0) in R, with the snRNA-seq data and lineage-level cell type annotations as references. The snRNA-seq atlas was first filtered to retain MSS samples with known sidedness and stage metadata. For computational efficiency, we filtered for genes that (1) had an average expression greater than 0.05 counts in at least one cell type, (2) were not mitochondrial or ribosomal genes and (3) were also found in the bulk RNA-seq dataset (final gene count = 21,955). We performed a linear regression analysis using the lm function in R to assess the association of the natural logarithm of the stromal-to-epithelial ratio, modeling age continuously with sidedness, neoadjuvant treatment history, sex, stage and bulk RNA-seq dataset as covariates. Samples with nonfinite log-transformed ratios were excluded from the regression model (n = 493 modeled samples). Proportions were visualized using ggplot2 (v4.0.2).

Differential expression analysis. Differential expression was tested on unstranded counts using the DESeq2 package (v1.42.1) in R (v3.4.0), with age scaled to unit variance (z-scored) and modeled continuously. Neoadjuvant treatment history, sidedness, sex, stage and bulk RNA-seq dataset were included as covariates (~neoadjuvant_treatment + sidedness + sex + stage + dataset + age_scaled). Prefiltering was performed to exclude genes lacking a minimum of five samples with at least ten reads, and independent filtering was set at $\alpha = 0.1$. Genes with an adjusted P value of <0.1 were considered differentially expressed. The overlap of these genes with the union of all DEGs from the snRNA-seq atlas was visualized using ggplot2 (v4.0.3) and ggrepel (v0.9.8). Differential expression in adjacent normal samples from TCGA-COAD and TCGA-READ was similarly assessed, using the following model design formula: ~sidedness + sex + age_scaled.

Multispectral immunofluorescence and image analysis in external validation cohorts

We used the prospective cohort incident-tumor biobank method (PCIBM)[83] on two US prospective cohort studies: the NHS, which enrolled 121,700 women aged 30–55 years in 1976 (ref. 35), and the HPFS, which enrolled 51,529 men aged 40–75 years in 1986 (ref. 36). Under the PCIBM, during decades-long follow-up, these large populations resulted in 4,474 patients with incident CRC for the current study. All 4,474 incident CRC cases were used in the IPW method[37] to adjust for selection bias caused by tissue data availability in the 887 CRC cases that underwent analysis. The study protocol was approved by the institutional review boards of Brigham and Women's Hospital and the Harvard T.H. Chan School of Public Health (protocol no. 2019P003588), as well as the boards of participating registries, as required. Informed consent was obtained from all study participants.

To determine MSI status, DNA was extracted from tumor regions of whole tissue sections using the QIAamp DNA Mini Kit (Qiagen), and a panel of ten microsatellite markers (D2S123, D5S346, D17S250, BAT25, BAT26, BAT40, D18S55, D18S56, D18S67 and D18S487) was assessed[84].

Epithelial and stromal cells in CRC tissues were quantified using a custom multispectral immunofluorescence assay based on tyramide signal amplification. The assay targeted ACTA2 (HGNC:130; actin α2, smooth muscle), CALD1 (HGNC:1441; caldesmon 1), CD34 (HGNC:1662), FAP (HGNC:3590; fibroblast activation protein-α), PDPN (HGNC:29602; podoplanin), KRT (keratin) and DAPI. Staining was performed using the Opal 6-Plex Manual Detection Kit (cat. no. NEL811001KT, Akoya Biosciences; see Supplementary Table 5 for antibody details). Images were acquired with the PhenoImager HT 2.0 (Akoya Biosciences, RRID: SCR_023772) and processed using inForm Tissue Analysis Software v3.0.0 (Akoya Biosciences, RRID: SCR_019155) for autofluorescence correction, spectral unmixing, nuclear segmentation and cell phenotyping.

Cores with nontumorous mucosa, marked necrosis or a limited number of tumor cells were excluded from the analysis. At

the single-cell level, cells were identified by nuclear segmentation based on DAPI signals, followed by cell phenotyping using a machine learning-based algorithm incorporating all markers. The cell phenotyping model was developed to detect and classify cells into the following types: epithelial cells (KRT$^+$), myocytes (CALD1$^+$), vascular endothelial cells (CD34$^+$), fibroblasts (ACTA2$^+$ or FAP$^+$), lymphatic endothelial cells (PDPN$^+$) and other cells (negative for all markers). For subsequent analyses, myocytes, vascular endothelial cells, fibroblasts and lymphatic endothelial cells were collectively defined as stromal cells. The stromal-to-epithelial cell ratio was calculated for each tissue microarray core and averaged across cores for cases with multiple samples to derive a single value for each of the 887 cases.

Statistical analyses were performed using SAS (v9.4; SAS Institute, RRID: SCR_008567). The Spearman correlation test was used to assess the association between age at diagnosis (continuous) and the stromal-to-epithelial cell ratio (continuous). The stromal-to-epithelial cell ratio was categorized into quartiles (Q1 (lowest)–Q4 (highest)), and logistic regression models were constructed to evaluate its association with age, adjusting for sex (female versus male), tumor location (proximal colon versus distal colon versus rectum) and American Joint Committee on Cancer (AJCC) stage (I–II versus III–IV). Cases with missing tumor location (0.4%) were assigned to the majority category, while those with missing AJCC stage (7.0%) were included as a separate category using an indicator variable. To adjust for selection bias due to tissue data availability among the 887 CRC cases included in the analysis, IPW based on all 4,474 CRC cases was incorporated into multivariable logistic regression models[37]. Validation analyses were then restricted to the 712 MSS cases.

Statistics and reproducibility

Sample size. All available EO-CRC specimens of sufficient size from the Young-Onset Colorectal Cancer Center tissue bank at DFCI were included. The study was designed to generate an snRNA-seq atlas with at least 20 samples per age decade whenever possible to enable modeling across the age continuum; however, we were able to collect only one sample from patients in their 20s and six samples from patients in the 90+ age group. No statistical method was used to predetermine the sample size. Only primary CRC tumors were included. The snRNA-seq atlas contained 54 EO-CRC samples (from patients aged <50 years) and 135 AO-CRC samples (from patients aged 50+ years), and the bulk RNA-seq dataset contained 83 EO-CRC samples. In the NHS/HPFS cohort, the distribution was as follows: <50 years, $n = 18$; 50–54 years, $n = 44$; 55–69 years, $n = 417$; and 70+ years, $n = 408$.

Data exclusions. Data exclusion did not occur at the sample level. Empty droplets and ambient RNA counts were removed from the count matrices, and nuclei with too few UMIs or genes, too large a percentage of UMIs mapped to mitochondrial genes, or too many UMIs or genes were excluded (see the Methods for details). DFCI EO-CRC bulk RNA-seq MSI-H samples and samples profiled in the snRNA-seq cohort were excluded from the analysis. For specific downstream analyses, data were subsetted based on clinical parameters (for example, left-sided, MSS) or excluded due to incomplete clinical metadata or insufficient nuclei counts per sample; further details are available in the relevant Methods sections. Among the NHS/HPFS CRC cases, those without available tumor tissue or multispectral immunofluorescence data were excluded (see the Methods for details).

Replication. This study included one tissue specimen per patient. Availability of resection tissue is limited, especially from patients with EO-CRC; therefore, replicate data could not be generated at the patient level. In the snRNA-seq cohort, we had biological replicates (patient tumors) within each age decade as follows: 20–29 years, $n = 1$; 30–39 years, $n = 20$; 40–49 years, $n = 33$; 50–59 years, $n = 30$; 60–69 years, $n = 31$; 70–79 years, $n = 31$; 80–89 years, $n = 37$; and 90+ years, $n = 6$. The DFCI EO-CRC bulk RNA-seq cohort distribution was as follows: 20–29 years, $n = 9$; 30–39 years, $n = 22$; 40–49 years, $n = 50$; and 50–59 years, $n = 2$. In the NHS/HPFS digital histopathology cohort, the distribution was as follows: <50 years, $n = 18$; 50–54 years, $n = 44$; 55–69 years, $n = 417$; and 70+ years, $n = 408$.

Randomization. The experiments were not randomized.

Blinding. Grouping by clinical metadata was central to the analysis design of this study. Investigators were not blinded to allocation during the experiments and outcome assessment.

Reporting summary

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

Data availability

Raw human sequencing data from the snRNA-seq atlas are available in dbGaP (https://dbgap.ncbi.nlm.nih.gov/home), a controlled access repository, under accession number phs004493.v1.p1. Processed counts matrices from Cell Ranger and CellBender, as well as an AnnData object containing the snRNA-seq atlas, are available in the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE312983. Clinical metadata for the samples in the snRNA-seq atlas are available in Supplementary Table 1. For confidentiality reasons, the ages of patients 90 years and older have been aggregated into a single category in the data object and in Supplementary Table 1. Requests to access these age data should be directed to K.N. (Kimmie_Ng@DFCI.HARVARD.EDU) and can typically be granted within 3 months after Steering Committee approval. Raw human sequencing data for the DFCI EO-CRC bulk RNA-seq atlas are available in dbGaP under accession number phs004807.v1. Associated processed data are available in GEO under accession number GSE333907. Clinical metadata for the samples from both the DFCI EO-CRC bulk RNA-seq cohort and the TCGA cohort used in this study are available in Supplementary Table 4. The NHS/HPFS digital histopathology data used for multispectral immunofluorescence validation cannot be shared publicly due to participant confidentiality and privacy concerns. Requests to access NHS/HPFS data must be submitted in writing. According to standard controlled access procedures, applications to use NHS/HPFS resources will be reviewed by our External Collaborations Committee to verify that the proposed use maintains the privacy of participants and the confidentiality of the data. Investigators wishing to use NHS/HPFS data are asked to submit a brief description of their proposed project. Refer to https://nurseshealthstudy.org/researchers (contact email: nhsaccess@channing.harvard.edu), https://gutsweb.org/collaborate-with-guts/ and https://hsph.harvard.edu/research/health-professionals/resources/for-external-collaborators/ for details. Once proposals are approved, external collaborators should seek guidance from the institutional review boards at their home institutions before conducting secondary analyses of NHS/HPFS data. Published data used in this manuscript include TCGA-COAD and TCGA-READ, which were downloaded from the NCI Genomic Data Commons (GDC) portal using TCGAbiolinks (v2.37.1) on March 5, 2026. MSI status was obtained directly from the GDC portal, and copy number alterations were obtained from cBioPortal (https://www.cbioportal.org/study/summary?id=coadread_tcga_pan_can_atlas_2018). The single-cell RNA-seq CRC atlas by Pelka et al.[26] (https://singlecell.broadinstitute.org/single_cell/study/SCP1162/human-colon-cancer-atlas-c295) and the single-cell RNA-seq Crohn's disease atlas by Kong et al.[27] (https://singlecell.broadinstitute.org/single_cell/study/SCP1884/human-cd-atlas-study-between-colon-and-terminal-ileum) were downloaded from the Broad Institute Single Cell Portal. The snRNA-seq atlas by Becker et al.[28] was downloaded from GitHub (https://github.com/winstonbecker/scCRC_continuum).

Code availability

All analysis code used in this article is available on GitHub at https://github.com/klarman-cell-observatory/EO-CRC.

References

71. Chafamo, D. et al. kco/bcl_convert. *Terra* https://app.terra.bio/#workflows/kco/bcl_convert/11 (2022).
72. Li, B. et al. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat. Methods* **17**, 793–798 (2020).
73. Cumulus Team. cumulus/Cellranger:2.3.0. *Dockstore* https://dockstore.org/workflows/github.com/lilab-bcb/cumulus/Cellranger:2.3.0 (2022).
74. Terra Community Workbench. *Terra* https://app.terra.bio (2026).
75. Fleming, S. J. et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* **20**, 1323–1335 (2023).
76. Fleming, S. cellbender/remove-background. *Terra* https://app.terra.bio/#workflows/cellbender/remove-background/11 (2023).
77. Cumulus Team. cumulus/Cumulus:2.3.0. *Dockstore* https://dockstore.org/workflows/github.com/lilab-bcb/cumulus/Cumulus:2.3.0 (2022).
78. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
79. Xu, C. et al. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell* **186**, 5876–5891 (2023).
80. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
81. Georges, C. *CTAT/infercnv_wf*. https://app.terra.bio/#workflows/CTAT/infercnv_wf/15 (2023).
82. Jia, K., Cui, C., Gao, Y., Zhou, Y. & Cui, Q. An analysis of aging-related genes derived from the Genotype–Tissue Expression project (GTEx). *Cell Death Discov.* **4**, 26 (2018).
83. Ogino, S., Patel, A. V., Ugai, S., Giannakis, M. & Ugai, T. Unique role of the prospective cohort incident-tumor biobank method in etiological research. *Lancet Reg. Health Am.* **53**, 101290 (2026).
84. Yamauchi, M. et al. Assessment of colorectal cancer molecular features along bowel subsites challenges the conception of distinct dichotomy of proximal versus distal colorectum. *Gut* **61**, 847–854 (2012).

Acknowledgements

We thank E. Chen for assisting with data management and data preprocessing, B. Shao for contributing to early cell–cell interaction analyses, H. S. Hwang for helping with early exploratory analyses, L. Kong and M. Kanai for contributing to discussions regarding compositional analysis, C. McCabe for providing project management and wet-laboratory support, M. Glean for contributing to experimental processing and K. A. Tooley for providing project management support.

Author contributions

C.B.M.P., K.N., A.C.A. and R.J.X. conceived the study. T.M.D., C.B.M.P., K.N., A.C.A. and R.J.X. designed the experiments and interpreted the results with assistance from J.D., J.S.T., M.G. and O.A. S.P.M., L.A.-Z. and E.F. performed experiments. C.B.M.P., O.A. and A.C.A designed the computational analyses. C.B.M.P., L.K. and J.C. performed the computational analyses. E.A., W.T., E.S., L.K.B., K.M., A.K., S.U., T.U., S.O., M.G. and K.N. collected and analyzed human clinical data. The manuscript was written by C.B.M.P., A.C.A. and K.N., and was edited with input and final approval from all authors.

Funding

This work was supported by the Gene Lay Institute of Immunology and Inflammation; the Helen Gurley Brown Foundation/Dana-Farber Cancer Institute; Project P Fund; Patrick Beauregard Fund; Poorvu-Jaffe Fund for Young-Onset Colorectal Cancer Research; the Thau Fund; the Eigner Fund for Young-Onset Colorectal Cancer; the Klarman Cell Observatory at the Broad Institute; the National Institutes of Health and National Cancer Institute (NIH/NCI R01 CA205406 to K.N.; NIH P30 DK043351 to R.J.X.; R01 CA285851 to M.G.; P01 CA87969, UM1 CA186107, P01 CA55075, UM1 CA167552, U01 CA167552, R01 CA151993 to S.O.; R35 CA197735 to S.O.); and the American Cancer Society Clinical Research Professor Award (CRP-24-1185864-01-PROF to S.O.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

A.C.A. is a member of the scientific advisory board for Compass Therapeutics, Zumutor Biologics and Cytospire. The interests of A.C.A. were reviewed and managed by Brigham and Women's Hospital and Mass General Brigham in accordance with the institutions' conflict-of-interest policies. S.O. served as a consultant for Sanofi Pasteur. R.J.X. is a cofounder of Convergence Bio, a scientific advisory board member at Nestlé and Magnet BioMedicine, and a board director at MoonLake Immunotherapeutics; these organizations had no roles in this study. J.D. is a member of BioRender's scientific advisory board. M.G. receives institutional research funding from Janssen and Sunbird Bio and has consulted for Nerviano Medical Sciences. M.G. has received honoraria from OncLive and PER. K.N. consults for or serves on the advisory board of Bayer, Pfizer, CytomX, Jazz Pharmaceuticals, Revolution Medicines, Agenus, Johnson & Johnson, AbbVie, Sanofi, GlaxoSmithKline, Flagship Pioneering, AstraZeneca, Amgen, CRICO and Manta Cares. K.N. receives institutional research funding from Pharmavite and Janssen and serves as an associate editor at *JAMA*. The interests of K.N. were reviewed and managed by the Dana-Farber Cancer Institute in accordance with its conflict-of-interest policies. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-026-04721-y.

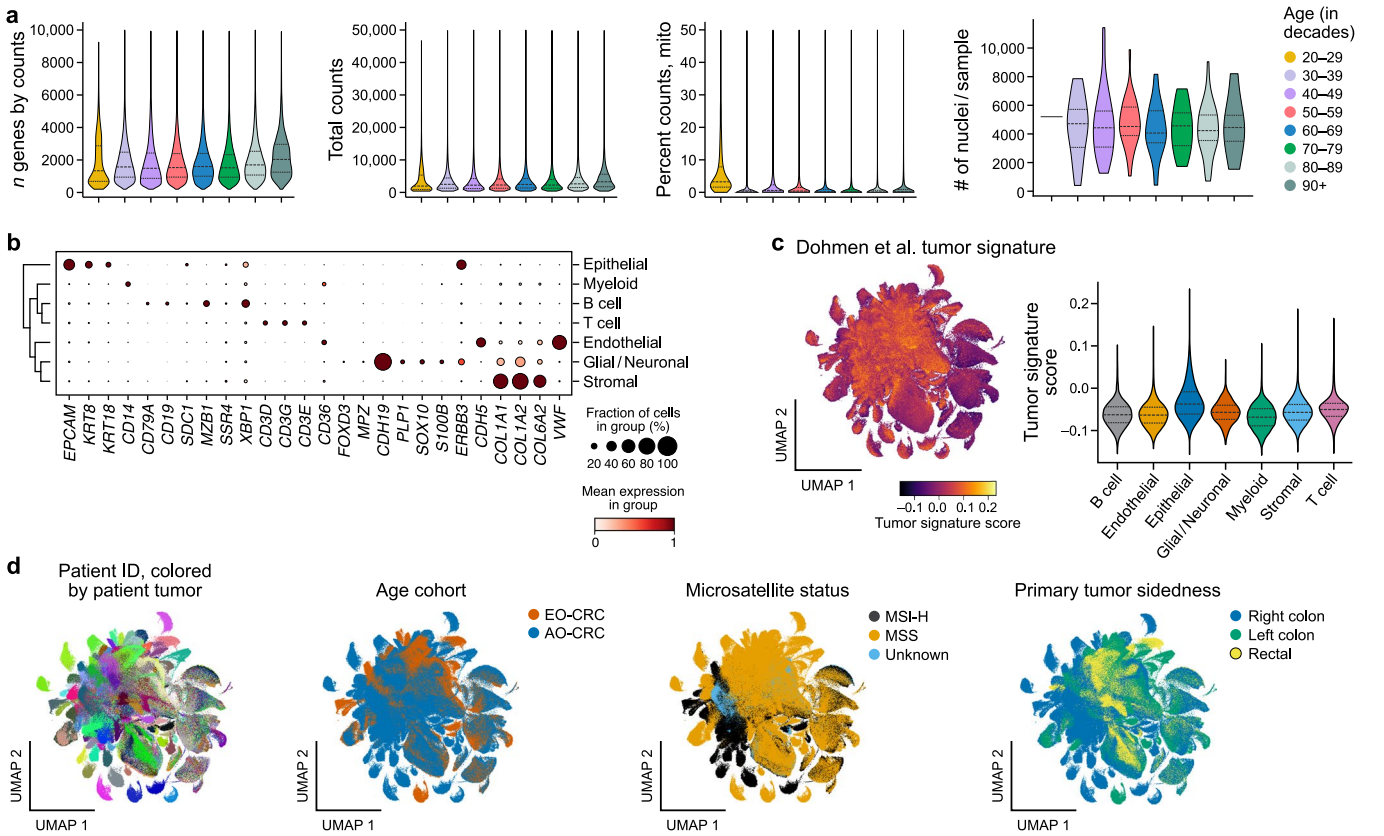
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-026-04721-y.

Correspondence and requests for materials should be addressed to Ramnik J. Xavier, Ana C. Anderson or Kimmie Ng.

Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Anna Ranzoni, in collaboration with the *Nature Medicine* team.

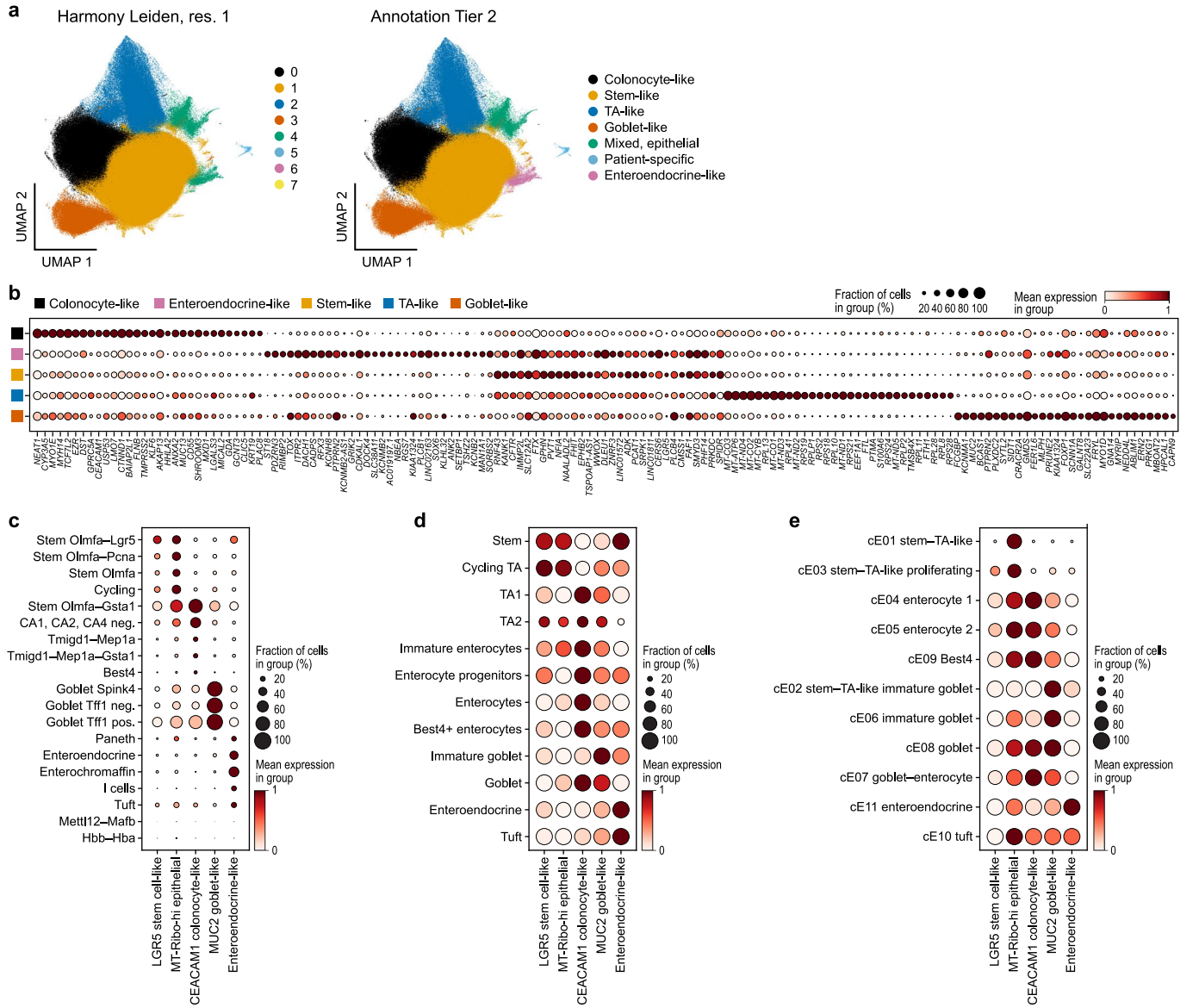
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41591-026-04721-y



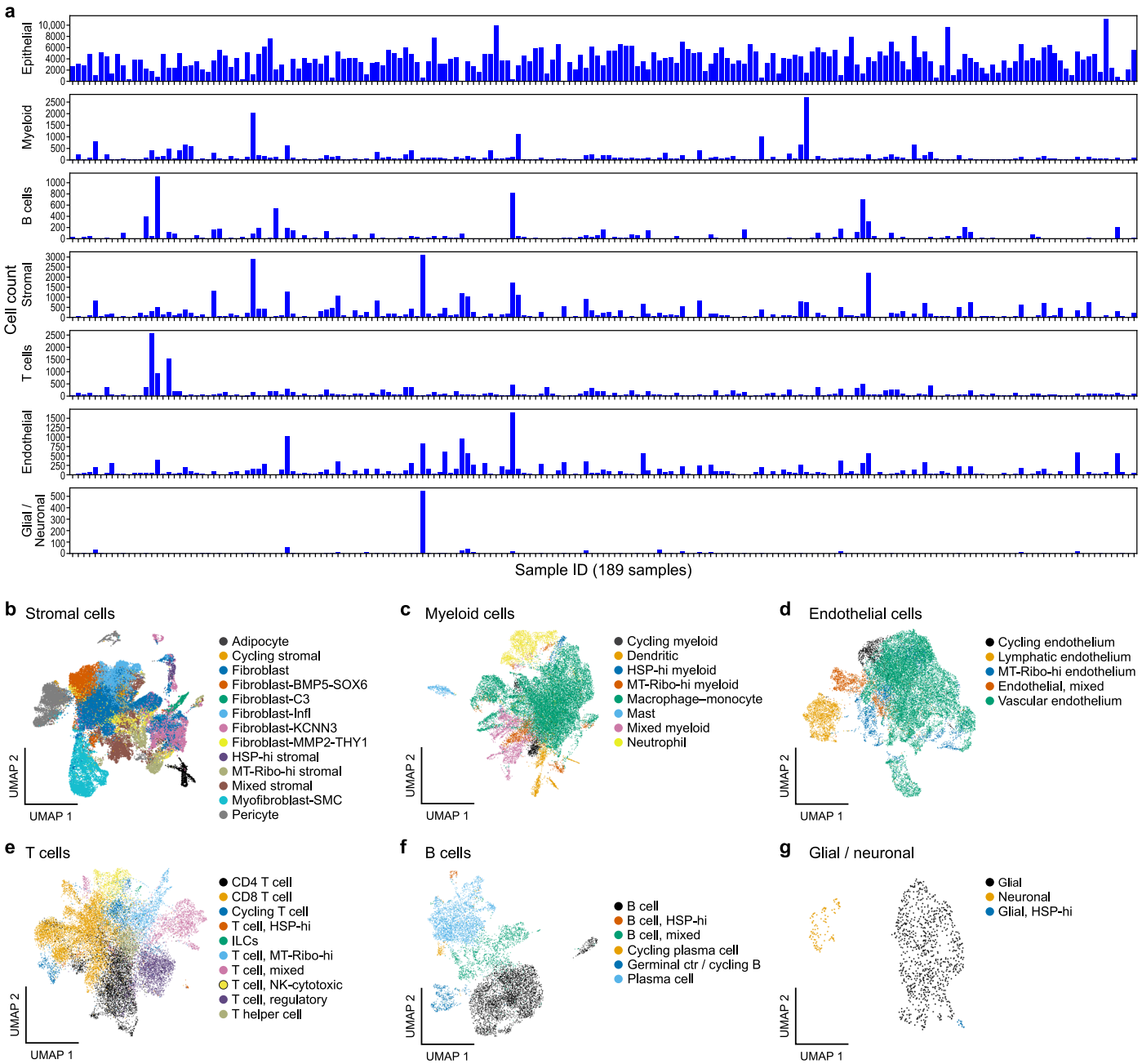
Extended Data Fig. 1 | scRNA-Seq CRC atlas quality control, annotation, and clinical features. a. Violin plots showing nuclei quality distribution (lines: median, 1st and 3rd quartiles) for nuclei passing quality control colored by age decade for number of genes per nucleus (left), total counts per nucleus (middle-left), the percentage of UMI mapping to mitochondrial genes per nucleus (middle-right), and number of nuclei per sample (right). **b.** Dot plot showing mean and percent expression for select marker genes, grouped by lineage annotation. **c.** UMAP embedding (no batch correction) of all nuclei colored by a tumor signature score[16] (left) and violin plot showing the score distribution (lines: median, 1st and 3rd quartiles) by lineage annotation (right). **d.** Same UMAP as in (**c**), colored by sample ID, age cohort, microsatellite status, and tumor location; each dot represents one nucleus. *n = 189 tumors, 850,222 nuclei* for all panels.

https://doi.org/10.1038/s41591-026-04721-y



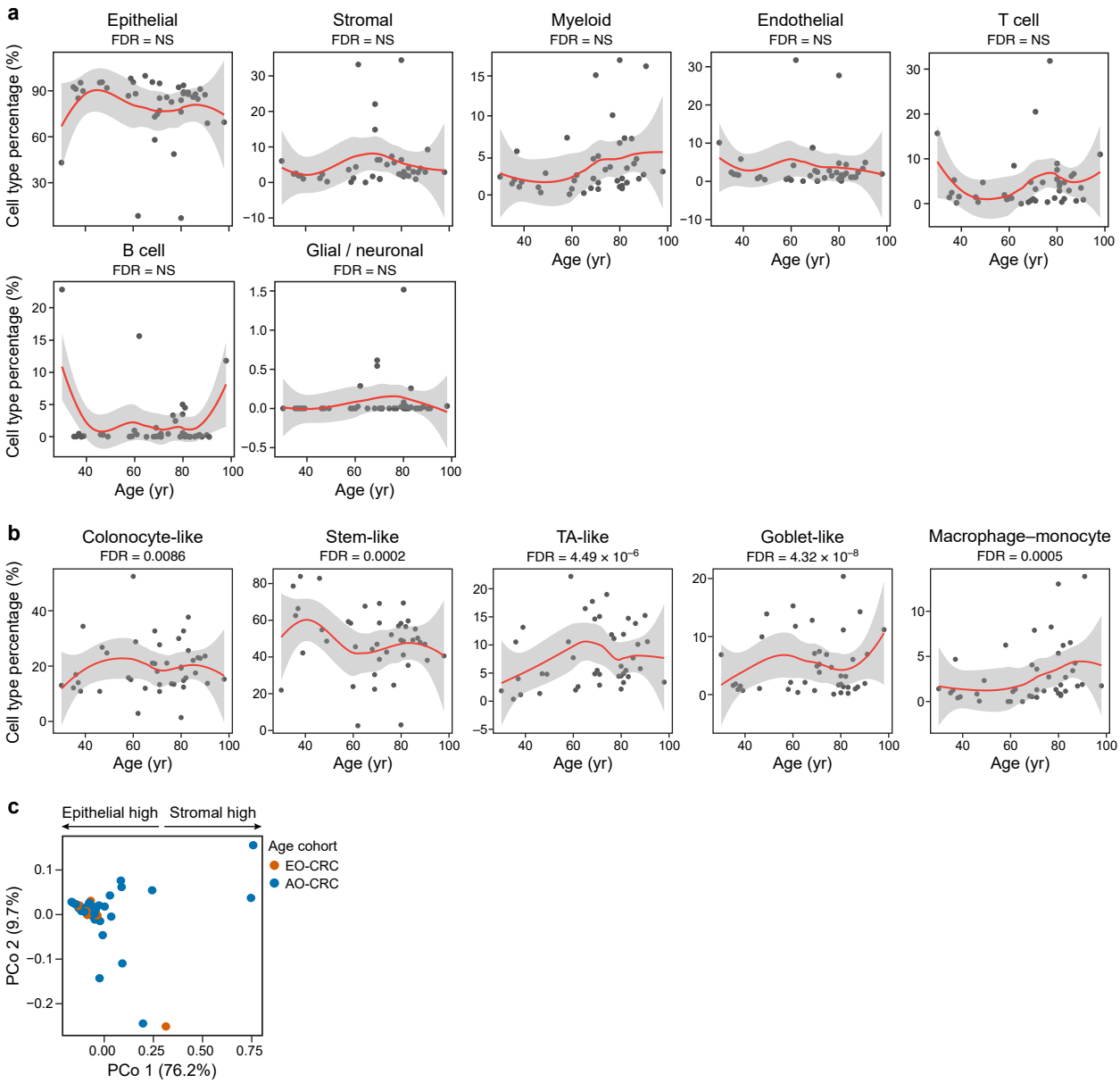
Extended Data Fig. 2 | Epithelial subset annotation, clinical features, and iCMS scores. a. UMAP embedding of malignant epithelial cells following patient batch correction using Harmony, colored by Leiden cluster resolution 1.0 (left) and final annotation (right) (*n = 189 tumors, 722,091 nuclei*). **b.** Dot plot showing mean and percent expression for the top 30 calculated granular cell type marker genes (Methods) for the five major epithelial subsets. **c-e.** Dot plots showing the mean and percent expression for gene signatures from Kong et al.[27] **(c)**, Becker et al.[28] **(d)**, and Pelka et al.[26] **(e)**, grouped by granular epithelial cell type from our atlas. Mixed-epithelial and the patient-specific cluster are not shown. **f.** UMAP embedding of epithelial cells (no batch correction, distinct from (**a**)), colored by sample ID and tumor location; each dot represents one nucleus (*n = 189 tumors, 722,091 nuclei*). **g.** Same UMAP as in (**f**), colored by iCMS2 and iCMS3 gene signature[24] score (*yellow = high score, purple = low score*).

https://doi.org/10.1038/s41591-026-04721-y



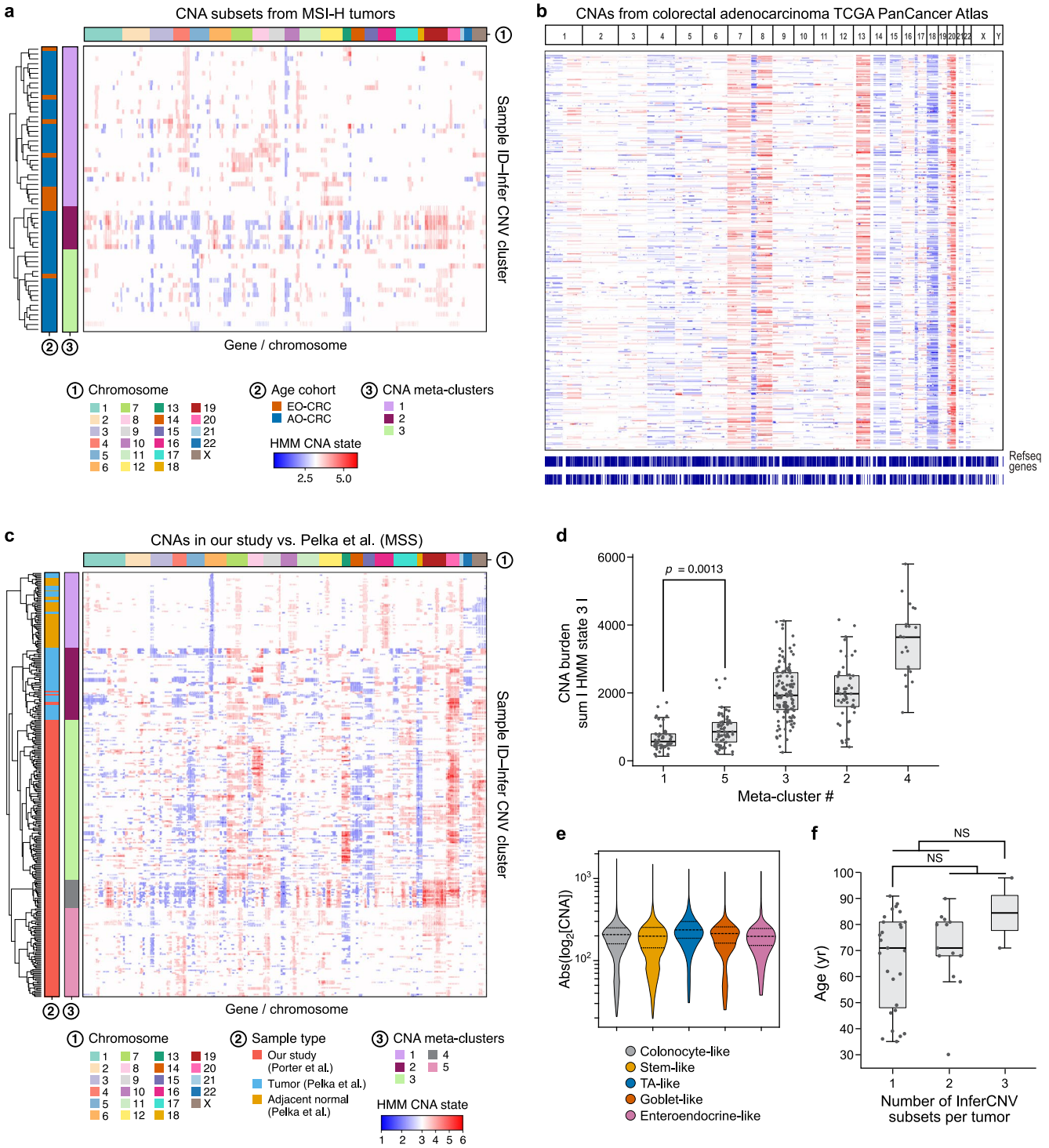
Extended Data Fig. 3 | Cell type distribution and non-epithelial lineage annotations. a. Bar plot showing the number of nuclei (y axis) per sample (x axis) for each lineage annotation (*n = 189 tumors*). **b-g**. UMAP embeddings (no batch correction) colored by granular subset annotations for Stromal (**b**; *n = 185 tumors, 47,712 nuclei*), Myeloid (**c**; *n = 187 tumors, 26,248 nuclei*), Endothelial (**d**; *n = 185 tumors, 22,705 nuclei*), T (**e**; *n = 185 tumors, 21,497 nuclei*), B (**f**; *n = 153 tumors, 8,990 nuclei*), and Glial/Neuronal (**g**; *n = 50 tumors, 979 nuclei*) cells; each dot represents a single nucleus. In all lineages except Glial/Neuronal, Harmony batch correction was used to aid annotation (Methods).

https://doi.org/10.1038/s41591-026-04721-y



Extended Data Fig. 4 | Compositional analysis of MSI-H tumor samples. **a,b**. Scatter plots showing cell type percentage per MSI-H tumor (n = 43 MSI-H tumors: 9 EO-CRC, 34 AO-CRC) for all seven cell type lineages **(a)** and granular cell types **(b)**; red line indicates fitted values from LOESS, with shaded regions showing the 95% confidence interval; FDRs from Dirichlet regression analysis modeling age as a continuous variable with sidedness, neoadjuvant treatment history, sex, and stage as covariates. **c**. Principal coordinates analysis (PCoA) of lineage level compositional profiles of MSI-H tumor samples, colored by age cohort (Bray–Curtis; Methods); arrows indicate cell lineages with greatest spearman correlation with PCo1.

https://doi.org/10.1038/s41591-026-04721-y

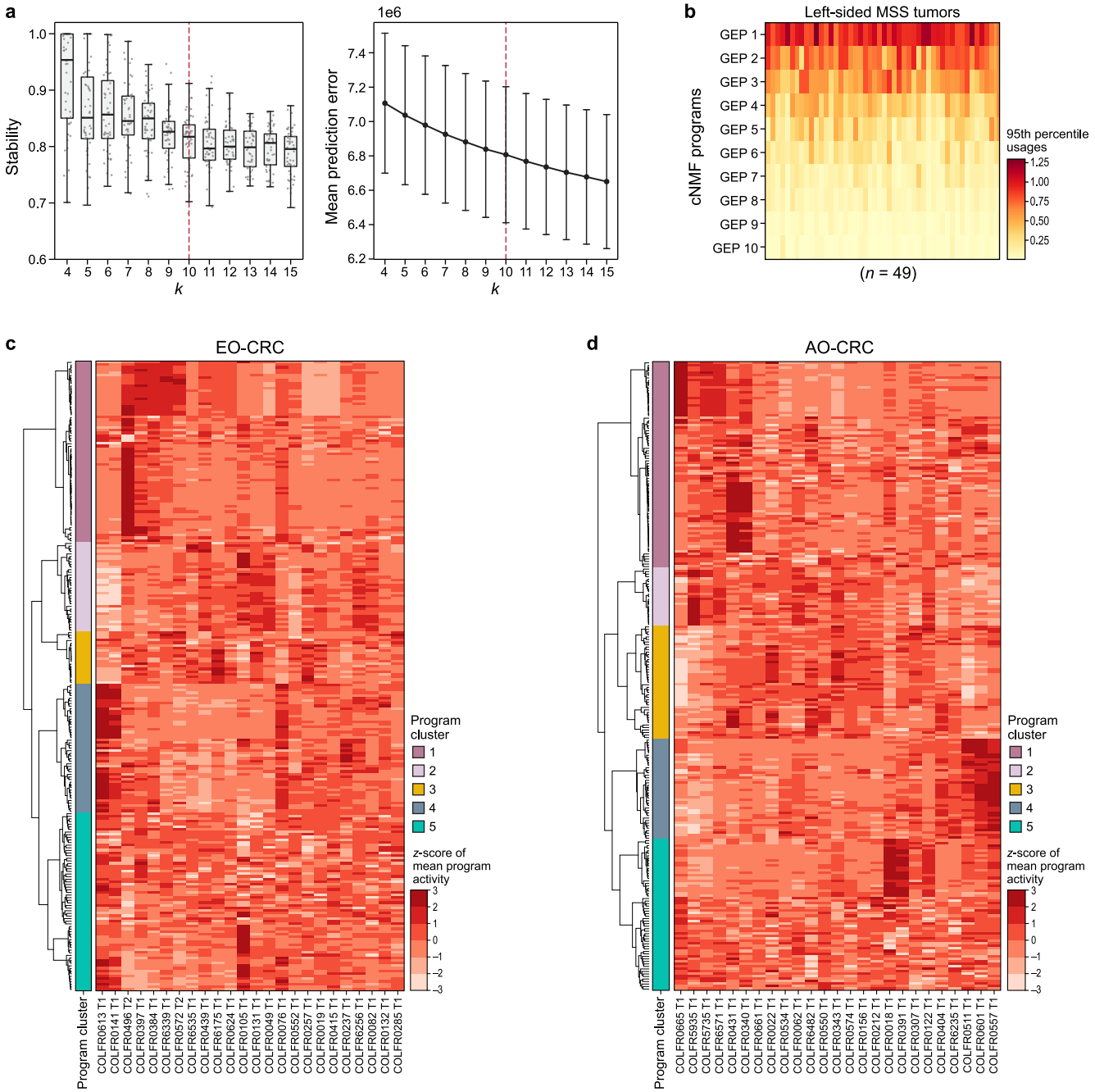


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41591-026-04721-y

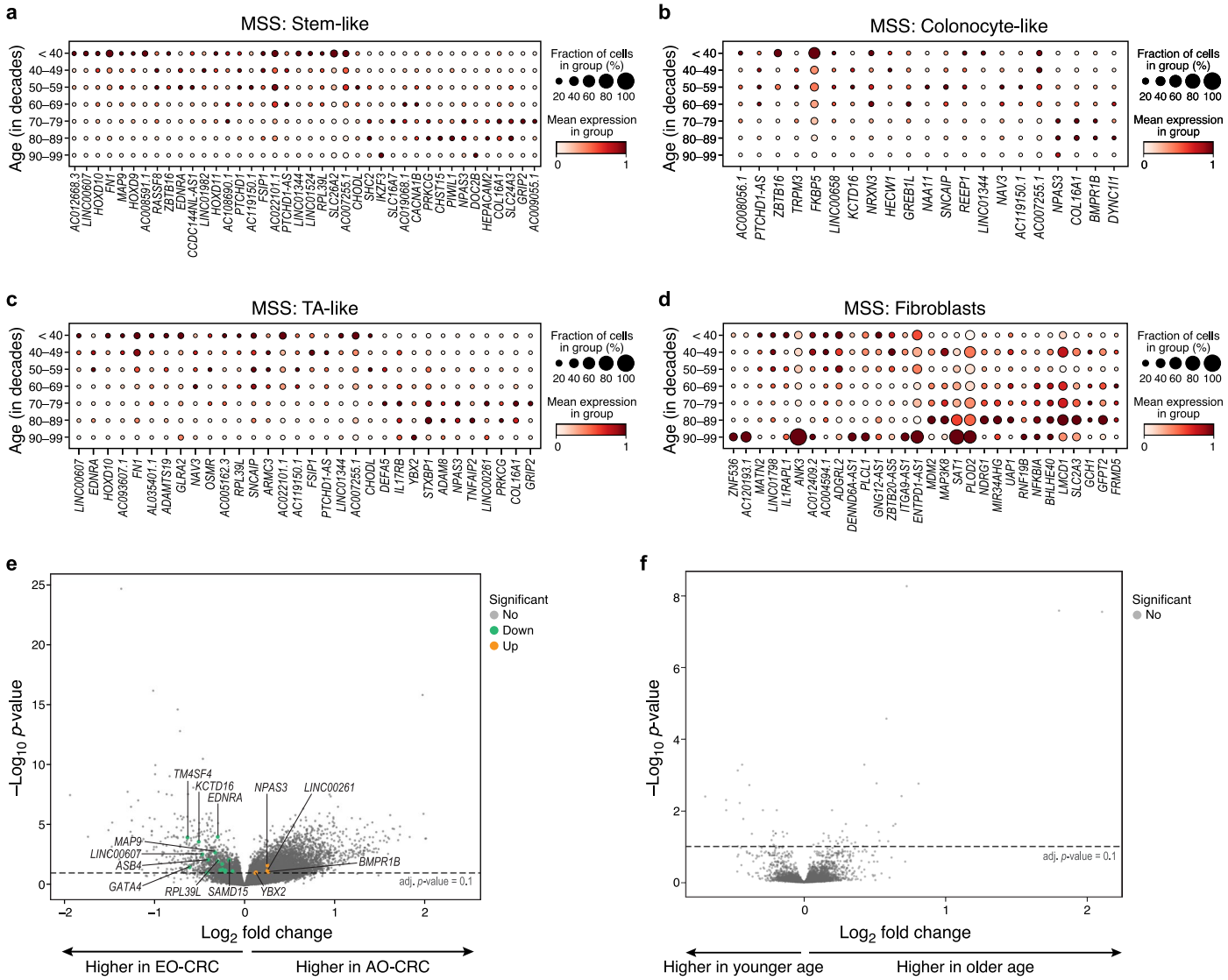
Extended Data Fig. 5 | CNA analysis of MSI-H tumor samples and comparison to Pelka et al.[26] normal. a. Heatmap of epithelial copy number alteration (CNA) for MSI-H samples for which CNVs could be predicted by inferCNV (*n = 9 EO-CRC tumors and 10 CNA subsets; 33 AO-CRC tumors and 49 CNA subsets*); each row corresponds to a tumor-level inferCNV subset (1-3 subsets per tumor, Methods) and each column corresponds to a gene, ordered by chromosomal loci. Colors correspond to HMM-derived CNA states (*red=duplication, values 4-6; blue=deletions, values 1-2; white=neutral/no call, value 3*). The top annotation bar (1): chromosome; left annotation bar (2): age cohort (EO-CRC=orange; AO-CRC=blue); right annotation bar (3): CNA meta-cluster. ChrY and ChrM are not shown. **b**. CNA heatmap from CBioPortal for colorectal adenocarcinoma TCGA PanCancer Atlas MSS samples (MANTIS score <0.4). **c**. CNA heatmap comparing CNAs from MSS tumors in our study to MSS tumors and adjacent normal in Pelka et al.[26] The top annotation bar (1): chromosome; left annotation bar (2): study and sample type (red = our study, blue = Pelka tumor, orange = Pelka adjacent normal); right annotation bar (3): CNA meta-cluster. ChrY and ChrM are not shown (*n = 41 total subsets across 36 adjacent normal samples and 64 subsets across 28 tumor samples from Pelka* et al.*; n = 210 total subsets across 133 tumors from our atlas*). **d**. Box plot (box: interquartile range; line: median; whiskers: 1.5x interquartile range) showing the CNA burden per CNA meta-cluster as in (**c**), calculated by summing the absolute value of the HMM state minus three (three indicates no duplication or deletion); significance between Pelka et al.[26] adjacent normal (Cluster 1, *n = 56 subsets*) and the lower CNA cluster from our atlas (Cluster 5, *n = 66 subsets*) was determined by Mann-Whitney U test. **e**. Violin plot showing the distribution (lines: median, 1st and 3rd quartiles) of CNA magnitude $|\log_2(\text{CNA})|$ by epithelial cell type in MSI-H tumors (as in (**a**)). **f**. Box plot (box: interquartile range; line: median; whiskers: 1.5x interquartile range) showing the number of CNA subsets (1-3) per tumor by age as in (**a**); significance from two-sided multivariable logistic modeling age association with >=2 (n.s.) or >=3 (n.s.) CNA subsets accounting for sidedness, neoadjuvant treatment history, sex, stage, mean UMI count per nucleus, and epithelial nuclei count per tumor; no adjustment for multiple comparisons was made; each dot represents one MSS tumor.

https://doi.org/10.1038/s41591-026-04721-y



Extended Data Fig. 6 | Determination of consensus MSS left-sided gene metaprograms by cNMF. a. Box plot (left) showing the distribution (box: interquartile range; line: median; whiskers: 1.5x interquartile range) of stability scores and line graph (right) showing the mean prediction error (dots) with standard error of the mean (error bars) for the 12 tested k values when running cNMF once per left-sided MSS tumor sample; red dashed line indicates final selected k = 10 value across all samples (*n = 49 tumor samples, 100 cNMF iterations per sample*). **b.** Heatmap showing the 95th percentile usage score from programs 1-10 across all left-sided MSS tumors (*n = 49 tumors*; rows = programs, columns = samples). **c,d.** Heatmap showing cNMF gene metaprograms (Methods) within each MSS left-sided sample with greater than 100 epithelial cells for EO-CRC (**c**, *n = 24*) and AO-CRC (**d**, *n = 25*), with 10 tumor-specific gene programs per sample scored across all samples in the age cohort. Rows: tumor-specific gene programs. Columns: tumor samples. Values: z-score of mean program activity across samples. Annotation bar: meta-gene program cluster assignments within a given age cohort.

https://doi.org/10.1038/s41591-026-04721-y



Extended Data Fig. 7 | Differential gene expression of select cell subsets and comparison to bulk RNA-Seq cohorts. a-d. Dot plots showing mean expression and percent expression across MSS tumor cells within each decade for selected significant DEGs (*FDR < 0.05*) from Stem-like (**a**), Colonocyte-like (**b**), TA-like (**c**), Fibroblasts (**d**) differential expression tests as in Fig. 5a–d; significance was determined using a two-sided DESeq2 Wald test with Benjamini-Hochberg correction, with age modeled as a continuous variable and sidedness, neoadjuvant treatment history, sex, and stage included as covariates. **e.** Volcano plot as in Fig. 5e with full range of DEGs shown; genes overlapping with DEGs from our snRNA-Seq atlas are labeled (green: higher in EO-CRC; orange: higher in AO-CRC; dashed line indicates FDR 0.1; n = 564 MSS tumors: 126 EO-CRC, 438 AO-CRC). Differential expression was determined using two-sided DESeq2 Wald test with age modeled continuously and sidedness, neoadjuvant treatment history, sex, stage, and bulk RNA-Seq dataset as covariates; multiple comparisons were corrected using Benjamini-Hochberg. **f.** Volcano plot showing DEGs associated with age from TCGA-COAD/READ adjacent normal tissue; differential expression determined using two-sided DESeq2 Wald test with age as a continuous variable and sex and sidedness as covariates (dashed line indicates FDR 0.1, no genes reached significance; n = 48 samples: 7 with age <50 years, 41 with age 50+ years); multiple comparisons were corrected using Benjamini-Hochberg.

nature portfolio

Corresponding author(s): Ramnik J. Xavier, Ana C. Anderson, Kimmie Ng

Last updated by author(s): Sept 9, 2026

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Software was not used for data collection.
Data analysis	SciPy (v1.14.1) statsmodels (v0.14.2) bclconvert (v3.10.5) Cellranger (v7.0.1) Cumulus (v2.3.0) Terra Cellbender (v0.2.0) ScanPy (v1.10.2) harmonypy (v0.0.10) CellTypist (v1.6.3) CMScaller (v2.0.1) decoupler (v1.8.0) biomaRt (v2.58.2) DirichReg (v0.7.1) pertpy (v1.0.5) labsdsv(v2.1.0) inferCNV (v1.19.1) Seaborn (v0.13.2) lme4 (v1.1.38)

cnmf (v1.6.0)
DESeq2 package (v1.42.1)
Seurat (v5.1.0)
stats (v4.3.3)
CellChat 2.0 (v2.1.2)
clusterProfiler (v4.10.1)
MSigDB (msigdbr, v25.1.0)
EnhancedVolcano (v1.20.0)
TCGAbiolinks (v2.37.1)
STAR (v2.7.5c)
MuSiC (v1.0.0)
SAS (version 9.4)
Additional R and python packages for manipulation and visualization

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

RNA-Seq atlas are available in dbGaP (dbgap.ncbi.nlm.nih.gov/home), a controlled access repository, under accession number phs004493.v1.p1. Processed counts matrices from cellranger and cellbender, as well as an anndata object containing the snRNA-Seq atlas, are available in the Gene Expression Omnibus (GEO) (ncbi.nlm.nih.gov/geo/) under accession number GSE312983. Clinical metadata for the samples in the snRNA-Seq atlas available in Supplementary Table 1. Note that for confidentiality reasons, the ages for patients ninety years and older have been aggregated into a single category in the data object and in Supplementary Table 1. Requests to access these age data should be directed to Kimmie Ng: Kimmie_Ng@DFCI.HARVARD.EDU, and can typically be granted within three months after Steering Committee approval. Raw human sequencing data for the DFCI EO-CRC bulk RNA-Seq atlas are available in dbGaP under accession number phs004807.v1. Associated processed data are available in GEO under accession number GSE333907. Clinical metadata for the samples from both the DFCI EO-CRC bulk RNA-Seq cohort and the TCGA cohort used in this study are available in Supplementary Table 4.

NHS/HPFS multispectral immunofluorescence validation data: Because of participant confidentiality and privacy concerns, the digital histopathology data cannot be shared publicly and requests to access NHS/HPFS data must be submitted in writing. According to standard controlled access procedures, applications to use NHS/HPFS resources will be reviewed by our External Collaborations Committee to verify that the proposed use maintains the protection of the privacy of participants and the confidentiality of the data. Investigators wishing to use NHS/HPFS data are asked to submit a brief description of the proposed project. Please refer to nurseshealthstudy.org/researchers (contact email: nhsaccess@channing.harvard.edu), gutsweb.org/collaborate-with-guts/, and hsph.harvard.edu/research/health-professionals/resources/for-external-collaborators/ for details. Once proposals are approved, external collaborators should seek Institutional Review Board guidance from their home institution for conducting secondary analyses of NHS/HPFS.

Published data: Published data used in this manuscript include TCGA-COAD and TCGA-READ, which were downloaded from the NCI Genomic Data Commons (GDC) portal using TCGAbiolinks (v2.37.1) on 03/05/2026; MSI status was obtained directly from the GDC portal and copy number alterations were obtained from CBioPortal: cbioportal.org/study/summary?id=coadread_tcga_pan_can_atlas_2018. The scRNA-Seq CRC atlas by Pelka et al. and the scRNA-Seq Crohn's Disease atlas by Kong et al. were downloaded from the Broad Institute Single Cell Portal:
Pelka et al.: singlecell.broadinstitute.org/single_cell/study/SCP1162
Kong et al.: singlecell.broadinstitute.org/single_cell/study/SCP1884/)
The snRNA-Seq atlas by Becker et al. was downloaded from Github: github.com/winstonbecker/scCRC_continuum.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	For the snRNA-Seq and bulk RNA-Seq cohorts, patient sex was collected through clinical records and as all available tissue was processed, sex was not considered in the study design. Sex, age, and other clinical metadata are reported in a disaggregated manner in Supplementary Table 1 (snRNA-Seq) and Supplementary Table 4 (bulk RNA-Seq). The age of patients 90 years and older has been aggregated into a single "90+" category. For the NHS/HPFS multispectoral immunofluorescence cohort, sex was self-reported and was only considered in the study design by nature of the cohort compositions (NHS is a cohort of females and HPFS is a cohort of males). Aggregated clinical metadata are reported in Supplementary Table 5.
Reporting on race, ethnicity, or other socially relevant groupings	For the snRNA-Seq and bulk RNA-Seq cohorts, subjects' race and ethnicity was collected through clinical records. Race and ethnicity are reported in Supplementary Table 1 and Supplementary Table 4 but were not used in the analyses.
Population characteristics	Subjects' age at collection ranged from 20s to 90+. All subjects were diagnosed with colorectal cancer. Associated clinical metadata are provided in Supplementary Table 1, Supplementary Table 4, and Supplementary Table 5. For confidentiality reasons, the ages for patients ninety years and older belonging to the snRNA-Seq cohort have been aggregated into a single category in the data object and in Supplementary Table 1. Similarly, patient metadata for the NHS/HPFS cohort have been aggregated in Supplementary Table 5. Details for obtaining access to the aggregated are in the Data Availability statement.

Recruitment	scRNA-Seq and bulk RNA-Seq cohorts: Every patient seen in the Gastrointestinal Cancer Center at DFCI is approached at their first visit, introduced to the DFCI Gastrointestinal Biobank by a Research Data Specialist, and consented to participate if they are willing and agreeable. Selection bias associated with DFCI being a tertiary care, academic cancer center should be noted, although the overall population of patients with cancer seen at DFCI is largely representative of the general New England population. Because academic cancer centers are often enriched in younger patients, this is actually a strength for the primary objective of this study, which is focused on characterizing the biology of young-onset colorectal cancer. Patients consented to participate in research on Dana-Farber/Harvard Cancer Center (DF/HCC) Protocols 03-189, 11-104, 14-091, 14-408, 17-000, and/or 20-000. For the NHS/HPFS cohort, patients were recruited from two U.S.-wide prospective cohort studies, namely the Nurses' Health Study and the Health Professionals Follow-Up Study, using the prospective cohort incident-tumor biobank method (PCIBM). Under the PCIBM, incident colorectal cancer cases were identified among the cohort participants and tissue specimens were collected from hundreds of hospitals for tissue-based studies. The study protocols were approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health (IRB protocol number: 2019P003588), and those of participating registries as required.
Ethics oversight	The protocols used to generate our snRNA-Seq and bulk RNA-Seq data were approved by the Dana-Farber/Harvard Cancer Center (DF/HCC) Institutional Review Board. Protocols related to the NHS/HPFS multispectoral immunofluorescence validation cohort were approved by the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health review boards.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	All available EO-CRC specimens of sufficient size from the DFCI Gastrointestinal Biobank. The study was designed to generate a snRNA-Seq atlas with at least 20 samples per age decade whenever possible to enable modeling across the age continuum, though we could only collect one sample from the 20s decade and six from the 90+ decade. No statistical method was used to predetermine sample size. Only primary CRC tumors were included. The snRNA-Seq atlas contained 54 EO-CRC (<50 years) and 135 AO-CRC (50+ years), and the bulk RNA-Seq dataset contained 83 EO-CRC samples. In the bulk RNA-Seq cohort, two samples were collected at age 50, despite an EO-CRC diagnosis. In the NHS/HPFS validation cohort distribution is: <50: 18, 50-54: 44 55-69: 417 70+: 408.
Data exclusions	Data exclusion did not occur at a sample level. Empty droplets and ambient RNA counts were removed from the count matrices, and nuclei with too few UMIs/genes, too large a percentage of UMIs mapped to mitochondrial genes, or too many UMIs/genes were excluded (see Methods for details). DFCI EO-CRC bulk RNA-Seq MSI-H samples and samples profiled in the snRNA-Seq cohort were excluded from analysis. For specific downstream analyses, data were subset based on clinical parameters (e.g., left-sided, MSS) or excluded due to incomplete clinical metadata or insufficient nuclei counts per sample; further details are available in the relevant Methods sections. Among the NHS/HPFS CRC cases, those without available tumor tissue or multispectral immunofluorescence data were excluded (see Methods for details).
Replication	This study contained one tissue specimen per patient. Availability of resection tissue is limited, especially from the EO-CRC patient samples, and thus data could not be generated in replicate on a patient level. In the snRNA-Seq cohort, we had biological replicates (patient tumors) within each decade as follows: 20s: 1, 30s: 20, 40s: 33, 50s: 30, 60s: 31, 70s: 31, 80s: 37, 90+: 6. The DFCI EO-CRC bulk RNA-Seq cohort distribution is: 20s: 9, 30s: 22, 40s: 50, 50s: 2. In the NHS/HPFS validation cohort distribution is: <50: 18, 50-54: 44 55-69: 417 70+: 408.
Randomization	The experiments were not randomized. .
Blinding	Grouping by clinical metadata was central to the analysis design of this study; investigators were not blinded to allocation during experiments and outcome assessment.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used: Please refer to Supplementary Table 5

Validation: Please refer to Supplementary Table 5

Plants

Seed stocks: NA

Novel plant genotypes: NA

Authentication: NA