

Polyclonal-to-monoclonal transition in colorectal precancerous evolution

<https://doi.org/10.1038/s41586-024-08133-1>

Received: 28 November 2023

Accepted: 27 September 2024

Published online: 30 October 2024



Zhaolian Lu^{1,2,13}, Shanlan Mo^{1,2,13}, Duo Xie^{1,3,13}, Xiangwei Zhai^{4,13}, Shanjun Deng^{4,13}, Kantian Zhou^{1,2}, Kun Wang^{1,5,6}, Xueling Kang^{1,2}, Hao Zhang^{1,2}, Juanzhen Tong^{1,2}, Liangzhen Hou^{1,3}, Huijuan Hu¹, Xuefei Li¹, Da Zhou^{5,6}, Leo Tsz On Lee^{3,7}, Li Liu⁴, Yaxi Zhu⁸, Jing Yu^{9,10}, Ping Lan^{9,10}, Jiguang Wang^{2,11}, Zhen He^{9,12}✉, Xiongfei He⁴✉ & Zheng Hu^{1,2}✉

Unravelling the origin and evolution of precancerous lesions is crucial for effectively preventing malignant transformation, yet our current knowledge remains limited^{1–3}. Here we used a base editor-enabled DNA barcoding system⁴ to comprehensively map single-cell phylogenies in mouse models of intestinal tumorigenesis induced by inflammation or loss of the *Apc* gene. Through quantitative analysis of high-resolution phylogenies including 260,922 single cells from normal, inflamed and neoplastic intestinal tissues, we identified tens of independent cell lineages undergoing parallel clonal expansions within each lesion. We also found polyclonal origins of human sporadic colorectal polyps through bulk whole-exome sequencing and single-gland whole-genome sequencing. Genomic and clinical data support a model of polyclonal-to-monoclonal transition, with monoclonal lesions representing a more advanced stage. Single-cell RNA sequencing revealed extensive intercellular interactions in early polyclonal lesions, but there was significant loss of interactions during monoclonal transition. Therefore, our data suggest that colorectal precancer is often founded by many different lineages and highlight their cooperative interactions in the earliest stages of cancer formation. These findings provide insights into opportunities for earlier intervention in colorectal cancer.

Deciphering the cellular origin and evolution that underlie the initial phases of cancer formation is crucial for efficient early screening and intervention^{1–3}. Nevertheless, precancerous stages often span several years, remaining occult for much of that time^{1,3,5}. This inherent complexity poses a formidable challenge for accurately identifying the earliest events in tumour initiation.

Colorectal cancer (CRC) is the third most prevalent cancer globally and the second leading cause of cancer-related deaths⁶. The well-established adenoma–carcinoma sequence characterizes the development of CRC, providing an ideal model to investigate early events in precancerous stages⁷. Cancer genomic sequencing typically unveils a single founding clone in a malignant tumour⁸. However, clonality analysis using somatic markers indicates that premalignant adenomas⁹ or even malignant CRCs¹⁰ can have polyclonal origins, wherein multiple distinct lineages expand concurrently within the same lesion. Despite these insights, our understanding of the origin and evolutionary dynamics in the earliest stages of lesion formation remains limited.

Prospective lineage tracing using evolving DNA barcodes has revolutionized our approaches to recording cell lineages in unprecedented resolution and high throughput^{11–14}. In this study, we harnessed a base editor-based lineage tracing technique, substitution mutation-aided lineage-tracing⁴ (SMALT), to systematically map the origin and evolution of intestinal precancers in mouse models at single-cell resolution. Through integrated analyses of single-cell phylogenies ($n = 260,922$ single cells from 112 normal and neoplastic samples), single-cell RNA sequencing (scRNA-seq, $n = 43,001$ single cells from 9 neoplastic samples) and a comprehensive cohort of humans with sporadic premalignant polyps (whole-exome sequencing (WES) for $n = 107$ colorectal polyp/CRC pairs, whole-genome sequencing (WGS) for $n = 29$ single glands from 5 colorectal polyps), we identified a common polyclonal-to-monoclonal trajectory in intestinal tumorigenesis in mouse and human. Additionally, our investigation revealed extensive intercellular interactions in early polyclonal lesions, highlighting the intercellular cooperation in early tumorigenesis.

¹Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

²SIAT-HKUST Joint Laboratory of Cell Evolution and Digital Health, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. ³Cancer Centre, Faculty of Health Sciences, University of Macau, Taipa, Macau, China. ⁴MOE Key Laboratory of Gene Function and Regulation, State Key Laboratory of Biocontrol, Innovation Center for Evolutionary Synthetic Biology, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. ⁵School of Mathematical Sciences, Xiamen University, Xiamen, China. ⁶National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China. ⁷Ministry of Education Frontiers Science Center for Precision Oncology, University of Macau, Macau, China. ⁸Department of Pathology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. ⁹Department of General Surgery, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China.

¹⁰Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. ¹¹Division of Life Science, Department of Chemical and Biological Engineering, State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology, Hong Kong SAR, China. ¹²Key Laboratory of Human Microbiome and Chronic Diseases (Sun Yat-sen University), Ministry of Education, Guangzhou, China. ¹³These authors contributed equally: Zhaolian Lu, Shanlan Mo, Duo Xie, Xiangwei Zhai, Shanjun Deng. ✉e-mail: hezh5@mail.sysu.edu.cn; hexiong1@mail.sysu.edu.cn; zheng.hu@siat.ac.cn

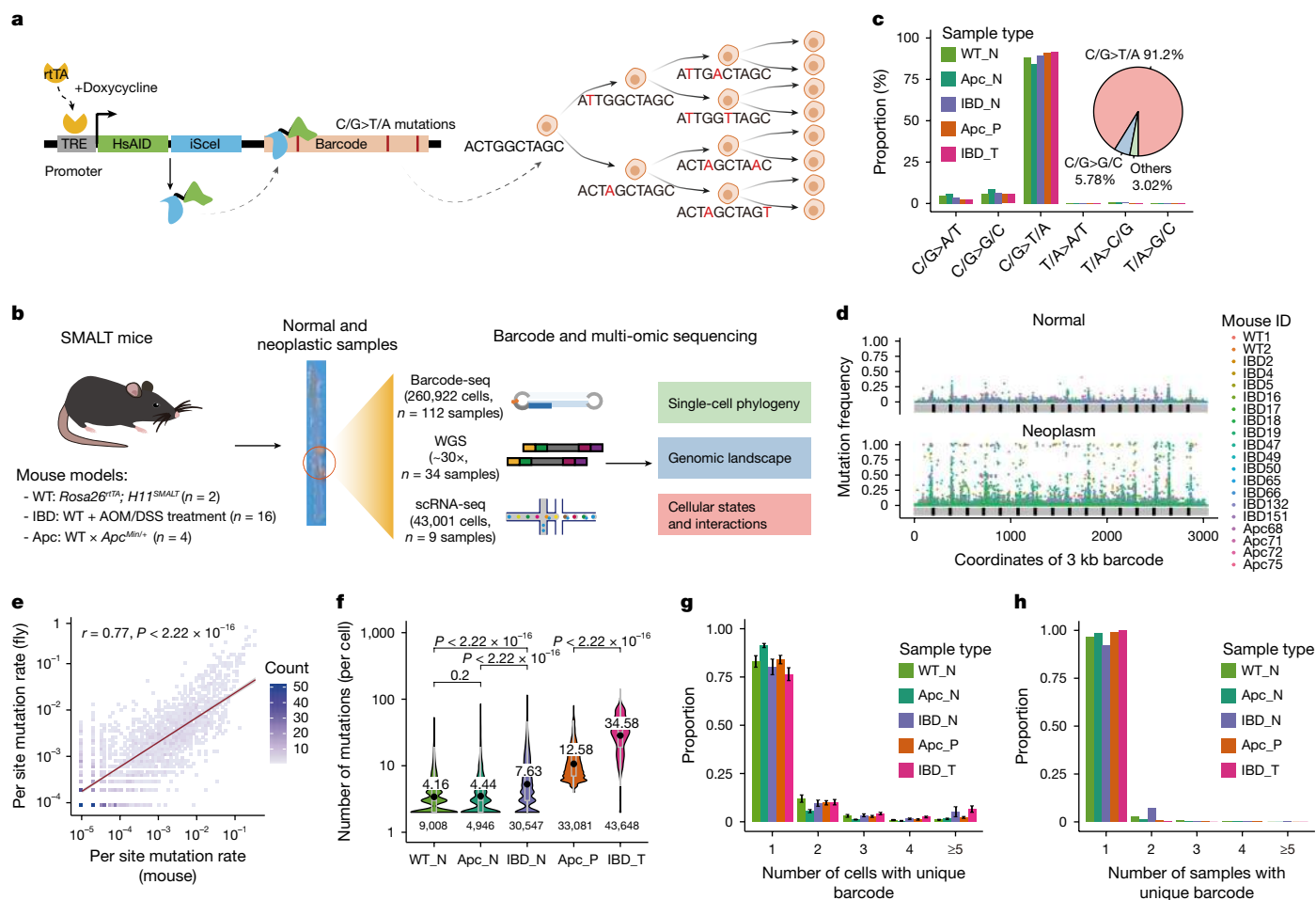


Fig. 1 | SMALT lineage tracing of mouse intestinal tumorigenesis.

a, Schematic of the SMALT lineage tracing system. **b**, Intestinal tumorigenesis with AOM/DSS or *Apc*^{Min/+} mice carrying the engineered SMALT system in the germline. Normal and neoplastic samples were collected for long-read sequencing of lineage barcodes (barcode-seq), WGS and scRNA-seq. WT, wild type. **c**, The relative proportions of distinct substitution types in barcodes across all samples. *Apc*_N, normal small intestine in *Apc*^{Min/+} mice; *Apc*_P, polyps in *Apc*^{Min/+} mice; *IBD*_N, normal tissue in AOM/DSS mice; *IBD*_T, neoplasms in AOM/DSS mice; *WT*_N, wild-type normal colon. **d**, Per site mutation frequency on barcodes across all samples. **e**, Correlation of per site mutation frequency

between mouse and fruit fly. Pearson's r and P value are shown. **f**, Violin plot showing the number of barcode mutations per cell in different tissues. The mean number of mutations and the number of cells are shown. P values are by two-sided Wilcoxon rank-sum test. **g**, The proportion of unique barcode alleles and the number of cells with the unique barcode in *WT*_N ($n = 4$), *Apc*_N ($n = 4$), *IBD*_N ($n = 19$), *Apc*_P ($n = 21$) and *IBD*_T ($n = 30$), respectively. Data are mean $\pm$ s.e.m. **h**, The proportion of unique barcode alleles and the number of samples with the unique barcode. In box plots, the horizontal line is the median, the box delineates the 25th to 75th centiles, and whiskers extend to 1.5 times the interquartile range.

SMALT lineage tracing and mouse models

The SMALT lineage tracing system consists of three key components⁴: HsAID, iScel (an inactive variant of the homing nuclease I-SceI) and a 3-kb DNA barcode (Fig. 1a). HsAID is an optimized *Homo sapiens* activation-induced cytidine deaminase with 10 amino acid substitutions relative to the wild type, which induces cytosine (C)-to-thymine (T) mutations around 30-fold more rapidly than the wild type⁴. Notably, the deaminated C (uracil (U)) can be converted to T only after DNA replication, which means that the number of C-to-T barcode mutations measures the relative number of cell divisions. HsAID is fused to iScel, which binds specifically to an 18-bp DNA motif, and thereby guides HsAID to target regions within the barcode (Fig. 1a and Extended Data Fig. 1a,b). The expression of HsAID–iScel is induced by doxycycline.

We induced intestinal neoplasms based on two classic mouse models, azoxymethane (AOM)/dextran sodium sulfate (DSS) and *Apc*^{Min/+}, in the SMALT mice (C57BL/6 J background), (Extended Data Fig. 1c–e and Methods). The AOM/DSS model is widely used to study CRC induced by inflammatory bowel disease (IBD) such as Crohn's disease or ulcerative colitis¹⁵. Histopathological examination indicated that many

neoplasms induced from a standard protocol were adenocarcinomas (Supplementary Figs. 1–3). However, WGS demonstrated that these had stable chromosomal karyotypes, with few instances of aneuploidy or copy number alterations (Supplementary Fig. 4). Therefore, AOM/DSS tumours resemble early-stage neoplasms in human CRCs. *Apc*^{Min/+} mice spontaneously develop multiple intestinal neoplasms owing to the loss of *Apc*, and are a good animal model for APC-driven intestinal tumorigenesis such as human familial adenomatous polyposis (FAP). WGS showed that the polyps that developed in *Apc*^{Min/+} mice lacked chromosomal aberrations, showing that they also represent an early stage in tumorigenesis (Supplementary Fig. 4).

We performed targeted barcode sequencing for $n = 30$ AOM/DSS neoplasms (designated *IBD*_T), 17 *Apc*^{Min/+} polyps (*Apc*_P), 26 samples from healthy colon, inflamed colon or *Apc*^{Min/+} small intestine, and 18 samples from 6 unaffected organs in 3 SMALT mice. We also performed WGS for $n = 11$ *IBD*_T samples and 4 *Apc*_P samples with matched healthy samples and scRNA-seq for $n = 9$ *IBD*_T samples (Supplementary Table 1). These comprehensive multiomics datasets allowed us to systematically delineate the single-cell phylogenies and genomic and microenvironmental landscapes in early stages of colorectal tumorigenesis (Fig. 1b).

High diversity of SMALT barcodes

We used high-fidelity PacBio long-read sequencing to recover the full length of the target 3-kb barcode through a unique molecular identifier (UMI)-based library preparation protocol (Supplementary Fig. 5). Since each cell contains only one copy of the 3-kb barcode in its genome, a PacBio consensus read represents a single cell, despite the library being prepared with bulk samples. Barcode mutations were detected using an in-house bioinformatics pipeline with sophisticated UMI correction and artefact filtering (Supplementary Fig. 6 and Methods).

As expected, more than 90% of mutations were indeed C/G-to-T/A, followed by C/G-to-G/C (5.78%) and others (3.02%) (Fig. 1c). Mutations were widely distributed across the entire 3-kb barcode (Fig. 1d), with increased mutation rates downstream of iSceI-binding motifs (Supplementary Fig. 7a). Of note, an average of 836 mutated sites (95% confidence interval, 752–920; Supplementary Fig. 7b) were observed across all cells in one sample, far surpassing the number that can be detected by CRISPR–Cas9 lineage tracing methods (typically 10–60 mutable sites)^{11,12,14}. Of note, the per site mutation rate in our mouse data correlated well with the rate detected in *Drosophila* embryos⁴ (Fig. 1e). Moreover, analysis of the mutational frequencies in normal cells identified 14 hotspot sites (Supplementary Fig. 7c). The somatic mutational spectrum in exonic regions of our AOM/DSS neoplasms from SMALT mice was highly similar to that observed in the SMALT-free AOM/DSS neoplasms¹⁶ (Supplementary Fig. 8), indicating minimal off-target editing.

Normal cells from lesion-adjacent tissues (Fig. 1f) or other organs (such as blood, liver and lung; Supplementary Fig. 9) exhibited lower barcode mutation counts than neoplastic cells. This was also true for CD45⁺ tumour-infiltrating immune cells (Extended Data Fig. 2a,b and Supplementary Fig. 10). These findings indicate that neoplastic cells can be effectively separated from normal cells on the basis of barcode mutation burden (Extended Data Fig. 2c, Supplementary Fig. 11 and Methods). After filtering, we observed an average of 4.16 to 34.58 barcode mutations per cell across different sample types (Fig. 1f and Supplementary Table 2). Notably, in both normal and tumour samples, approximately 90% of cells exhibited a unique combination of mutations, indicating extremely high diversity of the SMALT lineage barcodes (Fig. 1g,h). Together, these data demonstrated high *in vivo* mutagenesis activity and high lineage barcoding capacity of the SMALT lineage tracing system.

Of note, AOM/DSS lesions (IBD_T) showed significantly higher mutation burdens than *Apc*^{Min/+} lesions (*Apc*_P) (Fig. 1f), despite a similar sampling time (around 30 weeks postnatal). We observed a 2.8-fold increase in barcode mutation burden in neoplastic cells versus normal cells in the *Apc*^{Min/+} model, and a 4.3-fold increase over normal cells in the AOM/DSS model (Fig. 1f). CRC cells have been estimated to undergo around two times as many mitoses as adjacent normal cells¹⁷. Thus, we estimate the mutation rate per cell division in neoplastic cells to be approximately 1.4-fold higher than normal cells in the *Apc*^{Min/+} model and 2.15-fold higher than normal cells in the AOM/DSS model. The greater increase of mutation rate in AOM/DSS lesions is probably owing to AOM being a mutagen that also induces C/G>T/A transitions, as HsAID does in the SMALT system.

The origin of inflammation-driven lesions

The origin of a tumour, whether it derives from a single progenitor (monoclonal origin) or from multiple different progenitors (polyclonal origin, also called multi-ancestral origin¹⁸), is a fundamental question in cancer research. Although previous studies have shown that both scenarios can occur in colorectal precancerous lesions^{19–21}, their prevalence, temporal dynamics and mechanisms remain largely unknown. We used the highly diverse SMALT barcodes to reconstruct single-cell phylogenies for 30 AOM/DSS neoplasms along with their adjacent normal cells (Methods), each including 375 to 9,180 (mean = 3,141) single cells per sample (Fig. 2a,b and Extended Data Fig. 3). Bootstrapping

analysis demonstrated the high robustness of phylogenetic reconstruction (bootstrapping confidence >85%; Fig. 2c). Compared with CRISPR–Cas9-based lineage trees^{12,14}, our SMALT trees exhibited 3.3 times more internal branching events, indicating the superior resolution of SMALT for phylogenetic mapping (Fig. 2d).

Monoclonality was characterized by the presence of a single dominant monophyletic clade of neoplastic cells (Fig. 2a and Extended Data Fig. 3) with shared clonal mutations (Supplementary Fig. 12a and Methods). Conversely, polyclonality was identified when neoplastic cells were dispersed into multiple phylogenetic clades and intermixed with normal cells (Fig. 2b and Extended Data Fig. 3), with no clonal mutations observed (Supplementary Fig. 12b). We found that the majority (66.7% (20 out of 30)) of AOM/DSS neoplasms were polyclonal (Extended Data Fig. 3). This indicates that parallel expansions of multiple distinct cell lineages in the same physical niche are common in inflammation-driven intestinal tumorigenesis.

The number of founding progenitors

The high-resolution single-cell phylogenies enabled us to quantify the number of progenitors that founded each neoplasm (denoted as *N_p*) using targeting coalescent analysis²² (TarCA) (Methods). In this method, a progenitor is defined as an ancestral cell that is capable of founding a clonally expanding population in the lesion (Fig. 2e and Supplementary Fig. 13), regardless of the underlying mechanisms of clonal expansions. As expected, *N_p* was estimated to be approximately 1 for each of the monoclonal lesions (Fig. 2e). For polyclonal lesions, *N_p* varied from 2 to 33, and these estimates were robust to downsampling of cells (Fig. 2e) and hotspot mutation events (Supplementary Fig. 13b). Because many founding lineages may be lost during growth owing to random drift and competition, the actual number of progenitors is likely to be higher than our estimates. This suggests that the early tumorigenesis is likely to be driven by multiple independent progenitors.

Monoclonal lesions exhibited significantly more barcode mutations than polyclonal ones (Fig. 2f). Because the AID-induced C-to-T mutations only occur after DNA replication, this suggests that a monoclonal lesion has, on average, undergone more cell divisions. In line with this, WGS also revealed a significantly higher genome-wide mutation burden in monoclonal lesions compared to polyclonal lesions (Fig. 2g and Supplementary Table 3). Crucially, putative driver mutations were also more common in monoclonal lesions (Fig. 2h). These data together suggest that monoclonal lesions may represent a more advanced stage of the colorectal tumorigenesis (Fig. 2e).

Clonal expansion and selection

To quantify clonal expansion in tissues, we defined a clonal expansion score that measures the phylogenetic similarity of two random cells (Supplementary Fig. 14a and Methods). As expected, normal colon epithelium had a low level of clonal expansions. Of note, inflamed colons exhibited greater clonal expansion than wild-type normal colons (Fig. 2i,j and Supplementary Fig. 14b). These mouse data substantiated earlier observations of significant clonal expansions in inflamed colon of patients with IBD²³. Monoclonal lesions showed greater expansion than polyclonal lesions in the mice (Fig. 2i), and a metric of proliferative fitness showed the same trend (Fig. 2k). Additionally, among putative driver genes in the WGS data, monoclonal lesions showed an increased ratio of nonsynonymous to synonymous mutation rates²⁴ (dN/dS ratio) compared with polyclonal lesions, especially for nonsense and splicing mutations (Supplementary Fig. 14c). This indicates that there is stringent selection during somatic evolution towards monoclonal transition. Spatial computational inference based on agent-based tumour simulations and approximate Bayesian computation²⁵ (Supplementary Fig. 15) revealed that the monoclonal AOM/DSS tumours (10 out of 30) also underwent strong subclonal selection even after monoclonal

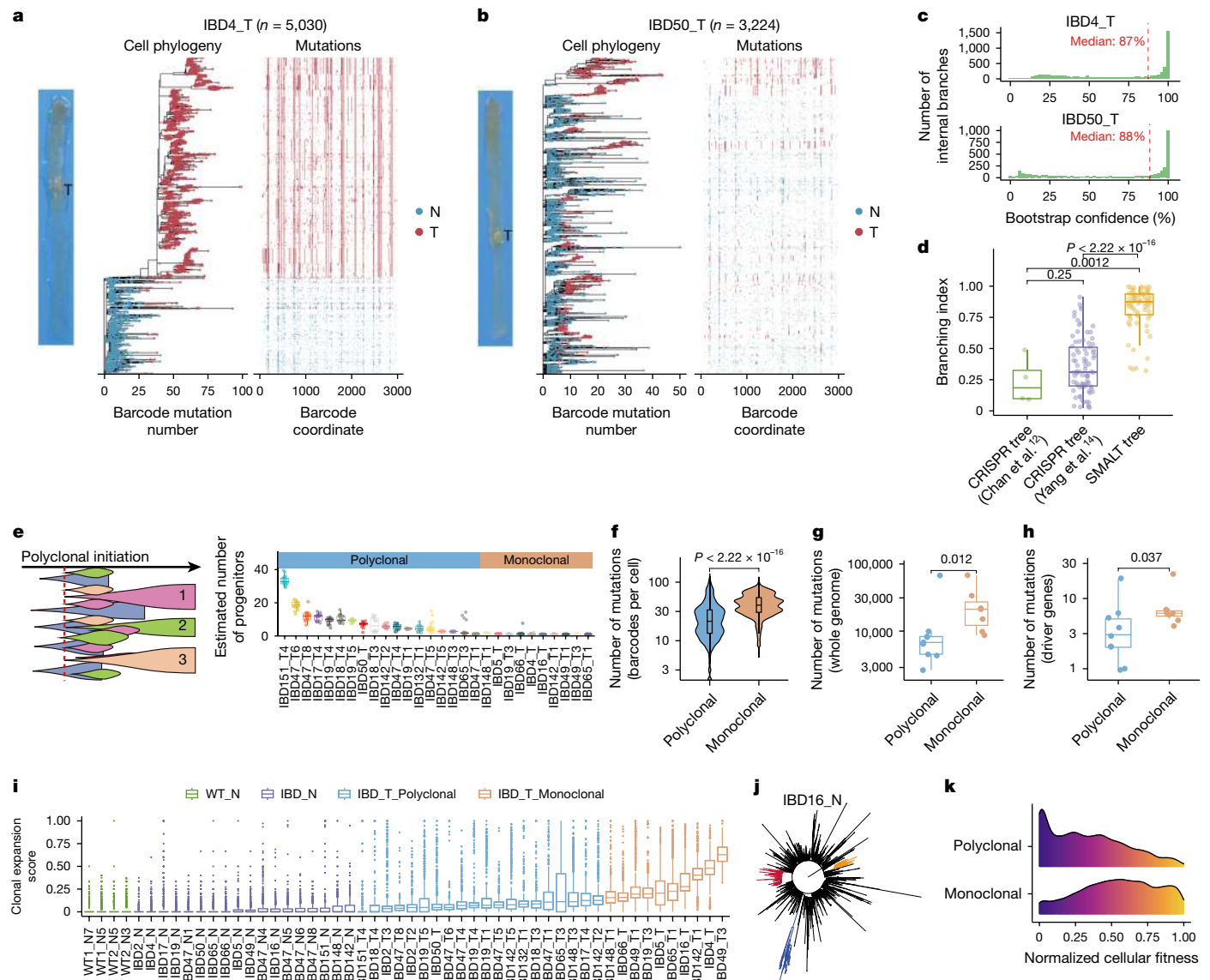


Fig. 2 | Single-cell phylogenies reveal the origin of inflammation-driven neoplasms. a, b, Single-cell phylogeny (left) and corresponding barcode mutations (right) for a representative monoclonal lesion (**a**; lesion IBD4_T) and a representative polyclonal lesion (**b**; lesion IBD50_T). **c**, Bootstrapping values of the phylogenetic tree for IBD4_T (top) and IBD50_T (bottom). **d**, The branching index of SMALT trees in this study ($n = 77$) compared with CRISPR-Cas9 lineage trees from two previous studies (ref. 12, $n = 4$; and ref. 14, $n = 85$). **e**, The number of founding progenitors (Np) estimated from single-cell phylogeny. For each lesion, Np was estimated 20 times using downsampled

cells. **f**, The barcode mutation count per cell in monoclonal ($n = 22,766$ cells) versus polyclonal ($n = 20,882$ cells) lesions. **g, h**, Total somatic mutation burden (**g**) or putative driver mutation burden (**h**) in WGS data of monoclonal ($n = 7$) versus polyclonal ($n = 9$) lesions. **i**, Clonal expansion scores calculated using 1,000 downsampled cell pairs ranked by the median clonal expansion scores within each sample type. **j**, A representative single-cell phylogeny for inflamed normal colon. Lineages exhibiting clonal expansions are highlighted in colour. **k**, Single-cell fitness scores in monoclonal versus polyclonal lesions. **d, f, h**, P values by two-sided Wilcoxon rank-sum test.

transition (selection coefficient (s) = 0.57–0.87, Supplementary Fig. 16). In addition, the inference by approximate Bayesian computation estimated the barcode mutation rate (u) in these inflammation-driven neoplasms to be 0.12–0.55 per cell division in the entire 3-kb barcode.

Polyclonal origin of lesions in *Apc*^{Min/+} mice

We next applied SMALT to *Apc*^{Min/+} mice. We sampled multiple polyps from each of four *Apc*^{Min/+} mice (Extended Data Figs. 1e and 4a). The integrated single-cell phylogenies (Supplementary Fig. 17a, b) with multiple polyps from each mouse exhibited highly intermixed lineage relationships, indicating that the phylogenies could not separate individual polyps. Despite the overall low genetic similarities among cells both within and between polyps (Supplementary Fig. 17c, d), cells

with a high barcode mutation burden (>20) showed higher intra-polyp similarity relative to inter-polyp similarity (Supplementary Fig. 17e–h). These data reveal high genetic diversity within a polyp, with evidence of clonal expansions in some lineages.

Of note, all individual polyps ($n = 17$) exhibited polyclonal origins (Extended Data Fig. 4a). Two regions (P5-1 and P5-5) located at the head and tail of a bar-shaped polyp underwent stronger but independent clonal expansions, as indicated by their monophyletic subtrees (Extended Data Fig. 4b, c), the mutational cellular frequencies in the single cells (Extended Data Fig. 4d) and their higher proliferative fitness (Extended Data Fig. 4e). The number of founding progenitors for each polyp was estimated to range from 4 to approximately 100 (Extended Data Fig. 5), indicating even higher polyclonality than that in AOM/DSS-induced lesions.

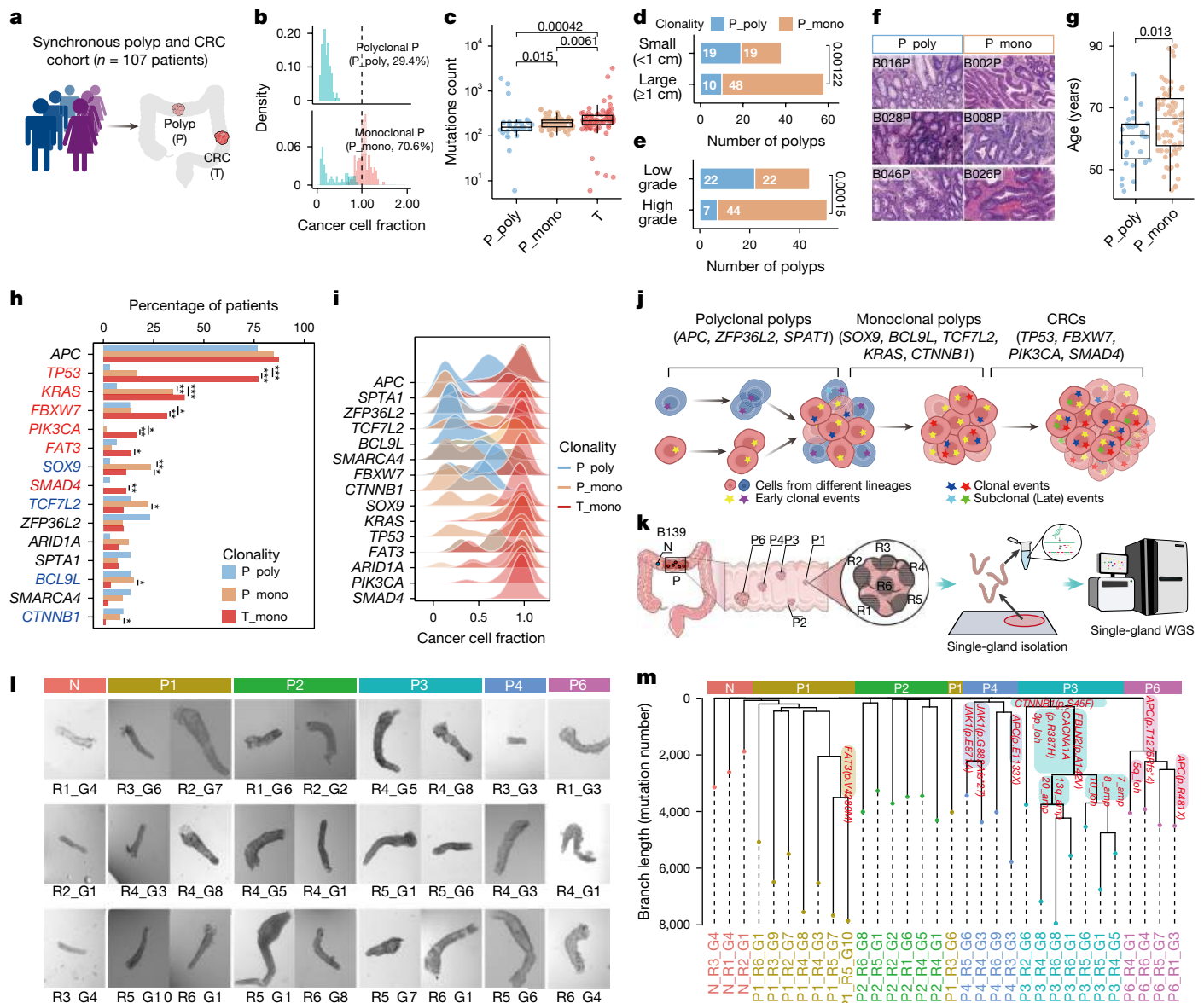


Fig. 3 | Polyclonal-to-monoclonal transition in human sporadic polyps. **a**, A human cohort with sporadic premalignant polyps, including 107 patients with synchronous polyps and CRC. **b**, The distribution of CCFs reveals the clonality of each lesion. One representative polyclonal polyp (P_poly, B046P) and one monoclonal polyp (P_mono, B002P) are shown. **c**, The total somatic mutation burden in P_poly ($n = 30$), P_mono ($n = 72$) or CRCs ($n = 86$) after removing samples with low purity (<0.25). **d**, Distribution of small (<1 cm) and large (≥ 1 cm) polyps. **e**, Distribution of low-grade and high-grade dysplasias. **f**, Representative images of haematoxylin and eosin (H&E) staining. Scale bar, 100 μ m. **g**, Age distribution of participants. **h**, Percentage of patients carrying indicated putative driver mutations. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, one-sided

Fisher's exact test. Highlighted genes have Benjamini–Hochberg FDR < 0.1 . **i**, The CCFs of putative driver genes. **j**, Schematic of polyclonal-to-monoclonal transition in the somatic evolution of premalignant polyp and its subsequent malignant transformation. **k**, Schematic of WGS of single glands from five premalignant polyps (P1–P4 and P6) in a patient with sporadic polyps (B139). N, normal tissue; R, regions within polyps. **l**, Images of individually isolated glands. **m**, Integrated phylogenetic tree including 3 normal glands and 29 neoplastic glands from 5 polyps. Putative driver mutations shared by multiple glands are labelled. P values by two-sided Wilcoxon rank-sum test (**c**, **g**) or Fisher's exact test (**d**, **e**). Graphics in **a**, **k** adapted from Servier Medical Art (CC BY 4.0).

We developed a method to estimate the timing of progenitors, which represents the start of clonal expansions for polyp initiation (Extended Data Fig. 5 and Supplementary Notes). The average timing of progenitors ranged from 59 to 130 (mean = 95) postnatal days for mouse *Apc68* and 14 to 70 (mean = 47) postnatal days for mouse *Apc72* (Extended Data Fig. 5g). Given that epithelial cells in mouse intestine divide approximately 1.5 times every day²⁶, we were able to estimate the number of cell divisions at the time of polyp initiation, which was 88–196 (mean = 142) postnatal divisions for *Apc68* and 21–105 (mean = 70) postnatal divisions for *Apc72* (Supplementary Fig. 18). Furthermore, given that the cell cycle time of human colonic epithelial cells is approximately 5 days²⁷, these numbers of cell divisions (21–196

divisions in two mice) roughly corresponded to ages of 105–980 days (0.3–2.7 years) for human patients with FAP. Therefore, our timing analysis suggests that neoplastic initiation in patients with FAP is likely to occur as early as in infancy.

Origin and evolution of human colorectal polyps

We also collected WES (mean depth $> 200\times$) data for a cohort of 107 treatment-naïve patients with sporadic premalignant polyps and synchronous CRCs (Fig. 3a, Supplementary Table 4 and Methods). Somatic single nucleotide variants (SSNVs), short insertion–deletions (indels) and somatic copy number alterations (SCNAs) were detected, and the

mutational cancer cell fractions (CCFs) were estimated by correcting the variant allele frequency (VAF) for tumour purity, ploidy and copy number variations^{25,28} (Supplementary Fig. 19, Supplementary Tables 5 and 6 and Methods).

Monoclonality or polyclonality of each lesion can be inferred by determining whether clonal SSNVs (those with CCF values of around 1) are present or absent (Fig. 3b). Here, 5 polyps and 21 tumour samples were removed in this analysis owing to low purity (<0.25). We found that polyclonality was more common in polyps (29.4% (30 out of 102)) than in CRCs (8.1% (7 out of 86)) (Fig. 3b and Supplementary Figs. 20 and 21). Fewer SSNVs and SCNAs were observed in polyclonal polyps than in monoclonal polyps, both of which exhibited lower mutational burdens than CRCs (Fig. 3c, Extended Data Fig. 6 and Supplementary Fig. 22). Notably, polyclonal polyps were relatively smaller in size (Fig. 3d) and more commonly exhibited low-grade dysplasia according to their pathological features (Fig. 3e,f and Methods). Furthermore, patients with polyclonal polyps were younger at the time of diagnosis (mean age = 60) than patients with monoclonal polyps (mean age = 66) (Fig. 3g). As all samples were from treatment-naïve patients, our results were not affected by treatment. Together, these findings support the notion that monoclonality signifies a more advanced stage of colorectal tumorigenesis, aligning with observations from lineage tracing studies in mouse models (Fig. 2e–h).

In our human cohort (Fig. 3a), the driver genes *TP53* (77.2% versus 12.7%), *KRAS* (40.5% versus 26.5%), *PIK3CA* (16.5% versus 1%), *FBXW7* (31.6% versus 13.7%) and *SMAD4* (11.4% versus 1%) were more commonly mutated in CRCs than in polyps (Fig. 3h and Supplementary Table 7). Of note, *KRAS* was frequently mutated in monoclonal polyps (34.7%) and CRCs (40.5%), but seldom mutated in polyclonal polyps (6.7%), suggesting that *KRAS* mutations confer a selective advantage for the clonal outgrowth and might be conducive for polyclonal-to-monoclonal transition. As expected, lower clonality of driver mutations was observed in polyclonal polyps compared with monoclonal polyps and CRCs (Fig. 3i).

We also found that *BCL9L*, *SOX9*, *TCF7L2* and *CTNNB1* were more frequently mutated in polyps than in CRCs (Fig. 3h). Notably, *SOX9* has also been found to be more frequently mutated in gastric intestinal metaplasia than in malignant gastric cancer²⁹, indicating *SOX9* mutation is preferentially selected at premalignant stage. Loss of *BCL9* or *BCL9L* has been found to suppress tumorigenesis, possibly owing to a shift in the gene expression of *APC*-mutant adenomas from stem cell-like towards differentiated³⁰. These findings suggest that the drivers of clonal expansion in precancerous stages might potentially impede malignant transformation, although the underlying mechanisms remain unclear. Finally, the overall selective strength for driver mutations, measured by dN/dS²⁴ (Methods), was higher in monoclonal polyps than in polyclonal polyps (Supplementary Fig. 23). Collectively, our genomic analysis of human premalignant polyps and malignant CRCs validates the findings observed in mouse models, in which early tumorigenesis in precancerous stages follows a trajectory of polyclonal-to-monoclonal transition (Fig. 3j).

Polyclonal origin validated by single-gland WGS

To take the advantage of glandular structures in colorectal lesions, as each of which represents a clonal cell population^{31,32}, we collected an additional patient with sporadic premalignant polyps (B139, male, 73 years old) and performed WGS (mean depth approximately 21x) for 29 neoplastic glands isolated from 5 separate polyps (4–8 single glands each) and 3 adjacent normal crypts (Fig. 3k,l, Supplementary Tables 4 and 8 and Methods). The SSNV burden in neoplastic glands was approximately twice than that in normal crypts (Extended Data Fig. 7). Phylogenetic analysis revealed that only two polyps (P3 and P6) contained a significant number of SSNVs and driver mutations (*CTNNB1* in P3 and *APC* in P6) in the trunk, indicating their monoclonal origin independently (Fig. 3m). However, in the other three polyps (P1, P2

and P4), the most recent common ancestor of glands from each polyp was near the phylogenetic root (Fig. 3m). Notably, no putative driver mutations were observed in the trunk, indicating the polyclonal origins of each lesion. Partially shared driver mutations were observed in two polyclonal polyps: P1 (*FAT3* in 2 out of 8 glands) and P4 (*JAK1* in 2 out of 4 glands), suggesting greater clonal expansion of certain lineages within a polyclonal lesion. Therefore, our single-gland genomic analysis clearly confirmed that polyclonal origins are also common in human sporadic premalignant polyps. Furthermore, given that 62% (18 out of 29) of the neoplastic glands exhibited non-*APC* driver mutations (for example, *CTNNB1*, *JAK1* or *FAT3*), the very early driver of clonal expansion in many human sporadic polyps is not *APC*.

Evolution of cell states and interactions

To further dissect the temporal cell-state dynamics and intercellular interactions that might underlie the polyclonal-to-monoclonal transition, we performed scRNA-seq for 9 AOM/DSS colorectal neoplasms whose SMALT barcodes had been sequenced (Supplementary Table 1).

In total, we obtained 45,620 high-quality cells, including 2,619 additional normal colon cells from a previous study³³ that we integrated to the current dataset. Eight main cell types were identified (Extended Data Fig. 8a–c, Supplementary Fig. 24a and Supplementary Tables 9 and 10). We observed marked increases of macrophage, neutrophil and endothelial cells and decreases of neoplastic epithelial cells as lesion clonality increased (Extended Data Fig. 8d,e). Here the clonality of a lesion was measured as 1/Np (Fig. 2e). Notably, epithelial cells and macrophages exhibited the highest transcriptomic diversity, which could be further classified into 10 and 7 fine-tuned subclusters, respectively (Fig. 4a, Extended Data Fig. 8f, Supplementary Fig. 24b and Supplementary Tables 11 and 12). Differential abundance analysis revealed substantial depletion of goblet and ‘stem-like’ epithelial cells and enrichment of *Trem2*⁺ and *Chil3*⁺ macrophages with increasing lesion clonality (Fig. 4b and Supplementary Fig. 24c,d). Supporting the notion that monoclonality represents a more advanced stage, we found that neoplastic epithelial cells in high-clonality lesions (Np ≤ 3) exhibited upregulation of cancer-associated molecular hallmarks, including those associated with MYC targets V1 and V2, *KRAS* signalling, epithelial–mesenchymal transition and inflammatory response (Extended Data Fig. 8g). Further, according to the unifying nomenclature of tumour-associated macrophages (TAMs) based on single-cell transcriptomics³⁴, *Trem2*⁺ macrophages were highly overlapped with lipid-associated TAMs (LA_TAMs) and immune regulatory TAMs (Reg_TAMs) (Fig. 4c, Supplementary Fig. 25 and Supplementary Tables 13 and 14), which have an immunosuppressive role in tumour microenvironments^{33,35}. Our data thus suggested that an immune-suppressive microenvironment potentially contributes to the monoclonal transition.

Cell–cell communication analysis based on ligand–receptor interactions using CellChat³⁶ revealed a marked elevation of ligand–receptor interactions between epithelial subtypes in early polyclonal lesions relative to normal colons and late monoclonal lesions (Fig. 4d,e and Supplementary Fig. 26). By comparing ligand–receptor interactions among neoplastic cells in early lesions to those in late lesions, we identified 14 ligand–receptor interactions that were significantly enriched (FDR < 0.05) in early polyclonal lesions (Fig. 4f, Extended Data Fig. 9, Supplementary Fig. 27 and Supplementary Table 15). Genes from 13 out of 14 ligand–receptor interactions were involved in extracellular matrix (ECM) organization and cell adhesion, 11 of which represented epithelial cell-derived components of the basement membrane, laminins³⁷, and their receptors. These results were validated using an orthogonal method of cell–cell communication analysis (Supplementary Fig. 28 and Methods). Notably, about 40% of the ligand–receptor interactions that were enriched in early lesions were contributed by *Krt20*⁺ neoplastic cells (Extended Data Fig. 10a,b). In addition to the overexpression of ECM genes, *Krt20*⁺ neoplastic cells also exhibited strong

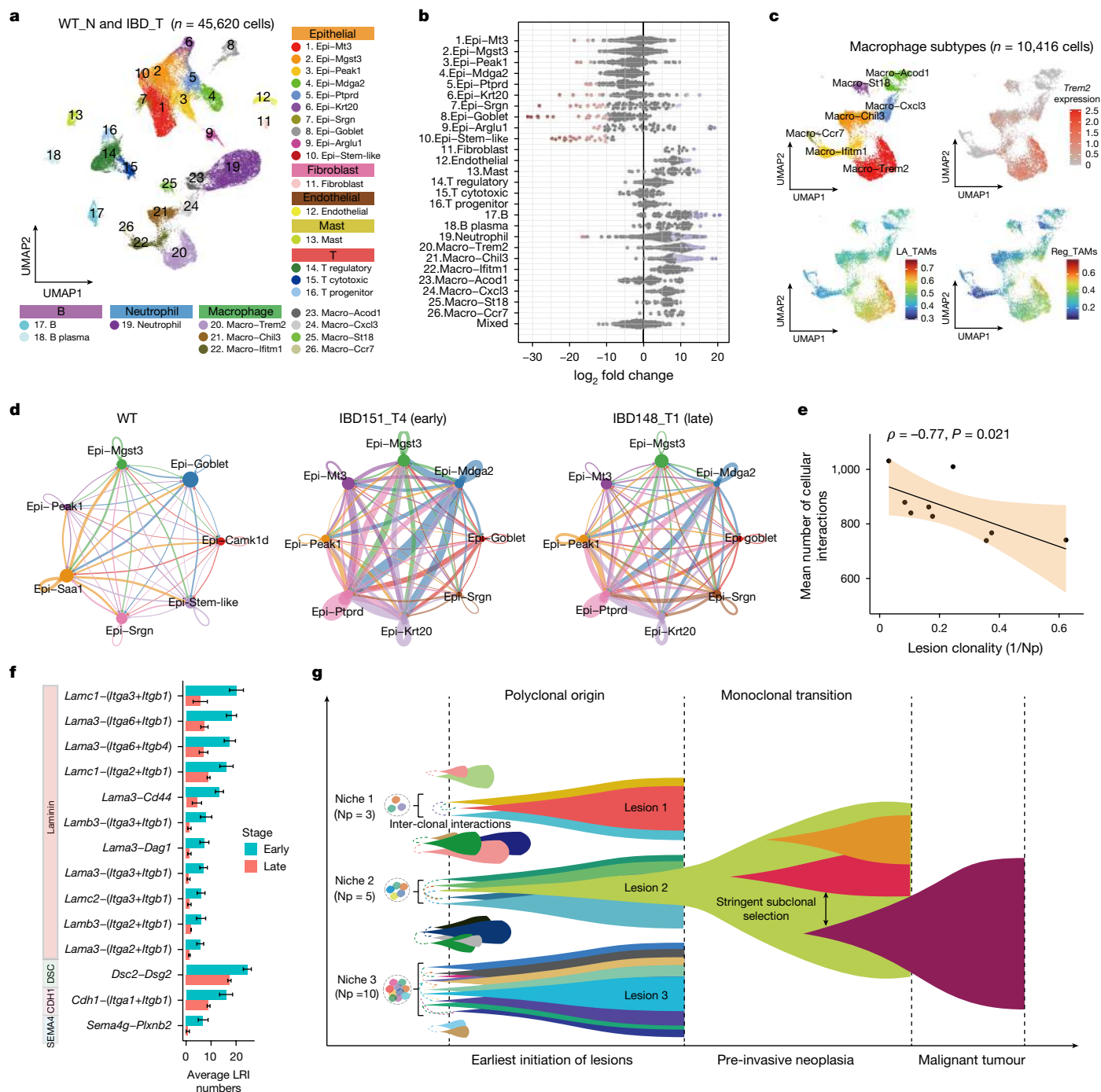


Fig. 4 | Intercellular interactions and polyclonal-to-monoclonal evolution model. a, scRNA-seq identifies 26 cell subclusters. **b**, Beeswarm plot of differential abundance for the cell subclusters along the increase of lesion clonality (measured by 1/Np). Each point represents a neighbourhood that contains a group of cells with similar transcriptomes. Cell neighbourhoods with spatial FDR < 0.1 are highlighted in red for decreased abundance and blue for increased abundance. **c**, Subclusters of macrophages and the expression of *Trem2*, signatures of LA_TAMs and Reg_TAMs. **d**, Cell-cell communication between neoplastic epithelial subclusters inferred by CellChat. The nodes represent epithelial subclusters. The thickness of edges represents the average number of ligand-receptor interactions between every two subclusters from the 50 downsamplings (Methods). **e**, Correlation between lesion clonality (1/Np)

and the average number of ligand-receptor interactions. Spearman's ρ and P value are shown. **f**, A total of 14 ligand-receptor interactions significantly enriched (FDR < 0.05) in the early polyclonal lesions (Np > 3) relative to late lesions (Np ≤ 3) belong to 4 pathways: laminin, desmocollin (DSC), CDH1 and SEMA4. Data are mean ± s.e.m. **g**, Schematic of polyclonal origins and monoclonal transition in early intestinal tumorigenesis. Each precancerous lesion is founded by many lineages undergoing parallel clonal expansions and strong inter-clonal interactions. The gradual loss of inter-clonal interactions and microenvironmental changes might facilitate subsequent polyclonal-to-monoclonal transition. Subclonal selection remains stringent after monoclonal transition, where malignant transformation requires a further clonal sweep.

pro-inflammatory characteristics, such as the production of interleukins IL-1, IL-6, IL-8 and IL-13, response to IFN β , and antigen processing and presentation. (Extended Data Fig. 10c). In summary, these results

strongly suggest that extensive intercellular interactions, probably via ECM organization and cell adhesion, occur early in inflammation-driven intestinal tumorigenesis. Future studies are warranted to elucidate how

these intercellular interactions promote neoplastic growth at the early stage of intestinal tumorigenesis.

Discussion

Polyclonal origins of premalignant lesions have been recognized in many tissues⁹. However, the molecular and evolutionary mechanisms that underlie this cryptic process are still unknown (Fig. 4g). Here our systematic single-cell phylogenetic mapping indicated that each colorectal lesion was typically initiated by several to tens of distinct clonally expanding lineages (Fig. 4g). This prompts a crucial question—whether polyclonal initiation involves positive inter-clonal interactions, such as clonal cooperation^{21,38–40}, that facilitate the overall neoplastic growth. The remarkably high number of initiating clones in our data supports the recruitment model involving short-range interactions³⁸ (Extended Data Fig. 10d), because random collisions by vast numbers of clones in the same physical location is statistically implausible. Our genome analysis of individual premalignant colorectal glands demonstrated that many neoplastic cells do not have canonical driver mutations (such as *APC* or *KRAS*). This suggests that these cells might be influenced or recruited by neighbouring clones with driver gene mutations. Supporting this hypothesis, Reeves et al.⁴¹ demonstrated that a *Hras*-mutant initiating clone can recruit neighbouring wild-type epithelial cells during the development of squamous carcinoma. The recruitment might be selectively favoured as it can alleviate the documented growth barrier of small populations, known as the Allee effect⁴². Our scRNA-seq data revealed increased intercellular communications in early inflammation-driven polyclonal lesions, probably via mechanisms of ECM organization and cell adhesion. Together, our study underscores the crucial role of intercellular interactions and communications in the initiating phase of tumorigenesis.

Collectively, our study has constructed a comprehensive lineage map for early colorectal tumorigenesis, providing a conceptual framework for understanding the origin and evolution of tumour initiation. The insights provided here are pivotal for developing intervention strategies to prevent tumour formation by targeting intercellular interactions. Future research should focus on unravelling the molecular crosstalk within the microenvironment of these interactions by linking cell states (for example, using single-cell multiomics) and lineage information in the same cells¹³. Additionally, efforts are needed to formulate a predictive model that can forecast cancer risk based on the molecular and evolutionary features of premalignant lesions^{43,44}.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-08133-1>.

- Curtius, K., Wright, N. A. & Graham, T. A. Evolution of premalignant disease. *Cold Spring Harb. Perspect. Med.* **7**, a026542 (2017).
- Rane, J. K., Frankell, A. M., Weeden, C. E. & Swanton, C. Clonal evolution in healthy and premalignant tissues: implications for early cancer interception strategies. *Cancer Prev. Res.* **16**, 369–378 (2023).
- Zhang, S. et al. Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets. *Signal Transduct. Target. Ther.* **9**, 149 (2024).
- Liu, K. et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat. Methods* **18**, 1506–1514 (2021).
- Curtius, K., Wright, N. A. & Graham, T. A. An evolutionary perspective on field cancerization. *Nat. Rev. Cancer* **18**, 19–32 (2018).
- Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. Colorectal cancer. *Lancet* **394**, 1467–1480 (2019).
- Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
- Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
- Parsons, B. L. Multiclonal tumor origin: evidence and implications. *Mutat. Res. Rev. Mutat. Res.* **777**, 1–18 (2018).

- Chen, B. et al. Very large hidden genetic diversity in one single tumor: evidence for tumors-in-tumor. *Natl. Sci. Rev.* **9**, nwac250 (2022).
- McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
- Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
- Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
- Yang, D. et al. Lineage tracing reveals the phylogenetics, plasticity, and paths of tumor evolution. *Cell* **185**, 1905–1923 e1925 (2022).
- Wirtz, S. et al. Chemically induced mouse models of acute and chronic intestinal inflammation. *Nat. Protoc.* **12**, 1295–1309 (2017).
- Pan, Q. et al. Genomic variants in mouse model induced by azoxymethane and dextran sodium sulfate improperly mimic human colorectal cancer. *Sci. Rep.* **7**, 25 (2017).
- Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).
- Leystra, A. A. et al. Multi-ancestral origin of intestinal tumors: Impact on growth, progression, and drug efficacy. *Cancer Rep.* **5**, e1459 (2022).
- Novelli, M. R. et al. Polyclonal origin of colonic adenomas in an XO/XY patient with FAP. *Science* **272**, 1187–1190 (1996).
- Merritt, A. J., Gould, K. A. & Dove, W. F. Polyclonal structure of intestinal adenomas in *Apc^{Min}* mice with concomitant loss of *Apc* from all tumor lineages. *Proc. Natl. Acad. Sci. USA* **94**, 13927–13931 (1997).
- Halberg, R. & Dove, W. F. Polyclonal tumors in the mammalian intestine: are interactions among multiple initiated clones necessary for tumor initiation, growth, and progression? *Cell Cycle* **6**, 44–51 (2007).
- Deng, S., Gong, H., Zhang, D., Zhang, M. & He, X. A statistical method for quantifying progenitor cells reveals incipient cell fate commitments. *Nat. Methods* **21**, 597–608 (2024).
- Olafsson, S. et al. Somatic evolution in non-neoplastic IBD-affected colon. *Cell* **182**, 672–684.e611 (2020).
- Martincoren, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e1021 (2017).
- Hu, Z. et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat. Genet.* **51**, 1113–1122 (2019).
- Potten, C. S. Stem cells in gastrointestinal epithelium: numbers, characteristics and death. *Philos. Trans. R. Soc. B* **353**, 821–830 (1998).
- van der Flier, L. G. & Clevers, H. Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annu. Rev. Physiol.* **71**, 241–260 (2009).
- Hu, Z., Li, Z., Ma, Z. & Curtis, C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nat. Genet.* **52**, 701–708 (2020).
- Huang, K. K. et al. Spatiotemporal genomic profiling of intestinal metaplasia reveals clonal dynamics of gastric cancer progression. *Cancer Cell* **41**, 2019–2037.e8 (2023).
- Gay, D. M. et al. Loss of BCL9/9l suppresses Wnt driven tumorigenesis in models that recapitulate human cancer. *Nat. Commun.* **10**, 723 (2019).
- Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
- Heide, T. et al. The co-evolution of the genome and epigenome in colorectal cancer. *Nature* **611**, 733–743 (2022).
- Vega, P. N. et al. Cancer-associated fibroblasts and squamous epithelial cells constitute a unique microenvironment in a mouse model of inflammation-induced colon cancer. *Front. Oncol.* **12**, 878920 (2022).
- Ma, R. Y., Black, A. & Qian, B. Z. Macrophage diversity in cancer revisited in the era of single-cell omics. *Trends Immunol.* **43**, 546–563 (2022).
- Katzenelenbogen, Y. et al. Coupled scRNA-seq and intracellular protein activity reveal an immunosuppressive role of TREM2 in cancer. *Cell* **182**, 872–885.e819 (2020).
- Jin, S. et al. Inference and analysis of cell–cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).
- Teller, I. C. & Beaulieu, J. F. Interactions between laminin and epithelial cells in intestinal health and disease. *Expert Rev. Mol. Med.* <https://doi.org/10.1017/S1462399401003623> (2001).
- Thliveris, A. T. et al. Polyclonality of familial murine adenomas: analyses of mouse chimeras with low tumor multiplicity suggest short-range interactions. *Proc. Natl. Acad. Sci. USA* **102**, 6960–6965 (2005).
- Tabassum, D. P. & Polyak, K. Tumorigenesis: it takes a village. *Nat. Rev. Cancer* **15**, 473–483 (2015).
- Lee, N. D., Kaveh, K. & Bozic, I. Clonal interactions in cancer: Integrating quantitative models with experimental and clinical data. *Semin. Cancer Biol.* **92**, 61–73 (2023).
- Reeves, M. Q., Kandyba, E., Harris, S., Del Rosario, R. & Balmain, A. Multicolour lineage tracing reveals clonal dynamics of squamous carcinoma evolution from initiation to metastasis. *Nat. Cell Biol.* **20**, 699–709 (2018).
- Johnson, K. E. et al. Cancer cell population growth kinetics at low densities deviate from the exponential growth model and suggest an Allee effect. *PLoS Biol.* **17**, e3000399 (2019).
- Martinez, P. et al. Dynamic clonal equilibrium and predetermined cancer risk in Barrett's oesophagus. *Nat. Commun.* **7**, 12158 (2016).
- Zhang, D., Zhang, A., He, X. & Deng, S. Variation in cancer risk between organs can not be explained by the degree of somatic clonal expansion. *Adv. Biotechnol.* **2**, 18 (2024).

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

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

© The Author(s), under exclusive licence to Springer Nature Limited 2024

Methods

Animal models and experiments

The SMALT lineage tracer cassettes were engineered into a male mouse embryonic stem cell line derived from C57BL/6J background and SMALT mouse was established. *Rosa26^{rtTA}* and *Apc^{Min/+}* mice were purchased from Saiye (Guangzhou) Biological Technology. All mice were genotyped by PCR. All mouse breeding and experiments described in this study were approved by Institutional Animal Care and Use Committee (IACUC) of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (protocols SIAT-IACUC-201207-HCS-HZ-A1491 and SIAT-IACUC-221111-HCS-LZL-A2218).

All mice were housed in specific pathogen-free (SPF) environments, with ad libitum access to food and water. The AOM/DSS inflammation-driven tumour was induced in transgenic male mice (*Rosa26^{rtTA};H11^{SMALT}*) by feeding AOM and DSS in combination. In particular, all mice were hybrid from *Rosa26^{rtTA}* and *H11^{SMALT}*, with age at 6–8 weeks and body weight ranged 18–20 g. Mouse food containing 5 mg per day doxycycline were fed to all mice three days before experimental starting points. Two mice were randomly selected as control without any further treatment, and 16 mice received a single intraperitoneal injection of AOM (Sigma-Aldrich) with a dose of 12.5 mg kg⁻¹ body weight. Three days later, the treated mice were given three cycles of 2% DSS (MP Biomedicals) treatment. Each cycle lasts for three weeks, with DSS-dissolved water for 7 days and tap water for 14 days. During each cycle, bottles were emptied on day 3 and day 5, and filled with fresh DSS-dissolved water. The *Apc^{Min/+}* male mice mated with *Rosa26^{rtTA};H11^{SMALT}* female mice and generated *Rosa26^{rtTA};H11^{SMALT};Apc^{Min/+}* males for *Apc*-driven polyp model establishment. Female mice (*Rosa26^{rtTA};H11^{SMALT}*) have been fed with 5 mg/day doxycycline for three days before mating. During the model establishment, body weight and bloody diarrhoea were tracked and recorded weekly.

Cell sorting

Cells were washed and resuspended in PBS. In the final wash, cells were resuspended in MojoSort buffer and adjusted the concentration no more than 1 × 10⁸ cells per ml. Ten microlitres of MojoSort mouse CD45 nanobeads (Biolegend, 480118, Biolegend) were mixed well with cells and incubated on ice for 15 min. After wash cells with MojoSort Buffer, the tube containing cells were placed in the magnet for 5 min for separations. CD45 positive cells were selected for a total of 3 separations, and unlabelled cells (CD45 negative cells) were purified twice. CD45 negative cells were subsequently purified by MojoSort Mouse CD326 selection kit (Biolegend, 480142). CD45 negative cells were labelled with 10 µl of TruStain FcX (anti-mouse CD16/32 antibody). After incubating at room temperature for 10 min, 5 µl of the Biotin anti-mouse CD326 (EpCAM) antibody was added, followed by another incubation on ice for 15 min. Mixtures were washed once and then incubated with 10 µl of Streptavidin Nanobeads on ice for 15 min. Tubes with washed cells were placed in the magnet for a total of 3 separations. High purity of sorted cells resuspended in PBS with 0.2% BSA-containing antibodies has been confirmed using FACS, where CD45 positive cells were labelled with Ms CD45 FITC 30-F11 (BD Pharmingen, 553079) and CD326 positive cells were labelled with ANTI-MO CD326 G8.8 APC (Thermo, 17-5791-82).

Mouse intestinal organoid experiments

A 12-week-old male *Apc^{Min/+}* (*Rosa26^{rtTA};H11^{SMALT};Apc^{Min/+}*) mouse was prepared for organoid derivation. Small intestine was dissected from the mouse and rinsed thoroughly with epithelial organoid basal medium (bioGenous B213151). The normal epithelium and neoplastic polyp were removed from the luminal surface. The collected tissues (normal or polyp) were digested in pre-cooled Epithelial Organoid Basal Medium containing 5 mM EDTA (bioGenous E219121) and incubated at 4 °C for 30 min. After removing EDTA by washing with epithelial organoid basal medium, the tissue fragments were vigorously agitated in pre-cooled

epithelial organoid basal medium to dislodge crypts, and the filtered through a 70-µm mesh filter. The dissociated tissue suspension was centrifuged at 300g for 5 min. Cell pellet was resuspended in a reduced growth factor matrix (Corning 356231). After the matrix solidified, mouse intestinal organoid complete medium (bioGenous K2001-M1) was then added. Organoids were cultured in a 37 °C incubator with 5% CO₂.

Library preparation for PacBio sequencing

Whole genome of sorted cells was isolated with DNeasy Blood and Tissue kit (Qiagen, 69504). We used 500 ng of genomic DNA as the template of amplification using KOD One PCR Master Mix (TOYOBO, KMM-101). We adopted a three-step strategy to ensure effectively amplify 3-kb target barcodes with UMIs and sample barcodes. First, we performed one cycle of lineage amplification using P1 primer which incorporates 14-nucleotide UMIs onto each original DNA molecule, followed by clean-up of PCR products with 0.9× AMPure XP beads (Beckman Coulter, A63881). Second, we conducted ten cycles of nested PCR amplification using P2 and P3 primers to enrich the indexed target molecules. Finally, the PCR products after clean-up were amplified using P4 and P5 primers which both contain 6-nucleotide symmetric sample barcodes to enable sample multiplexing.

Sequencing libraries for each sample were pooled with the same amount of DNA molecules to yield approximately equal coverage per cell per sample. Three micrograms of pooled 3-kb barcode sequencing libraries were used for PacBio SMARTbell library preparation (FraserGen, Sequel IIe platform). Roughly 200 Gb raw data and over 1 million HIFI reads were generated per sequencing run.

Long-read sequencing data processing

The 3 kb barcode library was processed using a modified pipeline as described⁴. In brief, adapters were first removed from the polymerase reads. Circular consensus sequencing (CCS) reads were generated with at least five passes using pbccs v6.2.0 (--min-length=1000 --num-threads=10 --by-strand --min-passes=3). The highly accurate CCS reads were then mapped to the reference 3-kb target barcode sequence using minimap v2.17⁴⁵ (-t 10 -A 4 -B 12 -O 10,15 -E 2,1 --score-N 0 --end-bonus 10 -a --MD -x map-pb), followed by manipulating sam files with samtools⁴⁶. Next, UMI, the 3-kb barcode sequence, and sample barcodes of each CCS read were annotated according to the mapping coordinates and the CCS mapped bam file was split into multiple fastq files accordingly. UMIs were grouped by usearch v11.0.667⁴⁷ (-id 0.95 -gapopen 3.01/2.0E -gapext 1.01/0.5E -match +2.0 -mismatch -20.0 -sizeout), removing CCS groups with less than 3 CCSs. For each CCS group, realigned reads were collapsed into a consensus CCS read, calling nucleotide substitutions with an overall mapping quality greater than 50 and a mutation frequency greater than 0.6. Finally, we obtained high-quality 3 kb barcodes for each sample.

High-quality CCS reads with at least two mutations on barcode were used for downstream analysis, where 7–49% of the reads were filtered in this step across normal and neoplastic tissues. Because non-neoplastic cells (for example, sorted immune cells) from neoplastic samples (AOM/DSS or *Apc^{Min/+}* lesions) or across different organs had a lower mutation burden than neoplastic cells (Extended Data Fig. 2), we used following two methods to filter out plausible normal cells from a neoplastic sample: (1) when a neoplastic sample exhibited a bimodal distribution of mutation counts, the cells in lower mutation cluster were removed; (2) without this characteristic bimodal distribution, the cells with mutation counts lower than the 75th percentile of mutation counts in adjacent normal cells were removed. After these filtering steps, in total we obtained 260,922 cells for downstream analysis in this study.

Phylogenetic reconstruction

Phylogenetic trees were reconstructed using a maximum-likelihood method implemented in IQ-TREE v2.2.7⁴⁸ (-T 5 -o ref-mGTR2 + FO + R10),

Article

with the original reference 3 kb barcode sequence as the phylogenetic root. The optimal substitution models GTR2 + FO + R10 was selected using IQ-TREE2 after evaluating all available substitution models. To evaluate the robustness of tree reconstruction, we ran 1,000 rounds of ultrafast bootstrap approximation (-B 1000) and 1,000 rounds of SH-like approximate likelihood ratio test (-alrt 1000). Multiple CPU cores (-T 50) were used to speed up the reconstruction process.

Using wild-type normal cells as a control, we evaluated the mutation frequency at each site across samples. Mutable sites in the 3-kb target barcode with a frequency greater than 0.04 and present at least two normal samples were defined as hotspot sites. In total, 14 sites were identified as hotspots with a mutation frequency roughly 20 times higher than that of all the other mutated sites.

WGS of mouse samples

High-quality genomic DNA from 34 mouse samples (including adjacent normal, polyps and tumours) were extracted using the DNeasy Blood and Tissue Kit (Qiagen, 69504) and DNA concentration was quantified by Qubit Fluorimeter (Invitrogen). Illumina NGS libraries were prepared according to the sequencer's protocols, and sequencing was performed on Illumina Novaseq PE150 platform. An average of 90 Gb of data (~30x coverage) for each mouse sample was generated.

WES of human sporadic polyp/CRC cohort

The human sporadic polyp/CRC cohort for WES consisted of 107 patients who had not received prior treatment or had family history of CRC, who were recruited from the Sixth Affiliated Hospital of Sun Yat-sen University. Synchronous tumour (T), polyp (P) and adjacent normal (N) samples were collected from each patient following written informed consent through sample collection protocols. All bi-specimen collection protocols were conducted in accordance with the principles of the Declaration of Helsinki and was approved by the institutional review board (IRB) of the Sixth Affiliated Hospital of Sun Yat-sen University (Number 2019ZSLYEC-06). All of samples were anonymously coded in accordance with local ethical guidelines. Freshly collected samples were stored in liquid nitrogen or directly stored as whole tissue blocks in liquid nitrogen.

A two-tier system was used for histological grading of the polyps into either low-grade or high-grade dysplasia, based on a combination of architectural and cytological features observed in H&E-stained images^{49–52}. High-grade polyps were defined as having more than 25% of the gland structure exhibiting one or more of the following features: (1) marked complex glandular crowding with glandular irregularity or cribriform architecture visible at low power; (2) substantial loss of cell polarity and pseudostratified nuclei; (3) markedly enlarged nuclei with prominent nucleoli, dispersed chromatin pattern, or atypical mitotic figures. Additionally, the size of the polyps was measured by the endoscopists. For subsequent statistical analysis, polyps were divided into two categories based on size: small (<1 cm) and large (≥1 cm). Clinical information, including size and grade, was recorded in Supplementary Table 4.

Genomic DNA from fresh frozen samples of 107 patients were extracted using the Magnetic Universal Genomic DNA Kit (Tiangen) and then quantified by Qubit 3.0 (Invitrogen). Sequencing libraries were generated using the NEB Next Ultra DNA Library Prep Kit (NEB) following manufacturer's recommendations and index codes were added to each sample. Exons were captured using SureSelect XT Human All Exon V6 kit (Agilent), and paired-end sequencing (150 bp) was performed on Illumina NovaSeq 6000 platform.

Detection of SSNVs and SCNAs

After the raw fastq files from WGS and WES were preprocessed with fastp v0.19.7⁵³, these cleaned sequencing reads were aligned to the reference genome (mouse WGS: NCBI build mm10; human WES: GRCh38) using the BWA-MEM algorithm⁵⁴ as implemented in BWA v0.7.17-r1188.

Following the GATK best practices and the associated set of tools v4.2.0.0^{55–57}, the aligned reads were processed by MarkDuplicates, BaseRecalibrator and ApplyBQSR in sequence. Sequencing depth were obtained using bamdst v1.0.9 (<https://github.com/shiquan/bamdst>). As described in previous studies^{25,28}, we used Mutect2⁵⁸ to identify all SSNVs and indels for each tumour/normal pair or polyp/normal pair in each individual. Artefactual variants or background germline variants were removed using a panel-of-normals (PoN) filter, which was generated by pooled germline variants from normal samples. High-confidence somatic variants were further extracted with Filter-MutectCall. These somatic mutations were subsequently annotated using ANNOVAR v.20200608⁵⁹ for further analysis.

We then used TitanCNA v1.28.0⁶⁰ to detect SCNAs and estimate purity and ploidy for each polyp and tumour sample. The samples with high purity (≥0.25) were kept for follow-up analyses. The CCF for each somatic mutation was calculated by adjusting its observed VAF by considering tumour purity, local copy number and multiplicity⁶¹ and the inferred time order between SCNAs and SSNVs, as previously described²⁸. In addition, mutations were classified as 'clonal' if the upper bound of the 95% confidence interval of CCF was equal to or greater than 1. Otherwise, they were classified as 'subclonal'. Then, somatic mutations were retained if they had ≥5 variant reads in a polyp or tumour for WES and ≥4 variant reads in polyp or tumour for WGS and total reads at variant position ≥10 in both normal and polyp or tumour. For human WES data, variants with CCF ≥ 0.04 in polyp or tumour and VAF < 0.01 in normal sample were retained, and private mutations in driver genes were further manually checked in the Integrative Genomics Viewer v2.15.1⁶² in both polyp and tumour within the individual to ensure that they were indeed private. For mouse WGS data, we detected and visualized the SCNAs using scripts from a mouse cancer genome sequencing data analysis pipeline⁶³.

Single-gland WGS

Five polyps from the transverse colon of a sporadic patient (B139, male, 73 years old), recruited from the Sixth Affiliated Hospital of Sun Yat-sen University, were dissected through colonoscopy. The method of single-gland isolation and WGS is based on adaptation of a previously published protocol³². In order to sample distinct spatial regions from this polyp lesion, we first cut out 6 pieces representing different regions from the fresh frozen tissue. Then, the tissue pieces were placed on a clean glass slide under the microscope with 20 µl PBS. Subsequently, individual crypts (or glands) were manually dissociated using a 23 G (0.6 × 25 mm) injection needle and aspirated using a pipette. Each gland was transferred to a PCR tube with 20 µl protease buffer⁶⁴. The glands with enough DNA (>1 ng) extracted were selected for WGS library construction using the Vazyme TruePrep DNA Library Prep Kit V2 (TD501/502/503, Vazyme, China) according to the manufacturer's instructions. Libraries were sequenced on an Illumina NovaSeq platform (Illumina). Each gland has at least 100 Gb of raw data to ensure sufficient sequencing depth (>20x) to support subsequent analysis (Supplementary Table 8).

The single-gland WGS data were also preprocessed using the same pipeline with our WES data analysis. To filter out false positive mutation callings from the single-gland WGS data, two callers Mutect2⁵⁸ and Strelka v2.9.2⁶⁵ were used and consensus variant call set was obtained using VariantFilter (<https://github.com/rschenck/VariantFilter>). Finally, high-confidence mutations with VAF ≥ 0.15 and variant reads ≥ 4 and total reads ≥ 7 in gland samples and no reads in the matched bulk normal sample. Phylogenetic trees of glands using the SSNVs were reconstructed using a neighbour-joining method⁶⁶ implemented in Biopython⁶⁷, under Python v3.8.13, with the reference sequence without mutations as the root.

Somatic copy number alterations (SCNAs) of each crypt relative to matched normal tissue samples were estimated by Sequenza v3.0.0 package for R⁶⁸. Joint segmentation on B-allele frequencies and depth

ratio values across all samples from a given crypt were used to determine a set of breakpoints to use for the subsequent analysis. The cutoff value for amplification (AMP) and deletion (DEL) were set as $\log_2(2.5/2)$ and $\log_2(1.5/2)$. Arm-level copy number was annotated using all segments with same status dropping in the more than >40% specific region. Finally, we identified several important arm-level SCNAs were labelled in the phylogenetic tree.

Clonal relatedness for polyp/CRC pairs

The patients with none or very few (<10) shared clonal SSNVs/indels implied an distinct origin for polyp and tumour from the same patient. Clonal relatedness for each polyp/tumour pair was further assessed using Breakclone v0.3.3⁶⁹, which incorporated both population frequency and allele frequency of mutations. Breakclone calculated a *P* value based on a permutation test and a clonal relatedness score (Supplementary Table 6). The clonal relationship between polyp and tumour in each patient was then classified as related ($P < 0.01$ and clonality score > 0.1), unrelated ($P > 0.05$ and clonality score < 0.05), or ambiguous (the remaining pairs). This analysis was also applied to multi-region sequencing data in two sporadic CRCs⁷⁰ as a positive control.

dN/dS analysis

The dN/dS ratio was estimated using the R package dndscv v0.0.1.0²⁴, with the default and recommended parameters for all mutations in polyps and CRCs. The dndscv model was applied separately to mutations sets within the three major subtypes (polyclonal polyps, monoclonal polyps, and CRCs). The putative driver CRC genes²⁸ were curated by merging all TCGA pan-cancer drivers⁷¹ ($n = 299$) and additional CRC-specific drivers ($n = 47$) annotated by the COSMIC Cancer Gene Census (<https://cancer.sanger.ac.uk/census>). Except for the global dN/dS estimates of the fitted models, dN/dS estimates of putative CRC driver genes were extracted using the genesetdnds function of dndscv. For mouse WGS data, dN/dS values were estimated for each gene. The inferred global rates were then used to estimate global dN/dS for the mouse one-to-one ortholog of a curated human CRC driver genes²⁸. The human-mouse one-to-one orthologs were downloaded from the Ensembl database⁷² (release 110) via BioMart (www.ensembl.org/biomart, last accessed 23 August 2023).

Estimating the number of progenitor cells

We used TarCA v0.1.0²² to estimate the number of founding progenitor cells for each neoplastic lesion with the cell phylogenetic trees reconstructed by SMALT. It should be noted that a progenitor is defined as a cell capable of founding a clonally expanding population in the neoplasm, regardless of the underlying mechanisms of clonal expansions. Thus, for a phylogenetic tree consisting of both normal and neoplastic cells, TarCA first calculated the probability (Pr) of two random neoplastic cells sharing a common progenitor in a monophyletic clade of the tree. As demonstrated by Liu et al.⁴, the effective number of progenitors (Np) for a neoplastic tissue equals to the inverse of this probability, namely:

$$Np = 1/Pr \quad (1)$$

Given a phylogenetic tree including both normal and neoplastic cells, Pr can be calculated as:

$$Pr = \frac{\sum C_{m_i}^2}{C_{N_s}^2} = \frac{\sum (m_i \times (m_i - 1))}{N_s \times (N_s - 1)} \quad (2)$$

Where m_i is the number of sampled cells of the i th monophyletic clade of neoplastic cells, and N_s is the total number of neoplastic cells in the sample.

We performed downsamplings to evaluate the robustness of Np estimates: (1) if a neoplastic sample had more than 1,000 cells, 1,000

neoplastic cells were downsampled 20 times and all normal cells were included each time; (2) if there were fewer than 1,000 cells in the neoplasm sample, m normal or neoplastic cells were downsampled 20 times. Here downsampling is performed for the cell type (normal or neoplastic) with higher cell number, whereas m is the smaller value between the number of normal cells and the number of neoplastic cells.

Single-cell RNA sequencing

Colons from nine AOM/DSS mice were cut longitudinally and washed with cold phosphate buffered saline (PBS). Tumours were collected and cut into small pieces. Single-cell suspension was prepared using MACS Tissue Dissociation Kits (Miltenyi Biotec, 130-096-730). In brief, the tissues were cut into small pieces and dissociated with digestion buffer and incubated at 37 °C for 30 min. The digestion solution was washed with DME. Cell suspensions were then filtered through a 70- μ m cell strainer, spun down at 800g for 5 min, and resuspended. Library generation for 10X Genomics v2 chemistry was performed using the Chromium Single Cell 5' Reagent Kits.

Raw scRNA-seq data were aligned to the reference genome (mm10) and the UMI was quantified for each sample using Cell Ranger v7.1 (10x Genomics). Two normal samples (GSE134255) obtained from Vega et al.³³ were integrated to our dataset as normal controls. The UMI matrices generated by Cell Ranger were processed into individual Seurat objects using the R package Seurat v5.0⁷³. To ensure high data quality, cells were retained if they had at least of 500 genes and less than 15% mitochondrial gene expression, and were identified as singlets by DoubletFinder v2.0.3⁷⁴. Genes expressed in less than 3 cells were filtered. In this way, 45,620 high-quality cells were retained and these cells were then normalized using Seurat's SCTransform v2 (SCT)^{75,76}, which is specifically tailored for droplet-based scRNA-seq data and is much more robust to technical effects than default normalization method. Finally, mitochondrial read counts were regressed from the data, and 3,000 highly variable genes were screened for subsequent dimensionality reduction and integration analysis.

Principal component analysis (PCA) was performed first and the top 50 significant principal components were selected using Seurat's ElbowPlot function. To minimize batch effects and preserve the genuine biological heterogeneity, four single-cell integration methods were compared: canonical correlation analysis, reciprocal PCA, Harmony⁷⁷ and FastMNN⁷⁷. Harmony demonstrated moderate integration ability in our data and the best biological interpretability in the subsequent cell clustering and annotation. After removing batch effects using Harmony v1.1.0 based on the top 50 PCA components, the FindNeighbors function was applied on 30 dimensions, computing k -nearest neighbour (KNN) and shared nearest-neighbour (SNN) graph based on the 'harmony space'. Afterwards, the FindClusters function was used with a resolution of 0.1 to perform the first-round clustering using the Louvain algorithm. Each cluster was carefully annotated using known markers leading to the identification of eight major cell types: neutrophils, macrophages, T cells, B cells, endothelial cells, epithelial cells, mast cells and fibroblasts (Supplementary Tables 9 and 10). Next, a second-round clustering was performed by setting the resolution to 0.4, resulting in the identification of 26 subclusters. Notably, epithelial cells and macrophages exhibited significant heterogeneity, dividing into 10 and 7 subclusters, respectively. Meanwhile, B cells were categorized into two subclusters (B and B plasma), whereas T cells were subdivided into three subclusters (T regulatory, T cytotoxic and T progenitor). The remaining cell types were not subjected to further subdivision. To further annotate the subclusters of epithelial cells and macrophages, we used the FindMarkers function to identify marker genes within each subcluster of the two main clusters. We selected a representative marker gene as a suffix for naming, distinguishing the subclusters from one another. Specifically, for epithelial cells, the subclusters were named as follows: Epi-Mt3, Epi-Mgst3, Epi-Peak1, Epi-Mdga2, Epi-Ptprd, Epi-Krt20, Epi-Srgn, Epi-Goblet, Epi-Arglu1, and

Article

Epi-Stem-like. Similarly, for macrophages, the subclusters were named: Macro-Trem2, Macro-Chil3, Macro-Ifitm1, Macro-Acod1, Macro-Cxcl3, Macro-St18, and Macro-Ccr7 (Supplementary Tables I1 and I2).

Differentially expressed gene analysis

Differentially expressed genes (DEGs) between subclusters were determined using the Seurat FindALLMarkers and FindMarkers function with the Wilcoxon rank-sum test. DEGs were retained if they expressed in more than 1% of cells within a subcluster, coupled with an average log-transformed fold change greater than 0.25. To annotate the cell type of each cluster, canonical expression markers discovered among the DEGs and documented in the literature were combined. Additionally, heat maps and dot plots were generated using Seurat's DoHeatmap and DotPlot functions. The expression of marker genes on the uniform manifold approximation and projection (UMAP) plot was visualized using Seurat's FeaturePlot functions.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was conducted using the clusterProfiler package⁷⁸ (version 4.8.3). The GSEA algorithm calculates an enrichment score (ES) reflecting the degree of overrepresentation at the top or bottom of the ranked list of the DEGs. Positive or negative enrichment scores indicate that the gene set is enriched at the top or bottom of the DEG ranking list, corresponding to activated (ES > 0) or suppressed (ES < 0) gene pathways. For our pathway enrichment analysis, we utilized the Gene Ontology gene sets (M5) and mouse orthologue hallmark gene sets available in the MSigDB database (<https://www.gsea-msigdb.org/gsea/msigdb>).

Gene set variation analysis and scoring

The Gene Set Variation Analysis⁷⁹ (GSVA) package v1.48.3 was used to score different gene sets. GSVA computes a score for a specific gene set in individual cells based on their gene expression data. Each of the seven classes of macrophages, as defined by the unified nomenclature, possesses its unique set of gene markers. The macrophages in our dataset were annotated into these seven classes based on the corresponding gene sets (Supplementary Table I3). Each cell was scored using the ssgsea method with parameters `kcdf = "Gaussian"`, `ssgsea.norm=TRUE`.

Differential cell abundance analysis

miloR v1.8.1⁸⁰ was used to identify changes in abundance across distinct cell types. The milo performs differential abundance testing by assigning cells to partially overlapping neighbourhoods on a KNN graph. Molo, by considering the change in cell abundance as a continuous process, can capture continuous trajectories without depending on pre-clustering cells into discrete groups. This approach allows us to observe temporal variations in abundance for different cell types using generalized linear models. Specifically, a KNN graph was constructed based on Seurat's 'harmony space', and 10% of the cells were selected using a sampling refinement algorithm. Each sampled cell, along with its neighbouring cells exhibiting close transcriptional spatial distances (determined by KNN distance), constitutes a neighbourhood. Within each neighbourhood, the number of cells within each neighbourhood was reasonably distributed between 50 and 100. The spatial FDR (corrected *P* value) for each significant abundance change neighbourhood did not exceed 0.1.

Cell-cell communication analysis

CellChat v1.6.1³⁶ was used to infer interactions between different epithelial cell types within each sample. CellChat provides distinct advantages in the analysis of ligand-receptor interactions. It considers not only the expression levels of ligands and receptors genes across cell subclusters but also the necessary structural components during ligand-receptor binding, such as multimeric ligand-receptor complexes, soluble agonists, antagonists and stimulatory and inhibitory membrane-bound co-receptors. The strength of ligand-receptor

interactions within a single sample was measured by the total number of inferred ligand-receptor pairs present in that sample. As the number of inferred ligand-receptor pairs is affected by the sample cell number, we performed 50 downsamplings, each including 689 epithelial cells (the smallest cell number among these 9 samples, IBD142_T5) for each sample. We used the average number of ligand-receptor pairs between each two epithelial subclusters from the 50 downsamplings to visualize the intercellular interactions in each sample. The average number of ligand-receptor pairs per sample in the 50 downsamplings was used to perform correlation analysis with the lesion clonality (1/Np).

We identified a total of 1,543 candidate ligand-receptor pairs from the CellChat database (CellChatDB) that had genes expressions among neoplastic epithelial cells. These ligand-receptor pairs represent all possible potential interactions among neoplastic cells. We thus quantified the occurrence of each of these ligand-receptor pairs individually and compared early low-clonality lesions (Np > 3, 6 samples) and late high-clonality lesions (Np ≤ 3, 3 samples) using the Wilcoxon rank-sum test. An adjusted FDR threshold of less than 0.05 indicates a significant difference in ligand-receptor pair numbers between the two stages. MultiNicheNetR v1.0.3⁸¹ was also used to identify significantly altered ligand-receptor pairs under two conditions. MultiNicheNetR emphasizes whether downstream target genes are regulated by the corresponding ligand after binding to its receptor. The establishment of a ligand-receptor pair requires simultaneous high expression of the ligand in the sending cell and the receptor in the receiving cell, along with regulated target genes in the receiving cell (Supplementary Table I6). Both CellChat and MultiNicheNetR were implemented with the parameters recommended by their respective authors.

Reporting summary

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

Data availability

All processed data generated in this study have been deposited and are available at Zenodo (<https://zenodo.org/records/11647317> (ref. 82)). The single-cell phylogenies generated from this study are displayed at <https://smalt-phylogeny.org/>. Raw data are publicly available from the National Genomics Data Center (NGDC) under the accession numbers PRJCA024217 and PRJCA023981. Public scRNA-seq data for two wild-type mouse colon samples can be accessed with GSE134255 (<https://www.ncbi.nlm.nih.gov/>).

Code availability

All computer code used in this study is available from the GitHub repositories <https://github.com/zhaolianlu/SMALT-mouse> and <https://github.com/zhaolianlu/Homo-preCRC>.

- Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
- Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
- Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
- Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
- Gschwantler, M. et al. High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics. *Eur. J. Gastroenterol. Hepatol.* **14**, 183–188 (2002).
- Sawhney, M. S., Dickstein, J., LeClair, J., Lembo, C. & Yee, E. Adenomas with high-grade dysplasia and early adenocarcinoma are more likely to be sessile in the proximal colon. *Colorectal Dis.* **17**, 682–688 (2015).
- Gibson, J. A. & Odze, R. D. Pathology of premalignant colorectal neoplasia. *Dig. Endosc.* **28**, 312–323 (2016).
- Kuo, E., Wang, K. & Liu, X. A focused review on advances in risk stratification of malignant polyps. *Gastroenterology Res.* **13**, 163–183 (2020).
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
55. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
56. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
57. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 1110 11–1110 33 (2013).
58. Benjamin, D. et al. Calling somatic SNVs and indels with Mutect2. Preprint at *bioRxiv* <https://doi.org/10.1101/861054v1> (2019).
59. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
60. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
61. Tarabichi, M. et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat. Methods* **18**, 144–155 (2021).
62. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
63. Lange, S. et al. Analysis pipelines for cancer genome sequencing in mice. *Nat. Protoc.* **15**, 266–315 (2020).
64. Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
65. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
66. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
67. Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
68. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
69. Lips, E. H. et al. Genomic analysis defines clonal relationships of ductal carcinoma in situ and recurrent invasive breast cancer. *Nat. Genet.* **54**, 850–860 (2022).
70. Li, J. et al. Genomic and transcriptomic profiling of carcinogenesis in patients with familial adenomatous polyposis. *Gut* **69**, 1283–1293 (2020).
71. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e318 (2018).
72. Martin, F. J. et al. Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).
73. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
74. 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.e324 (2019).
75. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
76. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.* **23**, 27 (2022).
77. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
78. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
79. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
80. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
81. Browaeys, R. et al. MultiNicheNet: a flexible framework for differential cell-cell communication analysis from multi-sample multi-condition single-cell transcriptomics data. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.06.13.544751v1> (2023).
82. Lu, Z. et al. Systematic lineage tracing unveils polyclonal origin and evolution in colorectal precancer. *Zenodo* <https://doi.org/10.5281/zenodo.11647317> (2024).

Acknowledgements The authors thank C. Curtis, Q. Nie, H. Ji, W. Zhai and Hu laboratory members for constructive discussions, and B. Chen for the guidance of gland isolations. This work was supported by National Key R&D Program of China (2021YFA1302500 to Z. Hu and X.H.), National Natural Sciences Foundation of China (82241236 and 32270693 to Z. Hu, 32293190 and 32293191 to X.H., 82273346 to Z. He and 32100486 to Z.L.), Guangdong Basic and Applied Basic Research Foundation (2021B1515020042 to Z. Hu) and Shenzhen Science and Technology Program (RCBS20210706092346032 to Z.L.).

Author contributions Z. Hu, X.H. and Z.L. conceived and designed the study. Z.L., S.D., K.Z. and J.T. performed mouse experiments. Z.L. analysed the targeted long-read sequencing data. S.M. and H.Z. conducted single gland-related experiments. S.M. and D.X. analysed the WGS and WES data. X.Z. analysed the scRNA-seq data. Z. Hu and D.X. performed computational modelling and inferences. X.K. performed organoid cultures. J.W. provided guidance on data analysis. L.T.O.L. provided guidance on mouse experiments. Z. He, P.L. and J.Y. identified patients with synchronous polyps and CRC, performed colonoscopic polypectomy and collected clinical specimens. Y.Z. conducted pathological diagnosis and analysis. Z. Hu, Z.L., S.M., D.X. and X.Z. wrote the manuscript with contributions from all co-authors. Z. Hu, X.H. and Z. He supervised the project.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-08133-1>.

Correspondence and requests for materials should be addressed to Zhen He, Xionglei He or Zheng Hu.

Peer review information *Nature* thanks James DeGregori, Richard Halberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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