Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer

Winston R. Becker, Stephanie A. Nevins, Derek C. Chen, Roxanne Chiu, Aaron M. Horning, Tuhin K. Guha, Rozelle Laquindanum, Meredith Mills, Hassan Chaib, Uri Ladabaum, Teri Longacre, Jeanne Shen, Edward D. Esplin, Anshul Kundaje, James M. Ford, Christina Curtis, Michael P. Snyder and William J. Greenleaf

To chart cell composition and cell state changes that occur during the transformation of healthy colon to precancerous adenomas to colorectal cancer (CRC), we generated single-cell chromatin accessibility profiles and single-cell transcriptomes from 1,000 to 10,000 cells per sample for 48 polyps, 27 normal tissues and 6 CRCs collected from patients with or without germline APC mutations. A large fraction of polyp and CRC cells exhibit a stem-like phenotype, and we define a continuum of epigenetic and transcriptional changes occurring in these stem-like cells as they progress from homeostasis to CRC. Advanced polyps contain increasing numbers of stem-like cells, regulatory T cells and a subtype of pre-cancer-associated fibroblasts. In the cancerous state, we observe T cell exhaustion, RUNX1-regulated cancer-associated fibroblasts and increasing accessibility associated with HNF4A motifs in epithelia. DNA methylation changes in sporadic CRC are strongly anti-correlated with accessibility changes along this continuum, further identifying regulatory markers for molecular staging of polyps.

The identification of genes and pathways that drive formation of invasive cancers has been the central focus of a number of large-scale genomics efforts. These efforts have cataloged the diversity and commonality of many genetic and transcriptional changes that accompany malignancy in diverse cancer types. However, most studies have focused on bulk profiling of advanced stage tumors and have largely ignored premalignant lesions. As a result, a detailed understanding of the progression of phenotypic changes that occur during the transition from normal to precancerous to cancerous state, as well as the molecular drivers of this transformation, remain unexplored.

CRC is an ideal system to study the continuum of phenotypic states along malignant transformation as it follows a stereotypical progression from normal to atypical to carcinoma that includes the formation of precancerous polyps, which can subsequently give rise to CRCs. A number of the changes associated with these transitions are nearly universal to all CRC malignancies, as typified by the adenoma-to-carcinoma sequence. For example, an estimated 80–90% of colorectal tumors are initiated by loss of APC, resulting in β-catenin stabilization and increased WNT signaling leading to intestinal hyperplasia. Subsequent mutations in other cancer driver genes such as KRAS, TP53 and SMAD4 result in the transformation to carcinoma.

Because APC mutations are almost universally the initiating event for polyps and CRCs, patients with familial adenomatous polyposis (FAP), who have germline mutations in APC, are a suitable population in which to study the natural progression of polyposis. These patients typically develop hundreds of polyps by early adulthood, and therefore an individual patient can provide numerous polyps of varied molecular ages and stages of progression, all arising in the same germline background.

To chart the regulatory and transcriptomic changes that occur on the phenotypic continuum from healthy colon to invasive carcinoma, as part of the Human Tumor Atlas Network, we profiled single-nuclei transcriptomes (single-nucleus RNA sequencing (RNA-seq) and single-cell assay for transposase-accessible chromatin using sequencing (ATAC-seq)) and epigenomes from polyps and CRCs. Many polyps were obtained from patients with FAP who underwent surgical colectomies, allowing both analysis of polyps with diverse sizes and locations of origin, and collection of neighboring unaffected colon tissue. From these single-cell datasets, we first catalog immune, stromal and epithelial cell types. We find large shifts in fibroblast subpopulations that occur along the transition from normal colon to CRC. We identify a subpopulation of exhausted T cells present only in CRC tissue. We observe a much larger fraction of cells exhibiting a stem-like state (both transcriptionally and epigenetically) within polyps and CRCs. We find that polyps populate an epigenetic and transcriptional continuum from normal colon to CRC characterized by sequential opening and closing of chromatin upregulation and downregulation of genes associated with the cancer state. We identify regulatory elements and transcription factors.
Articles NATURE GENETICS

Articles NATURE GENETICS

(TFs) associated with different stages of transformation from normal colon to carcinoma, including early increases in accessibility of regions containing TCF and LEF motifs and loss of accessibility in regions containing KLF motifs. In the final stage of this pathway, malignant transformation, we observe increased accessibility in regions containing HNF4A motifs. Finally, we show that accessibility changes in polyps are strongly anti-correlated with DNA methylation changes in sporadic CRC, and identify a subset of these

Fig. 1 | Single-cell atlas of expression and chromatin accessibility in CRC development. a, Summary of the samples in this study. The bar chart shows the number of normal/unaffected colon tissues (gray), adenomas (purple) and CRCs (red) assayed for each patient. Locations of samples assayed from a single patient are indicated on the colon on the upper right. These data include deep profiling of four patients with FAP from whom we assayed 8–11 polyps, 0–1 carcinomas and 4–5 matched normal (unaffected) tissues. From non-FAP donors, we collected data on normal colon (9 samples from 2 donors), polyps (1 sample from 1 donor) and CRC tissues (4 samples from 4 patients). b,c, UMAP representations of all snRNA-seq (b) and scATAC-seq (c) cells colored by whether the cells were isolated from normal/unaffected colon tissues, adenomas or CRCs. d,g, UMAP representations and annotations of immune (d) and stromal (g) cells. e,h, Fraction of each immune (e) and stromal (h) cell type isolated from normal (green), unaffected (blue), polyp (purple) and CRC (red) samples. The color gradations within each color represent the contributions of each single sample (for example, each shade of red is a single CRC). f, CODEX images of eight polyps and two CRCs where cells are labeled with dark blue, CD3 is labeled in green and PD1 is labeled in light blue. All samples tested are shown in f. CODEX imaging of individual specimens was not reproduced. Representative sections of images of the entire specimen are shown in the figure. DC, dendritic cell; Fib., fibroblast; GC, germinal center; ILC, innate lymphoid cell; Myofib., myofibroblast/smooth muscle; NK, natural killer.
regions that change their accessibility state early in the malignant continuum, suggesting potential strategies for detection of pre-malignant polyps.

**Results**

**Mapping molecular changes across malignant transformation.** We generated single-cell data for 81 samples collected from eight FAP and seven non-FAP donors (Fig. 1a and Supplementary Tables 1 and 2). For each tissue, we performed matched scATAC-seq and snRNA-seq (10x Genomics). We obtained high-quality single-cell chromatin accessibility profiles for 447,829 cells from 80 samples, with a mean transcription start site (TSS) enrichment of ~8 for most samples (Extended Data Fig. 1a). After removing low-quality snRNA-seq cells and samples, we obtained single-cell transcriptomes for 201,884 cells from 70 samples (Extended Data Fig. 1b). Whenever there was sufficient tissue, we generated microscopic pathology data (Extended Data Fig. 2a and Supplementary Table 2) and found the majority of polyps were tubular adenomas, the most common polyp type identified in colonoscopies.

When all snRNA-seq cells (Fig. 1b) and scATAC-seq cells (Fig. 1c) are projected into low-dimensional subspaces, stromal and immune cells generally cluster by cell type whereas epithelial cells largely separate into distinct clusters comprising cells derived from polyps, unaffected tissues or CRCs. As a result, we annotated immune and stromal cells by subclustering cells from all samples, and analyzed epithelial cells separately.

**T cells and myeloid cells are enriched in polyps and CRC.** The immune compartment comprised B cells, T cells, monocytes, macrophages, dendritic cells and mast cells (Fig. 1d). We examined expression of known marker genes (Extended Data Fig. 1c) to annotate snRNA-seq data, and examined chromatin activity scores—a measure of accessibility within and around a given gene body—associated with marker genes to annotate the scATAC cells (Extended Data Fig. 1d). We identified a cluster of exhausted T cells in the scATAC data that exhibited high gene scores of T cell exhaustion marker genes and accessibility at exhausted T cell motifs, and was labeled as exhausted T cells by a published dataset (Extended Data Fig. 3a–g and Methods)23. The cell types identified were present in nearly all samples, although some cell types were enriched or depleted in specific disease states (Fig. 1e and Extended Data Figs. 2b,c, 3h and 4). Significant differences in cell-type abundance were identified with Wilcoxon testing and a generalized linear model-based method called Milo16, which produced consistent results. For example, regulatory T cells (Tregs) were enriched in polyps relative to unaffected tissue, while naïve B, memory B and germinal center cells were enriched in unaffected tissues relative to polyps (Extended Data Fig. 4a,b). Enrichment of myeloid cells and specific types of T cells and depletion of B cells was recently reported in a group of 22 mismatch repair-proficient and 13 mismatch repair-deficient CRCs35, and we observe similar shifts in the tumor immune composition in precancerous polyps.

The enrichment of (1) Tregs in both polyps and CRC and (2) exhausted T cells in CRC suggests mechanisms of immune evasion in the precancerous and cancerous states48. T cell exhaustion, which occurs in response to chronic antigen stimulation and is characterized by reduced cytokine production and increased expression of inhibitory receptors, is thought to be a primary mechanism of immune evasion by cancers39,40. To further support the observation of T cell exhaustion only occurring in CRC, we performed CODEX imaging of CD3 and PD1 and found low or undetectable PD1 expression in eight polyps but found PD1 expression in both CRC samples tested (Fig. 1f).

Within the stromal compartment, we identified glial cells, adipose cells and multiple types of endothelial cells and fibroblasts (Fig. 1g). Fibroblast subtypes include crypt fibroblasts (WNT2B or RSPO3 high), villus fibroblasts (WNT5B high) and myofibroblasts (ACTA2 and TAGLN high) (Extended Data Figs. 1f and 5a)41,42. Consistent with previous results, we observe high expression of BMP signaling genes in villus fibroblasts (Extended Data Fig. 5a). In agreement with recent reports that crypt fibroblasts secrete semaphorins to support epithelial growth, we observe one fibroblast cluster with high expression of semaphorins (Extended Data Fig. 5a)43. This cluster of fibroblasts exhibited the highest expression of RSPO3, a factor that supports the intestinal stem cell niche44. We also observe a cluster of cancer-associated fibroblasts (CAFs) consisting almost exclusively of cells from CRCs, and a scATAC cluster of fibroblasts enriched for cells from polyps and CRCs with accessibility around some of the same genes as CAFs, which we term pre-cancer-associated fibroblasts (preCAFs) (Fig. 1h and Extended Data Figs. 2d,e and 4). These observations suggest that phenotypically distinct fibroblasts exist in polyps and tumors, and thus may play a role in tumorigenesis in precancerous lesions.

We next integrated our scATAC-seq and snRNA-seq datasets to enable analyses of regulatory elements and TFs potentially driving gene expression. We aligned the datasets with canonical correlation analysis (CCA) and assigned RNA-seq profiles to each scATAC-seq cell (integrated expression)35. We then labeled scATAC cells with the nearest snRNA-seq cells, which closely agreed with manual immune (Extended Data Fig. 1i) and stromal (Extended Data Fig. 5b) annotations. Finally, we identified peaks highly correlated to gene expression of proximal genes in our datasets, which resulted in 52,443 stromal peak-to-gene links (Extended Data Fig. 5c,d).

**scATAC reveals preCAF population.** CAFs promote cancer development and progression through diverse mechanisms including matrix remodeling, signaling interactions with cancer cells and perturbation of immune surveillance26–28. We observe a CAF cluster with high expression of known CAF marker genes FAP and TWIST1 (Extended Data Fig. 5a)45,46. Among the most significant snRNA-seq markers for CAFs were FAP, VCAN and COLIA2, which are involved in extracellular matrix remodeling and upregulated in multiple cancers30–32 (Fig. 2a). Specific expression of these genes by CAFs suggests fibroblasts participate in unique extracellular matrix remodeling in cancerous tissues that does not occur in normal colon or precancerous polyps.

While CAFs are known to promote CRC progression, we next explored the role of fibroblasts in precancerous lesions. Because the preCAF cluster was enriched for cells from polyps, we examined accessibility around marker genes for CAFs and found many of these genes more accessible in preCAFs than other fibroblast subtypes. For example, CAFs secrete WNT2 to promote cell proliferation and angiogenesis in CRCs31,46. CAFs and preCAFs exhibit the greatest accessibility at the WNT2 TSS (Fig. 2b), suggesting that chromatin changes promote expression of WNT72 in CAFs and preCAFs. We also observed that preCAFs demonstrated higher integrated expression of multiple CAF marker genes than other fibroblast subtypes (Extended Data Fig. 5e). We computed global CAF accessibility scores for all fibroblast subtypes (Methods) and found that preCAFs had the highest median CAF scores other than CAFs (Extended Data Fig. 5f). Further, accessibility in CAFs was most correlated with preCAFs; however, the correlation with one crypt fibroblast subtype was only slightly lower (Extended Data Fig. 5g). Together, this highlights the similarities between CAFs and preCAFs and suggests that preCAFs may perform similar functions to CAFs.

**RUNX1 is associated with widespread accessibility in CAFs.** We found that CAF marker peaks were enriched for JUN/FOS and CEBP motifs and preCAF marker peaks were enriched for JUN/FOS and FOX motifs (Fig. 2c,d and Methods). To nominate TFs
driving changes in chromatin accessibility in different stromal cell types, we identified TFs with the highest correlation between their gene expression and the chromatin accessibility activity level of its DNA motif (Fig. 2e, x axis). Amongst the most correlated TFs were RUNX1, RUNX2 and CEBPB. We next plotted the expression and motif activities of these TFs on the Uniform Manifold Approximation and Projection (UMAP) representation of the stromal cells and in violin plots grouped by each cell type (Fig. 2f), and noted that chromatin activity levels for RUNX1 and RUNX2, which have similar motifs, are highest in CAFs and preCAFs. However, RUNX1 is primarily expressed in CAFs and preCAFs, while RUNX2 has much lower expression in CAFs, suggesting that RUNX1 is a stronger driver of accessibility at RUNX motifs than is RUNX2 in CAFs.

Fig. 2 | Epigenetic regulators of preCAFs and CAFs. a, Dot plot representation of significant (MAST test) marker genes for CAFs. b, Genomic tracks for accessibility around WNT2 and RUNX1 for different stromal cell types. Peaks called in the scATAC data and peaks-to-gene links are indicated below the tracks. For example, a regulatory element ~50 kb away from the WNT2 TSS that is most accessible in CAFs whose accessibility is highly correlated to gene expression of WNT2 is indicated below the tracks. Marker peaks (Wilcoxon FDR ≤ 0.1 and log2FC ≥ 0.5) for each fibroblast subtype are indicated below the tracks. c, Marker peaks (Wilcoxon FDR ≤ 0.1 and log2FC ≥ 0.5) for each stromal cell type. Significance is determined by comparing each cell type with a background of all other cell types. d, Hypergeometric enrichment of TF motifs in stromal cell marker peaks. e, Plot of maximum difference between chromVAR deviation z-score, depicting TF motif activity, against correlation of chromVAR deviation and corresponding TF expression. TFs with maximum differences in chromVAR deviation z-score in the top quartile of all TFs and a correlation of greater than 0.5 are indicated in red. f, RNA expression (top) and chromVAR deviation z-scores (bottom) for selected TFs. The RNA expression plotted is the expression in the nearest RNA cell following integration.
Consistent with the expression of these genes, we observed the greatest accessibility around the RUNX1 TSS in CAFs and preCAFs (Fig. 2b). When comparing gene scores for each stromal cell type with all other stromal cells, preCAFs had significantly higher RUNX1 gene scores (log, fold-change (log2FC) > 1 and false discovery rate (FDR) < 0.01), and no other cell types met this significance threshold. When identifying accessibility closest to RUNX1, we found five significant marker peaks for preCAFs and four for CAFs (Fig. 2b).

Polyps are enriched for stem-like epithelial cells. We examined the epithelial cells that initially clustered by unaffected, polyp or CRC disease state (Fig. 1b,c and Extended Data Fig. 6e). To analyze these data, we first constructed RNA-seq and ATAC-seq references.
The regulatory trajectory of malignant transformation. a. Malignancy continuum for snRNA-seq (left) and scATAC-seq (right). Principal components were computed on the log2FC values between stem-like cells from each sample and normal colon stem cells for the set of peaks and genes that were significantly differential (Wilcoxon FDR ≤ 0.05 and |log2FC| ≥ 1.5 for peaks; MAST test for genes) in at least two samples. A spline was fit to the first two principal components (red) and samples were ordered based on their position along the spline. b. Genomic alterations in common driver genes defined in the malignancy continuum defined in the correspondence matrix. Samples are colored based on if they are derived from polyps or CRCs. c. Genomic alterations in common driver genes defined in the malignancy continuum defined in the correspondence matrix. Samples are colored based on if they are derived from polyps or CRCs. d. Genomic alterations in common driver genes defined in the malignancy continuum defined in the correspondence matrix. Samples are colored based on if they are derived from polyps or CRCs. e. Heatmap of all genes (e) and peaks (f) that were significantly differentially expressed (MAST test, adjusted P ≤ 0.05) and |log2FC| ≥ 1.5 for peaks; MAST test for genes) in at least two samples. A spline was fit to the first two principal components (red) and samples were ordered based on their position along the spline. The regulatory trajectory of malignant transformation. f. Log2FC in expression of genes defined in the malignancy continuum defined in the correspondence matrix. Samples are colored based on if they are derived from polyps or CRCs. g. Hypergeometric enrichment of TF motifs in k-means clusters of peaks defined in the malignancy continuum defined in the correspondence matrix. Samples are colored based on if they are derived from polyps or CRCs. h. Log2FC in expression of genes defined in the malignancy continuum defined in the correspondence matrix. Samples are colored based on if they are derived from polyps or CRCs.
We classified all epithelial cells based on the nearest normal cells in the projection and found that cells originating from polyps and CRC samples are enriched for stem-like epithelial cells and depleted for mature enterocytes, suggesting that epithelial cells increasingly demonstrate a stem-like phenotype during the transformation from normal to polyp (Fig. 3b–d and Extended Data Fig. 4a,b). We speculate that the populations of stem-like cells in the polyps and CRCs likely represent the ‘cancer’ stem cells in these tissues. Expression of previously described intestinal stem cell and colon cancer stem cell marker genes in these stem-like populations is discussed in detail in a Supplementary Note and Extended Data Fig. 7a.

To quantify the degree of stemness in individual cells within samples, we assigned scores quantifying stemness for each snRNA-seq and scATAC-seq sample and ordered samples by the distribution of stem scores within each sample (Methods and Fig. 3e). As expected, unaffected samples have generally lower stem scores. A number of polyps clustered near the unaffected tissues, suggesting that they are relatively benign. However, cells from most polyps and CRCs typically had higher stem scores, with some demonstrating a larger spread of stemness and others with much tighter distributions of stem scores, indicating that some polyps may be more heterogeneous. Similar results were observed when ordering samples based on the nearest normal cell type in the projection into the normal colon subspace (Methods and Extended Data Fig. 7h).

**Stem-like cells form a potential malignancy continuum.** We next compared the gene expression and chromatin accessibility of polyp and CRC stem-like cells with normal stem cells to identify the aberrant gene expression and regulatory programs in precancerous and cancerous lesions. After computing differential peaks between stem-like cells from each sample and cells from the nearest normal cell type, we computed the principal components of the log₂FC for stem-like cells from each sample and cells from the nearest normal cells in the samples, or if other epithelial cells also exhibit a similar continuum, which would be consistent with these cells being the only malig-

**Fig. 5 | Dynamics of cell-type representation in malignant transformation.** a–h. Fraction of cell type in each scATAC sample plotted against position of the sample in the malignancy continuum defined in Fig. 4d for stem-like cells (a), enterocytes (b), immature goblet cells (c), goblet cells (d), Tregs (e), exhausted T cells (f), preCAFs (g) and CAFs (h). Samples are colored based on if they are derived from unaffected tissues, polyps or CRCs. Fractions are computed by dividing the number of cells of a given cell type by the total number of cells in the compartment (epithelial versus immune versus stromal).

i. Stacked boxplot representation of the fraction of epithelial cells of each cell type for each scATAC sample along the malignancy continuum.
microscopic pathology and genomic alterations (Fig. 4b) is discussed in the Supplementary Information.

After computing the trajectory, we repeated the differential analysis using all unaffected samples rather than normal samples to increase the total number of patients and cells in the background group. We observe that the absolute number of significantly differential peaks and genes gradually increased along the malignancy continuum—with adenocarcinoma samples exhibiting the largest number of differential peaks and genes (Fig. 4c,d).

**Gene expression changes along the malignant continuum.** We examined gene expression changes along this malignancy continuum...
by selecting genes differentially expressed in at least two samples then clustering these genes into ten \( k \)-means clusters (Fig. 4e). These clusters correspond to groups of genes that become differentially expressed at distinct stages of malignant transformation. For example, clusters 1–4 comprise genes upregulated in stem-like cells in early-stage polyps when compared with unaffected stem cells. Members of cluster 4 include \( OLFM4 \), a marker of intestinal stem cells\(^{24} \), indicating that \( OLFM4 \) expression increases in stem-like cells from polyps as they approach malignancy. Cluster 4 also includes \( GPX2 \), a glutathione peroxidase known to be upregulated in CRC that functions to relieve oxidative stress by reducing hydrogen peroxide, facilitating both tumorigenesis and metastasis\(^{37} \) (Fig. 4h). The upregulation is not donor dependent, and we observe the same trend across all donors in our study (Extended Data Fig. 6g). We observed translation Gene Ontology terms enriched in cluster 4 and splicing and RNA-processing Gene Ontology terms enriched in cluster 2 (Extended Data Fig. 6k). Clusters of genes that gradually reduce expression along the transition from normal colon to cancer (clusters 6–9) and genes specific to malignant transformation are discussed in a Supplementary Note and Extended Data Fig. 8a.

**Polyps demonstrate increased activity of TCF and LEF.** To identify groups of polyps associated with invasive transformation, we clustered the 36,374 peaks significantly differential compared with the nearest unaffected cell type in at least two samples into ten \( k \)-means clusters (Fig. 4i), revealing five clusters that become more accessible and five clusters that become less accessible at different stages of the transition to cancer. To identify TFs driving chromatin accessibility changes in the transition from normal colon to CRC, we computed hypergeometric enrichment of motifs in each cluster of peaks from Fig. 4f (Fig. 4g) and ensured the stability of these results (Extended Data Fig. 7b–g).

TCF and LEF family motifs were enriched in all clusters that became more accessible across the malignancy continuum (clusters 1–5), consistent with the fact that loss of APC leads to \( \beta \)-catenin accumulation in the nucleus, which interacts with TCF and LEF TFs to drive WNT signaling\(^{16,17,36} \). This regulatory transformation is gradual across the malignant continuum—new peaks containing TCF and LEF motifs continue to open at all stages of colon cancer development, as does overall accessibility aggregated across TCF and LEF motifs, suggesting that WNT signaling gradually increases throughout this transformation, over and above what is observed in normal stem cell populations.

Cluster 3 peaks, which became more accessible in later-stage polyps and CRC, also exhibited enrichments of ASCL2 motifs (Fig. 4g). ASCL2 is a master regulator of intestinal stem cell fate, and induced deletion of ASCL2 leads to loss of LGR5\(^+ \) intestinal stem cells in mice\(^{38} \). Consistent with a linkage between a more stem-like state in polyp epithelium and more advanced malignant continuum scores, ASCL2 expression gradually increases as polyps approach malignant transformation (Fig. 4h), again indicative of a ‘super stem’-like phenotype, wherein master regulators of stem state are even more active than they are in normal stem cells.

Motifs lost along the malignancy continuum include HOX family motifs, KLF motifs and GATA motifs (Fig. 4g), and specific KLF TFs along the malignancy continuum are discussed in detail in a Supplementary Note and Extended Data Fig. 8d,e. Clusters 4 and 5 exhibit large accessibility increases only in CRC samples, and the greatest enrichment for HNF4A motifs (Fig. 4g). This observation suggests differential usage of HNF4A in polyps, where it decreases to drive WNT signaling, versus in CRC, where it is upregulated to drive cancer-specific accessibility differences (Supplementary Note and Extended Data Fig. 8b,c).

**Remodeling of cellular composition along malignant continuum.** We calculated the fractional contributions of each cell type to each sample as a function of position in the malignancy continuum, and found some cell types were highly correlated with progression along the malignancy continuum. For example, the fraction of stem cells within a sample gradually increases throughout malignant transformation (Fig. 5a,i). Similarly, the number of mature enterocytes decreases as polyps transform to carcinomas (Fig. 5b,i). Milo analysis revealed that neighborhoods of stem-like cells tend to be significantly more abundant at the end of the malignancy continuum (Extended Data Fig. 4b). In the secretory compartment, which primarily consists of immature and mature goblet cells, we observe a fractional increase in immature goblet cells in many polyps. In carcinomas we see a pervasive lack of differentiation into the secretory lineage, effectively eliminating immature and mature goblet cells (Fig. 5c,d,i). This observation is consistent with previous work reporting a depletion of goblet cells in nonmucinous colon adenocarcinomas\(^{39} \). Previous work has also found that knockout of MUC2 leads to the formation of more adenomas and carcinomas in mice\(^{40} \), suggesting that the loss of immature and mature goblet cells may even contribute to tumorigenesis.

Outside the epithelial compartment, we also observe changes in cellular composition across the transition from unaffected to polyp to carcinoma. Within the stromal compartment, the fraction of preCAF s gradually increases, while CAFs only appear in CRCs (Fig. 5g,h). Within the immune compartment, Tregs are increased in the more malignant polyps and CRCs, while exhausted T cells only appear in CRCs (Fig. 5e,f and Extended Data Fig. 4b). Tregs are known to suppress the antitumor immune response and are typically present at high levels in the tumor microenvironment\(^{41} \). The gradual increase in Tregs may be a mechanism of immune evasion in precancerous polyps. We discuss possible cell–cell interactions between stromal and epithelial cells along the malignant continuum in a Supplementary Note and in Extended Data Fig. 8f,g.

**Comparing CRC DNA methylation changes with continuum accessibility.** Aberrant DNA methylation is a primary mechanism of tumorigenesis in CRC\(^{42,43} \), but the timing and extent to which methylation changes drive changes in chromatin accessibility before and during malignant transformation is not known. We identified differentially methylated probes between normal and CRC samples (Extended Data Fig. 9d) in The Cancer Genome Atlas (TCGA) DNA methylation data (Illumina 450K array)\(^{44} \). For the ~89,000 chromatin accessibility peaks from epithelial cells that overlap at least one 450K array probe, we determined how many overlapped at least one hypermethylated site, at least one hypomethylated site or no differentially methylated sites. We then divided the peaks into groups based on whether they were members of significantly upregulated or significantly downregulated clusters identified in Fig. 4b.

For peaks overlapping hypomethylated probes, approximately one-third (534) belonged to clusters that became significantly more accessible along the continuum, while <0.5% (5) became significantly less accessible (Fig. 6a). We saw similar correspondence for peaks overlapping hypermethylated probes, with approximately one-quarter (754) becoming less accessible, and <0.5% (9) becoming more accessible. Therefore, hypermethylation and hypomethylation in CRC nearly perfectly predict that accessibility at that site will either decrease or increase (respectively), or remain unchanged. In peaks not meeting the significance threshold, we still observe less aggregate accessibility within peaks overlapping hypermethylated probes and more accessibility when they overlap hypomethylated probes (Fig. 6b). However, we also observe that 79.4% (2,096) of significantly more accessible and 76.3% (2,440) of less accessible peaks overlap nondifferential probes, implying that a majority of chromatin accessibility changes are likely not driven by methylation.

We next plotted the number of differential peaks overlapping hypermethylated and hypomethylated probes across the malignancy continuum (Fig. 6c), and found that changes in chromatin
accessibility that occur in regions that are ultimately differentially methylated in CRC accumulate along the transition from normal to cancer, with the greatest number observed in late-stage polyps and CRC.

Among regions that overlap hypermethylated probes in CRC that become less accessible in polyps, some have been previously reported cancer-specific hypermethylated loci. For example, the promoter region and multiple distal regulatory elements near the *ITGA4* gene are accessible in normal colon, unaffected FAP colon and very early-stage polyps, but become closed early in the progression to CRC and remain closed even in low-grade polyps (Fig. 6d). The gene with the most nearby differential peaks overlapping hypermethylated probes in our dataset was *NR5A2*. Multiple peaks near this gene become less accessible along the malignancy continuum (Fig. 6d) and expression of *NR5A2* also gradually decreases along the malignancy continuum (Extended Data Fig. 6h). *NR5A2* is a nuclear receptor that has been linked to a wide range of functions including inflammation and cell proliferation. The hypermethylation, decrease in accessibility, and decrease in gene expression of *NR5A2* suggests that the pro-inflammatory state that may be triggered by the loss of *NR5A2* might have a role in tumorigenesis.

Hypermethylated DNA regions in CRC have also been incorporated into CRC screening tests, including hypermethylation of the promoter regions of *BMP3* and *NDRG4* (ref. 51). We observe multiple distal elements around *BMP3* that become inaccessible in the middle of the malignancy continuum (Extended Data Fig. 9a). We observe many regions with a similar behavior: sharp increases or decreases in accessibility at a specific point along the malignancy continuum. We speculate that testing for accessibility, or methylation, at these loci may enable staging of polyps along the malignancy continuum. This approach also identifies methylation markers/loci (for example, *GRASP*, *CIDEB*) for specific malignant transformation in CRC (Extended Data Fig. 9b,c), and differential genes whose promoters overlap CRC methylation changes (Extended Data Fig. 9e).

**Discussion**

**Strategies to identify individuals in a premalignant stage**. Where interventions might be highly efficacious, promise tractable means to prevent cancer deaths. However, most previous work profiling genetic, epigenetic and transcriptomic changes that occur in malignancy has focused on advanced tumors rather than premalignant lesions. Our single-cell atlas of colon cancer tumorigenesis fills this gap by identifying key changes in chromatin accessibility, gene expression and tissue composition that occur along this transformation, and provides a wealth of potential targets for prevention, diagnosis and treatment of malignancy.

Analysis of both the composition and cell state of precancerous epithelial cells revealed that an increasing fraction of epithelial cells occupy a stem-like state as polyps approach malignancy, but these cells also exhibit underlying dysfunctional epigenetic and gene expression programs distinct from normal stem cells. We also identify compositional and cell state changes of noncancerous cells, including fibroblasts and immune cells, which may influence the tumor microenvironment to drive cancer progression. Within the fibroblast compartment, we identified a population of fibroblasts enriched in polyps and adenocarcinoma samples that shared many features with CAFs. We discuss the possibility that preCAFs may be on the path of becoming CAFs in a Supplementary Note and Extended Data Fig. 10.

Gene expression changes along the malignancy continuum implicate mechanisms of cancer initiation and nominate diagnostic and therapeutic targets. We find that expression of *GPX2*, which encodes a glutathione peroxidase, is gradually upregulated across the transformation from normal tissue to malignancy. Even in premalignant tissues, *GPX2* is upregulated, suggesting that its role in reducing the oxidative environment may be needed for progression along this continuum. As a result, we speculate that expression of *GPX2* could serve as a marker for the degree of polyp malignancy, and that inhibitors of GPX enzymes—such as tiopronin—may be relevant treatment strategies for CRC or premalignant lesions.

We identified several TFs associated with chromatin accessibility changes as polyps transition to malignancy. In stem-like cells obtained from polyps, we observe that many TCF and LEF motifs inaccessible in normal intestinal stem cells become accessible, suggesting that WNT signaling increases along the malignant continuum over and above that of normal stem cells. The final step in cancer formation is malignant transformation, and we identify diverse changes in chromatin accessibility, gene expression and tissue composition associated with this transformation (Figs. 4 and 5 and Extended Data Figs. 4, 6 and 8). Our data indicate that HNF4A may be a key regulator of malignant transformation—as both the expression of this gene becomes upregulated and chromatin regions containing HNF4A motifs become accessible only after the transformation to CRC. In normal colon, HNF4A motifs are more accessible in mature enterocytes than stem cells, but HNF4A is upregulated in CRC stem cells and drives chromatin accessibility changes not observed in normal stem cells.

Previous work identified epigenetic factors associated with CRCs, and DNA methylation state markers are at the core of widely adopted screening tests for CRC. While it is unknown when in tumorigenesis these methylation state changes occur, our dataset revealed that these regions become inaccessible in the middle of the continuum. Because we observe that hypermethylated regions are strongly decreased in their accessibility and hypomethylated regions are strongly enriched for increased accessibility, our dataset can be used to stratify differentially methylated regions expected to be present early in the malignant continuum, those expected to be present later in the continuum and those expected to be present only in CRC, opening the possibility for a stage-specific molecular screening.

Currently, clinical guidelines for the timing and frequency of follow-up endoscopic screening after polypectomy depend on the size and degree of dysplasia of the polyps removed. We found variability in the degree of transformation in polyps with the same dysplasia classification. This work presents a strategy to order premalignant polyps by their degree of malignancy, which we speculate may be useful for staging polyps and assessing clinical risk. Perhaps the most straightforward approach to this type of staging is to use the fraction of Tregs or stem-like cells (through immunohistochemical labeling of cell-type-specific markers such as LGR5) as a proxy for position along the malignancy continuum (Fig. 4 and Extended Data Figs. 6 and 8). However, determining if these molecular features correlate with patient outcomes will require substantial future clinical investigation.

This work demonstrates that adenomatous polyps traverse a strikingly consistent epigenetic and transcriptional trajectory as they progress to CRC. These results lead us to question if a similar relatively uniform molecular phenotypic trajectory is common to other premalignant lesions that precede other cancers, or if other pre-cancers might traverse multiple diverse pathways on the way to malignant transformation. Further, we might also ask if CRCs continue to follow a consistent pathway to metastasis or if multiple distinct pathways are taken once malignant transformation occurs. We anticipate that similar single-cell integrative methods can be deployed to answer these questions.

**Online content**

Any methods, additional references, Nature Research 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/s41588-022-01088-x.

Received: 11 May 2021; Accepted: 28 April 2022; Published online: 20 June 2022

References

1. Weinstein, J. N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1061–1073 (2013).
2. International Cancer Genome Consortium et al. International network of cancer genome projects. Nature 464, 933–938 (2010).
3. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020).
4. Fodde, R., Smits, R. & Clevers, H. APC, signal transduction and genetic instability in colorectal cancer. Nat. Rev. Cancer 1, 55–67 (2001).
5. Aoki, K. & Taketo, M. N. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. J. Cell Sci. 120, 3327–3335 (2007).
6. Leslie, A., Carey, F. A., Pratt, N. R. & Steele, R. J. C. The colorectal adenoma-carcinoma sequence. Br. J. Surg. 89, 845–860 (2002).
7. Vogelstein, B. et al. Genetic alterations during colorectal-tumor development. N. Engl. J. Med. 319, 525–532 (1988).
8. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–767 (1990).
9. Mori, Y. et al. Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. J. Natl. Cancer Inst. 81, 229–233 (1992).
10. Logan, C. Y. & Nusse, R. The Wnt signaling pathway in development and disease. Annu. Rev. Cell Dev. Biol. 20, 781–810 (2004).
11. Schacht, E. M., Leach, B. I. & Dow, L. E. Wnt signaling and colorectal cancer. Curr. Colorectal Cancer Rep. 13, 101–110 (2017).
12. Grenen, J. et al. Identification and characterization of the familial adenomatous polyposis coli gene. Cell 66, 589–600 (1991).
13. Galiatsatos, P. & Foulkes, W. D. Familial adenomatous polyposis. Am. J. Gastroenterol. 101, 385–398 (2006).
14. Rozenblatt-Rosen, O. et al. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. Cell 181, 236–249 (2020).
15. Yost, K. E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. Nat. Med. 25, 1251–1259 (2019).
16. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using nearest neighbor k graphs. Cell 180, 6717–6730 (2014).
17. Schabath, M. B. et al. Defining ‘T cell exhaustion’. Nat. Rev. Immunol. 20, 637–654 (2020).
18. Vigneron, S. et al. Stromal versican regulates tumor growth by promoting angiogenesis. Cancer Med. 8, 6370–6382 (2019).
19. St Laurent, C. et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902.e21 (2019).
20. Fier, L. Gvander et al. OLFM4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. Gastroenterology 157, 15–17 (2009).
21. Emminck, B. L. et al. Gpax2 suppression of H2O2 stress links the formation of differentiated tumor mass to metastatic capacity in colorectal cancer. Cancer Res. 74, 6717–6730 (2014).
22. Cadogan, K. M. & Waterman, M. L. TCF/LEFs and Wnt signaling in the nucleus. Cell 175, 1847–1858 (2019).
23. L. Fler, G. et al. Transcription factor achaete-scute-like 2 controls intestinal stem cell fate. Cell 156, 903–912 (2019).
24. Leow, C. C., Romero, M. S., Polakis, P. & Gao, W.-Q. Hth1, down-regulated in colon adenocarcinomas, inhibits proliferation and tumorigenesis of colon cancer cells. Cancer Res. 64, 6050–6057 (2004).
25. L. Vezich, A. et al. Colorectal cancer in mice genetically deficient in the mucin Muc2. Science 295, 1726–1729 (2002).
26. Plitas, G. & Rudensky, A. Y. Regulatory T cells in cancer. Annu. Rev. Cancer Biol. 4, 459–477 (2020).
27. A. Ashktorab, H. et al. Human DNA methylation and colorectal cancer. Curr. Colorectal Cancer Rep. 10, 425–430 (2014).
28. Hinoue, T. et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. 22, 271–282 (2012).
29. C. Lengauer, C. et al. DNA methylation and genetic instability in colorectal cancer cells. Proc. Natl Acad. Sci. USA 94, 2545–2550 (1997).
30. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337 (2012).
31. Barault, L. et al. Discovery of methylated circulating DNA biomarkers for comprehensive non-invasive monitoring of treatment response in metastatic colorectal cancer. Gut 67, 995–2005 (2018).
32. Fernandez-Marcos, P. I., Auwerx, J. & Schoonjans, K. Emerging actions of the nuclear receptor LRH-1 in the gut. Biochem. Biophys. Acta 1812, 947–955 (2011).
33. Imperiale, T. F., Ransohoff, D. F. & Iizikowita, S. H. Multitarget stool DNA testing for colorectal-cancer screening. N. Engl. J. Med. 371, 187–188 (2014).
34. Hall, M. D. et al. Inhibition of glutathione peroxidase mediates the collateral sensitivity of multidrug-resistant cells to tiopronin. J. Biol. Chem. 289, 21473–21489 (2014).
35. Gupta, S. et al. Recommendations for follow-up after colonoscopy and polypectomy; a consensus update by the US Multi-Society Task Force on Colorectal Cancer. Gastrointest. Endosc. 91, 463–485.e5 (2020).

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

Experimental methods. Description of FAP donors. We collected samples from the following groups of patients: FAP (eight patients), routine colonoscopy screening (one patient), sporadic CRC (four patients) and healthy controls (two patients). This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. FAP tissue was collected at the time of partial or full colectomies for four patients and during screening colonoscopies for four patients. From non-FAP patients, one sporadic polyp was obtained during a standard colonoscopy as part of routine screening, four sporadic CRCs were obtained from the Stanford Tissue Bank and nine normal tissue samples were collected from brain-dead organ donors under consent.

Patient-matched normal colon mucosa, polypos and adenocarcinomas were flash frozen in liquid nitrogen at time of collection and stored at −80 °C. One polyp (A002-C-320) and one adenocarcinoma (A001-C-007) were embedded in optimal cutting temperature medium (OCT) and stored before storage at −80 °C. Polyps were sectioned by a board-certified pathologist for presence of dysplasia, including low or high grade with corresponding percentages. A small number of polyps (seven polyps) were exhausted for molecular assays so pathology reads were not obtained. The isolation −80 °C. One polyp −80 °C. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP.

Experimental methods. Description of FAP donors. We collected samples from the following groups of patients: FAP (eight patients), routine colonoscopy screening (one patient), sporadic CRC (four patients) and healthy controls (two patients). This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. FAP tissue was collected at the time of partial or full colectomies for four patients and during screening colonoscopies for four patients. From non-FAP patients, one sporadic polyp was obtained during a standard colonoscopy as part of routine screening, four sporadic CRCs were obtained from the Stanford Tissue Bank and nine normal tissue samples were collected from brain-dead organ donors under consent.

Patient-matched normal colon mucosa, polypos and adenocarcinomas were flash frozen in liquid nitrogen at time of collection and stored at −80 °C. One polyp (A002-C-320) and one adenocarcinoma (A001-C-007) were embedded in optimal cutting temperature medium (OCT) and stored before storage at −80 °C. Polyps were sectioned by a board-certified pathologist for presence of dysplasia, including low or high grade with corresponding percentages. A small number of polyps (seven polyps) were exhausted for molecular assays so pathology reads were not obtained. The isolation −80 °C. One polyp −80 °C. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP.

Methods

Experimental methods. Description of FAP donors. We collected samples from the following groups of patients: FAP (eight patients), routine colonoscopy screening (one patient), sporadic CRC (four patients) and healthy controls (two patients). This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. FAP tissue was collected at the time of partial or full colectomies for four patients and during screening colonoscopies for four patients. From non-FAP patients, one sporadic polyp was obtained during a standard colonoscopy as part of routine screening, four sporadic CRCs were obtained from the Stanford Tissue Bank and nine normal tissue samples were collected from brain-dead organ donors under consent.

Patient-matched normal colon mucosa, polypos and adenocarcinomas were flash frozen in liquid nitrogen at time of collection and stored at −80 °C. One polyp (A002-C-320) and one adenocarcinoma (A001-C-007) were embedded in optimal cutting temperature medium (OCT) and stored before storage at −80 °C. Polyps were sectioned by a board-certified pathologist for presence of dysplasia, including low or high grade with corresponding percentages. A small number of polyps (seven polyps) were exhausted for molecular assays so pathology reads were not obtained. The isolation −80 °C. One polyp −80 °C. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP.

Methods

Experimental methods. Description of FAP donors. We collected samples from the following groups of patients: FAP (eight patients), routine colonoscopy screening (one patient), sporadic CRC (four patients) and healthy controls (two patients). This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. FAP tissue was collected at the time of partial or full colectomies for four patients and during screening colonoscopies for four patients. From non-FAP patients, one sporadic polyp was obtained during a standard colonoscopy as part of routine screening, four sporadic CRCs were obtained from the Stanford Tissue Bank and nine normal tissue samples were collected from brain-dead organ donors under consent.

Patient-matched normal colon mucosa, polypos and adenocarcinomas were flash frozen in liquid nitrogen at time of collection and stored at −80 °C. One polyp (A002-C-320) and one adenocarcinoma (A001-C-007) were embedded in optimal cutting temperature medium (OCT) and stored before storage at −80 °C. Polyps were sectioned by a board-certified pathologist for presence of dysplasia, including low or high grade with corresponding percentages. A small number of polyps (seven polyps) were exhausted for molecular assays so pathology reads were not obtained. The isolation −80 °C. One polyp −80 °C. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP.

Methods

Experimental methods. Description of FAP donors. We collected samples from the following groups of patients: FAP (eight patients), routine colonoscopy screening (one patient), sporadic CRC (four patients) and healthy controls (two patients). This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. FAP tissue was collected at the time of partial or full colectomies for four patients and during screening colonoscopies for four patients. From non-FAP patients, one sporadic polyp was obtained during a standard colonoscopy as part of routine screening, four sporadic CRCs were obtained from the Stanford Tissue Bank and nine normal tissue samples were collected from brain-dead organ donors under consent.

Patient-matched normal colon mucosa, polypos and adenocarcinomas were flash frozen in liquid nitrogen at time of collection and stored at −80 °C. One polyp (A002-C-320) and one adenocarcinoma (A001-C-007) were embedded in optimal cutting temperature medium (OCT) and stored before storage at −80 °C. Polyps were sectioned by a board-certified pathologist for presence of dysplasia, including low or high grade with corresponding percentages. A small number of polyps (seven polyps) were exhausted for molecular assays so pathology reads were not obtained. The isolation −80 °C. One polyp −80 °C. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP. This study was approved by the Stanford Institutional Review Board and informed consent was obtained from all patients. All patients with FAP had clinical FAP.
After generation of a normal epithelial reference, we next aimed to project single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger, v.3.1.0) done with the Cell Ranger Pipeline (https://support.10xgenomics.com/snRNA-seq: initial processing. Initial processing of snRNA-seq data was done with the Cell Ranger Pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger, v3.1.0) by first running cellranger mk Faq to demultiplex the bcl files and then running cellranger count. Since nuclear RNA was sequenced, data were aligned to a Cellranger count. Since nuclear RNA was sequenced, data were aligned to a cellranger count. Since nuclear RNA was sequenced, data were aligned to a cellranger count. Since nuclear RNA was sequenced, data were aligned to a cellranger count. Since nuclear RNA was sequenced, data were aligned to a cellranger count. Since nuclear RNA was sequenced, data were aligned to a genome using normal stem cells as the background for differential testing and produced the log(counts per million) transformation with \(\text{edgeR::cpm(mat,log\_f} = 0.75\text{in} = \text{FALSE,}\) which were filtered out and, from the remaining genes, the top 1,600 most variable genes were identified. We then computed the term frequency–inverse document frequency (TF-IDF) transformation on these genes and performed SVD on the transformed matrix, and provided dimensions 1–8 of this reduction as input to Seurat’s shared nearest neighbor clustering with resolution of 0.1. We summed counts from each cluster and then computed the inverse document frequency from the previous TF-IDF transformation. These were then used to compute the new TF-IDF. The resulting LSI dimensionality reduction was provided as input to compute a UMAP representation of the data and the cells were clustered using a resolution of 1.0. The resulting clusters were then annotated based on expression of known marker genes. The progenitor cell annotations for colon epithelial cells were done following the procedure described previously\(^{30,31}\). Brefly, when computing the TF-IDF transformation on normal colon epithelial cells, we stored the colSums, rowSums and SVD. To project cells from additional samples into this subspace, we first zero out rows based on the initial TF-IDF rowSums. We next calculated the term frequency by dividing by the column sums and computed the inverse document frequency from the previous TF-IDF transformation. These were then used to compute the new TF-IDF. The resulting LSI approach with only eight dimensions, allowed us to denoise the data and limit batch effect, which was useful for the projections.

This final LSI dimensionality reduction was provided as input to compute a UMAP representation of the data and the cells were clustered using a resolution of 1.0. The resulting clusters were then annotated based on expression of known marker genes. The progenitor cell annotations for colon epithelial cells were done following the procedure described previously\(^{30,31}\). Brefly, when computing the TF-IDF transformation on normal colon epithelial cells, we stored the colSums, rowSums and SVD. To project cells from additional samples into this subspace, we first zero out rows based on the initial TF-IDF rowSums. We next calculated the term frequency by dividing by the column sums and computed the inverse document frequency from the previous TF-IDF transformation. These were then used to compute the new TF-IDF. The resulting LSI approach with only eight dimensions, allowed us to denoise the data and limit batch effect, which was useful for the projections.

This final LSI dimensionality reduction was provided as input to compute a UMAP representation of the data and the cells were clustered using a resolution of 1.0. The resulting clusters were then annotated based on expression of known marker genes. The progenitor cell annotations for colon epithelial cells were done following the procedure described previously\(^{30,31}\). Brefly, when computing the TF-IDF transformation on normal colon epithelial cells, we stored the colSums, rowSums and SVD. To project cells from additional samples into this subspace, we first zero out rows based on the initial TF-IDF rowSums. We next calculated the term frequency by dividing by the column sums and computed the inverse document frequency from the previous TF-IDF transformation. These were then used to compute the new TF-IDF. The resulting LSI approach with only eight dimensions, allowed us to denoise the data and limit batch effect, which was useful for the projections.
normal and colorectal adenocarcinoma samples were computed using TCGAAnalyze_DMR with a $P$ value cutoff of $10^{-4}$ and a mean difference in $\beta$ value cutoff of 0.25 to determine significance. Overlaps between DNA methylation probes and our peak set were identified with the GenomicRanges function FindOverlaps in R.

Supplementary methods. Please see the Supplementary methods for additional methodologic details, including details on CODEX multiplex imaging.

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

Data availability
Sequencing data have been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE20349. Original data generated in this study are also available on the Human Tumor Atlas Network (HTAN) Data Portal (unaffected FAP tissues, polyps and CRCs) at https://data.humantumoratlas.org/ under the HTAN Stanford Atlas and the HuBMAP data portal (normal colon tissues; https://portal.hubmapconsortium.org/) under the Stanford TMC. Unique IDs for accessing the HTAN datasets are listed in Supplementary Table 3 and unique IDs for accessing the HuBMAP datasets are listed in Supplementary Tables 4 and 5. Receptor ligand pairs from the Fantom5 database were downloaded from https://fantom.gsc-riken.jp/5/suppl/Ramolowski_et_al_2015/ (ref. 50). Clustered TF motifs can be downloaded from https://www.vierstra.org/resources/motif_clustering/downloads (ref. 51). Seurat objects for previously published single-cell colon data were downloaded from https://github.com/cssmillie/ulcerative_colitis (ref. 22). Counts matrices and T cell annotations for cells from BCC are available on GEO with accession number GSE123813 (ref. 48). TCGA DNA methylation data can be downloaded from the GDC data portal (https://portal.gdc.cancer.gov/)

Code availability
Scripts for processing the single-cell data, generating the malignancy continuum and analyzing the WGS data are available on GitHub (https://github.com/winstonbecker/scCRC_continuum).

References
54. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. Science 362, eaav1898 (2018).
55. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. 53, 403–411 (2021).
56. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. 37, 925–936 (2019).
57. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16, 1289–1296 (2019).
58. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. Nature 583, 729–736 (2020).
59. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. 8, 329–337.e4 (2019).
60. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophag. Nat. Immunol. 20, 163–172 (2019).
61. Fink, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology 16, 1–13 (2015).
62. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 44, e71 (2016).
63. Ramlowsk, J. A. et al. A draft network of ligand-receptor-mediated multicellular signalling in human. Nat. Commun. 6, 7866 (2015).

Acknowledgements
This work is supported by NIH grants no. U2C CA233311 (M.P.S., J.M.F., C.C., W.J.G.), no. 1U54HG010426 (M.P.S., W.J.G.), no. RM1-HG007735, no. UM1-HG009442, no. UM1-HG009436, no. R01-HG0099091 and no. U19-A1057266 (to W.J.G.). This work used the Genome Sequencing Service Center by Stanford Center for Genomics and Personalized Medicine Sequencing Center, supported by NIH grant no. S10OD025212 and NCI/NIH/NIH-NIDDK grant no. P30DK116074. W.R.B. was supported in part by the Stanford MSTP training grant no. T32GM070363 and the SIGF affiliated with ChEM-H. S.A.N. was supported in part by the Stanford Graduate Fellowship and the NIH/NIH Stanford Genome Training Program 5T32 000044. Y. Lin of Washington University in St. Louis provided the B001 and B004 tissue samples as part of the HuBMAP Consortium. The intestine image in Fig. 1a was created with BioRender.com.

Author contributions
W.R.B., S.A.N., M.P.S. and W.J.G. conceptualized the study. Single-cell experiments were performed by S.A.N., W.R.B., R.C. and D.C.C. WGS experiments and analysis were performed by A.M.H. Pathology reads were performed by T.L. and J.S. W.R.B. analyzed the single-cell data and generated the figures. R.L., M.M., A.M.H., U.L., E.D.E., R.C., S.A.N. and W.R.B. were responsible for sample collection. T.K.G. generated the histology slides and performed CODEX multiplexed imaging. H.C. performed sequencing. W.R.B., S.A.N. and W.J.G. wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. A.K., C.C., E.D.E., J.M.F., M.P.S. and W.J.G. were responsible for funding acquisition. E.D.E., A.K., J.M.F., C.C., M.P.S. and W.J.G. supervised the project.

Competing interests
W.J.G. is a consultant and equity holder for 10x Genomics, Guardant Health, Quantapore and Ultima Genomics, and cofounder of Protillion Biosciences, and is named on patents describing ATAC-seq. M.P.S. is a cofounder and scientific advisor for Personalis, Qbio, January.ai, Filtricine, Mirvet and Proteo, and an advisor for Genapsys. A.K. is a consultant with Illumina, Inc. The remaining authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01088-x.
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01088-x.
Correspondence and requests for materials should be addressed to Michael P. Snyder or William J. Greenleaf.
Peer review information Nature Genetics thanks Arjun Raj, Amos Tanay, Itai Yanai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Quality control and annotation of single-cell datasets. (a) Violin plots of TSS-enrichments for all scATAC cells from each sample. Samples are labeled by patient (for example A001, A002, etc), source (C = Colectomy, E = Colonoscopy, A = Autopsy, T = Tissue Bank), dissociation (D = dounce, S2 = S2 singulator). Replicates performed on additional sections of the same polyp are indicated with a R. (b) Violin plots of the percent of RNA that is mitochondrial RNA per sample and the number of UMIs sequenced for cells from each sample. Samples are labeled the same as in S1A, except all tissues were dounced so the dissociation method is not included. Boxplots represent the median, 25th percentile, and 75th percentile of the data and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot. (c) UMAP projection of scATAC immune cells colored by gene activity scores reflecting accessibility within and around immune marker genes. (d) UMAP projection of snRNA-seq immune cells colored by expression of immune marker genes. (e) UMAP projection of snRNA-seq immune cells colored by automated labeling of snRNA-seq immune cells with SingleR. (f) UMAP projection of scATAC stromal cells colored by gene activity scores of stromal marker genes. (g) UMAP projection of snRNA-seq stromal cells colored by expression of marker genes. (h, i) UMAP projection of scATAC immune cells where cells are labeled by the nearest snRNA-seq cell from Smillie et al or this study after integrating the respective datasets with CCA. (j) UMAP projections of four scATAC samples with nuclei isolated with both douncing and the S2 Singulator, colored by disease state (top) and dissociation method (bottom). (k) Fraction of epithelial cells of each cell type for the 4 samples where nuclei were isolated with douncing and the S2 singulator. (l) Differential peaks between scATAC stem cells isolated from two sections of the same polyp that were processed with either the S2 singulator or douncing. (m) UMAP representation of stromal cells following Harmony batch correction on LSI dimensions. (n) Violin plots of gene module scores for interferon gamma gene sets for immune cells from different disease states. (o) Violin plot of gene module scores for an interferon gamma gene set for different immune cell types.
Extended Data Fig. 2 | See next page for caption.
**Extended Data Fig. 2** | Cellular composition of samples in this study. (a) Metadata collected for different samples in this study. (b, c) Stacked bar plot representation of the fraction of all immune cells in each sample composed of each cell type for the scATAC (B) and snRNA-seq (c) datasets. Each column represents a single sample, with each color representing a different cell type present in the sample. (d, e) Stacked bar plot representation of the fraction of all stromal cells in each sample composed of each cell type for the scATAC (d) and snRNA-seq (e) datasets. Each column represents a single sample, with each color representing a different cell type present in the sample.
Extended Data Fig. 3 | T-cell annotation and donor contributions to clusters of cells. (a, b) UMAP projection of all T-cells identified in the scATAC data. Points on the UMAP represent single-cells and are colored by tissue of origin (c) and cell type annotations (d). (c) UMAP projection of scATAC T-cells with cells colored by gene activity scores depicting chromatin accessibility surrounding BATF, CTLA4, PDCD1, and TOX. (d) ChromVAR deviation z-scores depicting TF motif activity of BATF and NR4A2 plotted on scATAC T-cell UMAP. (e) UMAP projection of scATAC T-cells colored by labeling of scATAC-seq T-cells with nearest snRNA-seq T-cells in BCC after integrating the datasets with CCA. (F) MA plot showing differential peaks (Wilcoxon test) between exhausted T-cells and CD8+ T-cells. Motifs with hypergeometric enrichment in peaks more accessible in exhausted T-cells are listed in the plot. (g) Genomic tracks depicting accessibility around CD8 locus in T-cell subtypes. Genes on the + strand are indicated in red and genes on the - strand are indicated in blue. (h) Stacked bar graph representation of the fraction of each cell type derived from each patient in the study. Cells from each patient have a different color and different shades of the same color represent individual samples from a given patient. (i) UMAP representation of all scATAC cells colored by patient of origin.
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Differential cell type abundance. (a) Boxplots depicting the fraction of cells in a given compartment (immune, stromal, or epithelial) that are composed of a given cell type. Each box represents data for a single disease state (N: Normal, U: Unaffected, P: Polyp, A: CRC) in this study. Wilcoxon p-values are listed above the plot and were corrected with Bonferroni correction for multiple hypothesis testing within each cell type. Wilcoxon comparisons were made to normal colon for stromal and epithelial cell types and unaffected FAP colon for immune cell types. The boxplots and statistics are derived from 8 normal samples, 18 unaffected samples, 48 polyp samples, and 6 CRC samples in the epithelial compartment, 8 normal samples, 18 unaffected samples, 47 polyp samples, and 6 CRC samples in the immune compartment, and 8 normal samples, 16 unaffected samples, 46 polyp samples, and 6 CRC samples in the stromal compartment. Boxplots represent the median, 25th percentile, and 75th percentile of the data, whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot, and all points are plotted. (b) Milo analysis of differential abundance changes between polyp and unaffected samples. The left plots show comparisons of polyp and unaffected samples, which were selected since they have the greatest number of samples. Neighborhoods that are significantly differentially abundant in polyps are colored in red and neighborhoods that are differentially abundant in unaffected samples are colored in blue. The plots on the right show comparisons along the malignancy continuum, with neighborhoods that are significantly differentially abundant early in the continuum shown in blue and neighborhoods that are significantly differentially abundant late in the continuum shown in red. All comparisons in A and B use the scATAC data, as we had a greater number of scATAC cells.
Extended Data Fig. 5 | Cell type specific expression and RNA-ATAC integration of stromal cells. (a) Dotplot representation of RNA expression of myofibroblast, stem maintenance, SEMA, CCL, BMP, and CAF genes by cells in different fibroblast subtypes. (b) Labeling of scATAC cells by aligning scATAC and snRNA-seq data with CCA and labeling scATAC cells based on nearest snRNA-seq cells. (c) Peak-to-gene linkages between scATAC and snRNA-seq stromal cells (correlation ≥ 0.45). Rows in the left heatmap represent peaks and are colored by accessibility while rows in the right heatmap represent genes and are colored by expression. (d) Hypergeometric enrichment of motifs in clusters of peaks from S5C. (E) Integrated gene expression of CAF marker genes for different stromal cell types. (f) CAF scores for different stromal cell types depicted as violin plots with overlying boxplots. CAF scores are a measure of global accessibility at CAF marker peaks, and were defined by first identifying CAF marker peaks relative to all other cell types and then computing the number of Tn5 insertions in those marker peaks for each stromal cell and normalizing by the number of fragments in each cell. (g) Pearson correlation between accessibility at all peaks between CAFs and all other stromal cell types. (h) Violin plots showing the distribution of RUNX1 scATAC gene scores for cells of each stromal cell type. Boxplots in (e), (f), and (h) represent the median, 25th percentile, and 75th percentile of the data, and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot.
Extended Data Fig. 6 | Characterization of normal colon epithelium and identification of changes along the malignancy continuum. (a) Upper 8 panels: UMAP projection of normal colon epithelial cells colored by scATAC gene activity scores of the epithelial marker genes RETNLB (immature goblet), MUC2 (goblet), FEV (enteroendocrine), RAB6B (enterocyte), SOX9 (stem), BEST4 (Best4+ enterocyte), LGR5 (stem), and ASCL2 (stem). Lower 4 panels: UMAP projection of normal colon epithelial cells colored by expression of marker genes EPCAM (general epithelial), SMOC2 (stem), BEST4, and MUC2. (b) Violin and boxplot representation of gene expression of stem marker genes by epithelial cell type. Asterisks indicate that gene expression is significantly upregulated when compared to all other cell types. Boxplots represent the median, 25th percentile, and 75th percentile of the data and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplots. (c) Labeling of scATAC-seq epithelial cells by nearest snRNA-seq cells following integration of the datasets with CCA. (d) Confusion matrix comparing annotation of scATAC cells using marker genes and labeling of scATAC cells with the nearest snRNA-seq cell following integration of scATAC and snRNA-seq datasets. (e) UMAP representation of snRNA-seq epithelial cells colored by disease state. (f) Results of computing the continuum on plasma cells and TA2 cells using the same method performed for stem cells. (g) \( \log_{10} \text{FC} \) in expression of ASCL2, HNF4A, and GPX2 in stem-like cells from each sample relative to stem-like cells in unaffected samples plotted against the malignancy continuum defined in 4D. Samples are colored based on the patient the sample was collected from. (h) \( \log_{10} \text{FC} \) in expression of NR3C2, NORAD, SLC4A4, LRIG3, NR5A2, and RPL13 as a function of malignancy continuum. Samples are colored based on if they are derived from polyps or CRCs. (i) Relationship between the malignancy continuum and percent of sample with any degree of dysplasia as determined by microscopic pathology. Samples are colored based on gross classification as a polyp (purple) or unaffected (green) tissue. Note that some samples classified as unaffected had dysplasia while some samples classified as polyps did not have dysplasia. (j) Relationship between the malignancy continuums defined from the scATAC and snRNA-seq datasets. Samples are colored based on gross classification as unaffected, polyp, or CRC. (k) Enrichment of gene ontology terms in clusters of differential RNA genes in Fig. 4.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Motif enrichment in differential peaks with different numbers of k-means clusters. (a) Expression of intestinal stem cell and colon cancer stem cell marker genes in stem cells, TA2 cells, TA1 cells, and Enterocytes by sample. Samples are ordered by the malignancy continuum defined in Fig. 4. (b, c) Heatmaps of all peaks that were significantly differentially accessible (Wilcoxon test, \(p_{\text{adj}} \leq 0.05\) & \(|\log_{2} \text{FC} | \geq 1.5\)) in \(\geq 2\) samples. Samples are ordered along the x-axis by the malignancy continuum defined in Fig. 4. Peaks are k-means clustered into 5 (b) or 15 (c) clusters. (d, e) Hypergeometric enrichment of TF motifs in k-means clusters of peaks defined in B (d) and C (e). (f) Heatmap of all peaks that were significantly differentially accessible in \(\geq 2\) samples between stem cells from a given sample and normal colon stem cells. Samples are ordered along the x-axis by the malignancy continuum defined in Fig. 4. Peaks are k-means clustered into 10 groups. (g) Hypergeometric enrichment of TF motifs in k-means clusters of peaks defined in F. (h) Heatmap representation of cell types in each epithelial sample as determined by the nearest normal cell after projecting the cells into the normal LSI subspace.
Extended Data Fig. 8 | Epigenetic and transcriptomic changes along the malignancy continuum. (A) Dot plot representation of genes differentially expressed in CRC relative to polyps. (B) UMAP projection of normal colon epithelial cells colored by motif activity of HNF4A. (C) Dot plot representation of HNF4A expression in different normal colon epithelial subtypes. (D) Log2FC in expression of KLF TFs relative to unaffected colon as a function of position along the malignancy continuum. Samples are colored based on if they are from polyps (purple) or CRC (red). (E) Dotplot representation of the expression of KLF TFs in normal colon epithelial cells. (F) Log2FC in expression of SDC1, SDC4, and RPSA along the malignancy continuum. Samples are colored based on if they are from polyps (purple) or CRC (red). (G) Dotplot representation of the expression of selected ligands by different stromal cell types.
Extended Data Fig. 9 | See next page for caption.
Extended Data Fig. 9 | Accessibility changes in regions hypermethylated in CRC. (a–c) Accessibility tracks around BMP3 (a), GRASP (b), and CIDEB (c), which are hypermethylated in CRC. (d) Adjusted p-value and mean difference in β-value cutoffs used to determine differentially methylated probes. P values were determined using the two-sided Wilcoxon test and were adjusted with the Benjamini-Hochberg method for multiple hypothesis testing. (e) Genes that were significantly differential along the malignancy continuum that also have differentially methylated probes within 500 bp of their TSS in TCGA 450 K methylation data. Genes are grouped into a heatmap of those with hypermethylated probes in their promoters and a heatmap of those with hypomethylated probes in their promoters.
Extended Data Fig. 10 | Trajectory analysis of preCAFs and CAFs. (a) Changes in most variable peaks, TF motif activity scores, and gene expression along the trajectory from villus fibroblasts to preCAFs to CAFs. (b) Changes in most variable peaks, TF motif activity scores, and gene expression along a control trajectory from CAFs to villus fibroblasts to preCAFs.
Reporting Summary

Nature Research 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 Research 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.

- 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
- A description of any 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

- Code for generating fragments files for scATAC and counts matrices for single cell RNA was obtained from 10x genomics (go.10xgenomics.com/scATAC/cell-ranger-ATAC and https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger).

Data analysis

- cellranger-atac-1.2.0 – alignment of ATAC data and generation of fragments files
- cellranger-3.1.0 – alignment of RNA data and generation of counts matrices
- macs2 2.1.1.20160309 – Software for peak calling
- R version 3.6.1 – R environment for all custom code except running singleR.
- R version 4.0.2 – R environment for running singleR.
- ArchR - 0.9.5 - Software for analysis of scATAC-seq data.
- Seurat_3.1.1 – Software for analysis of scRNA-seq data.
- DoubletFinder_2.0.3 – Software for doublet removal for scRNA-seq
- BSGenome.Hsapiens.UCSC.hg38_1.4.1 – Package containing genomic DNA sequences
- TCGAbiolinks_2.12.6 – Software for analysis of DNA methylation data.
- limma_3.40.6 – Software used for GO enrichments
- Rcpp_1.0.4.6 – Software for C++ in R
- SingleR_1.4.1 – Software for automated cell annotation
- edgeR_3.26.8 – Used for analysis of single-cell data.
- harmony_1.0 – Software used for batch correction.
- Mutect2 (GATK v4) – Used for analysis of whole genome data.
- Custom code for generating the malignancy continuum is available on GitHub (https://github.com/winstonbecker/scCRC_continuum).

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data has been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE201349. Original data generated in this study are also available on the Human Tumor Atlas Network (HTAN) Data Portal (unaffected FAP tissues, polyps, and CRCs; https://data.humantumoratlas.org/ under the HTAN Stanford Atlas) and the HuBMAP data portal (normal colon tissues; https://portal.hubmapconsortium.org/ under the Stanford TMC). Unique IDs for accessing the HTAN datasets are listed in Supplementary Table 3 and unique IDs for accessing the HuBMAP datasets are listed in Supplementary Tables 4 and 5. Receptor ligand pairs from the Fantom5 database were downloaded from https://fantom.gsc.riken.jp/5/suppl/Ramilowski_et_al_2015/. Clustered TF motifs can be downloaded from https://www.vierstra.org/resources/motif_clustering#downloads. Seurat objects for previously published single-cell colon data were downloaded from https://github.com/cssmillie/ulcerative_colitis. Counts matrices and T-cell annotations for cells from BCC are available on GEO with accession number GSE12381315. TCGA DNA methylation data can be downloaded from the GDC data portal (https://portal.gdc.cancer.gov/).

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 | Sample size was set based on the availability of polyps collected. We aimed to maximize the number of polyps assayed, as we were primarily interested in identifying features of the precancerous state. A smaller number of normal tissues and CRCs were collected to facilitate analysis of the polyps and to help define the precancerous continuum. The sample size in this study was sufficient to define a continuum from normal to cancer and to identify significant changes in composition between disease states. |
| Data exclusions | All datasets generated that did not fail experimentally (e.g. overloaded sample) were included in the study. |
| Replication | Replicate single cell ATAC datasets were generated for 4 samples and produced highly concordant results. Replicates for additional single-cell experiments were not performed as technical replicates are less informative than additional samples with different disease states. Further, using multimodal data (scATAC and snRNA) allows us to highlight results that are concordant between the two methods. Selected findings (e.g. PD1 expression) were also validated with orthogonal assays. |
| Randomization | There was no randomization into experimental groups as experiments on all samples in this study were performed the same way. |
| Blinding | No blinding was performed in this study that focused on deep characterization of polyps, CRCs, and normal tissues at a single point in time. No differential clinical intervention was performed or was being compared in this study. |

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 | Methods |
|----------------------------------|---------|
| Involved in the study | Involved in the study |
| n/a | n/a |
| Antibodies | ChIP-seq |
| Eukaryotic cell lines | Flow cytometry |
| Palaeontology and archaeology | MRI-based neuroimaging |
Antibodies

Antibodies used

- **CD3-BX015 (UCHT1)—Cy5-RX015; Supplier: Akoya Biosciences, Clone: UCHT1; Catalog number: 4350008**
- **PD-1/CD279-BX014 (EH12.2H7)—Atto 550-RX014; Supplier: Akoya Biosciences, Clone: EH12.2H7; Catalog number: 4250010**

Validation

The CD3 and PD-1 antibodies were pre-validated and conjugated by Akoya Biosciences, which were purchased for running the CODEX experiments.

Human research participants

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

### Population characteristics

Patients in this study include patients with FAP seen at Stanford healthcare, 1 patient without FAP undergoing routine screening colonoscopy at Stanford healthcare, patients with CRC with samples deposited in the Stanford Tumor bank, and patients without FAP or colon cancer at WUSTL. Ages of patients included ranged from 20–78. The study included 9 male and 6 female patients. 6 patients were White, 5 were Hispanic/Latino, 1 White/Asian, 1 Black, and 1 Other. Please see the supplemental table for additional information.

### Recruitment

For the FAP patients, given that this is a relatively rare condition, no criteria other than FAP were required for recruitment. Eligible patients undergoing colonoscopy, pouchoscopy, or colectomy were identified in the Stanford Cancer Genetics clinic, the Stanford Gastroenterology and Hepatology service, or the Stanford Adult and Pediatric Surgery service. Eligible patients were notified of their eligibility to participate in research. Over the phone, the description, risks, benefits, and alternatives of participating in the research study were described and a copy of the full consent form was sent to them via email. On the day of the procedure, they met with a clinical research coordinator to answer questions and sign the consent. This recruitment strategy only includes patient’s seen at Stanford healthcare, which is one potential source of selection bias. For polyps from non-FAP patients, they were obtained from patients undergoing screening colonoscopy at Stanford with no attempt to exclude donors on the basis of age, gender, or sex. Healthy controls were recruited at WUSTL, and no attempt was made to exclude healthy donors on the basis of age, gender, or sex. A subset of CRC samples were obtained from the Stanford tumor bank, again with no attempt to exclude patients on the basis of age, gender, or sex.

### Ethics oversight

The study was approved by the Stanford IRB and informed consent was obtained from all patients.

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