nature genetics

Article

https://doi.org/10.1038/s41588-026-02692-x

Prevalence and chronology of colibactin-associated mutational processes and their microbiome spectra in Japanese colorectal cancer

Received: 12 December 2025

Accepted: 23 June 2026

Published online: 10 August 2026

Check for updates

Satoshi Shiba[1,2,15], Shinichi Yachida[1,3,4,15] ✉, Sayaka Mizutani[5,6,15], Yasushi Totoki[1,15], Hiromi Nakamura[2,7,15], Natsuko Hama[2,7], Norikatsu Miyoshi[8], Yasuhito Arai[2,7], Mihoko Saito-Adachi[2,7], Hirokazu Kimura[1], Yoshito Hayashi[9], Hiroyuki Takamaru[10], Koji Tanaka[8], Rie Hayashi[8], Hirofumi Rokutan[11], Shoko Ikuta[1,12], Yukihide Kanemitsu[13], Yuichiro Doki[8], Hidetoshi Eguchi[8], Satoshi Hattori[14], Yutaka Saito[10], Takuji Yamada[5] ✉ & Tatsuhiro Shibata[2,7] ✉

The incidence of colorectal cancer (CRC) has risen in recent decades, with a disproportionate increase observed among younger individuals in Japan and other countries. The etiological contribution of the gut microbiota to CRC pathogenesis is recognized, yet the mechanisms involved remain to be fully clarified. Here we integrated whole-genome sequencing (WGS) and transcriptome profiling of CRC with whole-genome metagenomic sequencing of fecal samples to interrogate host–microbiome interactions at high resolution. Application of interpretable artificial intelligence enabled the stratification of CRC into four distinct microbiome-informed subtypes. WGS analysis identified mutational signatures SBS88 and ID18, linked to colibactin exposure, as early clonal events detected in 44.8% of non-hypermutated patients. Notably, these signatures were significantly more frequent among patients born after the 1960s. Microbiome-based subclassification revealed subtype-specific clinical and molecular features. Collectively, our findings indicate that colibactin exposure constitutes a prevalent and potentially modifiable risk factor for CRC in the Japanese population.

The worldwide incidence of CRC has shown a concerning upward trend, particularly in Japan and several other regions[1,2]. Although this increase is commonly attributed to the adoption of a Westernized lifestyle, the prevalence in Japan exceeds that in Western countries, suggesting the contribution of additional, region-specific factors. Notably, the incidence of early-onset CRC, defined as diagnosis before the age of 50 years, has been steadily rising across the globe since the 1990s[3,4].

A growing body of evidence implicates the gut microbiota as a critical determinant of CRC pathogenesis[5]. Certain bacterial species, including *Fusobacterium nucleatum*, *Peptostreptococcus stomatis* and *Parvimonas micra*, are frequently detected in the stools and tumor tissues of patients with CRC and may serve as potential diagnostic markers[6–8]. Among the microbial metabolites, colibactin has attracted particular attention. This genotoxin, synthesized by *Escherichia coli* and other enteric bacteria harboring the *pks* genomic island[9–11], induces DNA interstrand crosslinks and double-strand breaks[9,10]. Colibactin-associated DNA damage imprints distinct tumor mutational signatures, notably single-base substitution (SBS)88 and short

A full list of affiliations appears at the end of the paper. ✉e-mail: syachida@cgi.med.osaka-u.ac.jp; takuji@bio.titech.ac.jp; tshibata@ims.u-tokyo.ac.jp

https://doi.org/10.1038/s41588-026-02692-x

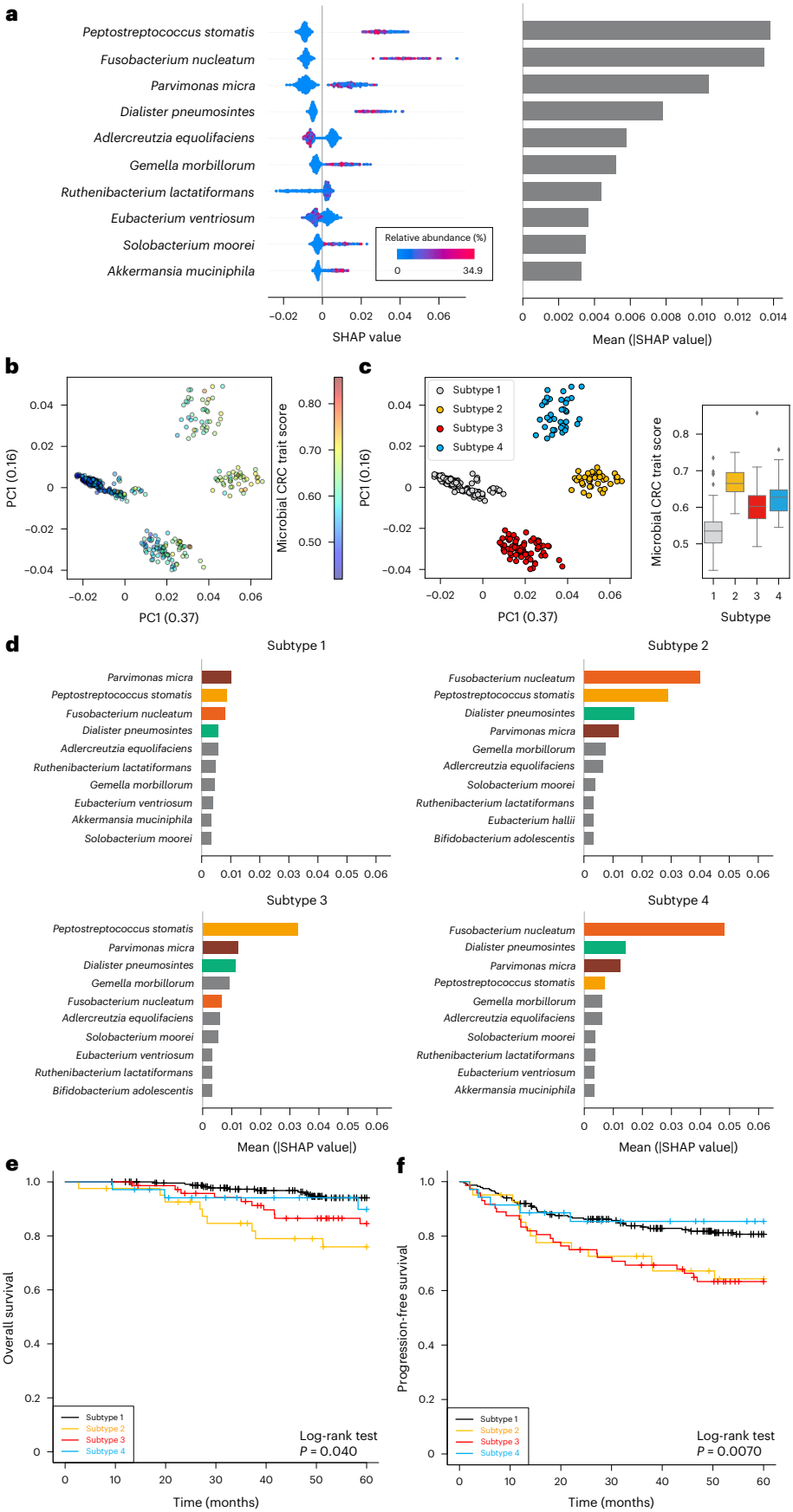


Fig. 1 | Subclassification of CRC based on the gut microbiome. a, The importance of global microbiome features in the CRC classifier. Each data point corresponds to an individual (sample) within the cohort (left). The plot shows the case samples from cohort 1 and cohort 2 ($n = 395$) and the HC samples from our previous report[8] ($n = 146$). The SHAP values, represented on the x axis, provide a quantitative measure of the influence of a bacterial species on the classifier's prediction for a given sample. The colors indicate the relative abundance of the species, with blue and red indicating low and high abundance, respectively. Among the species that have the greatest influence on the CRC classifier, including *P. stomatis*, *F. nucleatum*, *P. micra* and *D. pneumosintes*, the majority of samples with high relative abundance (indicated in red) are positioned on the positive side of the x axis. This indicates that a higher prevalence of these bacteria is associated with an increased likelihood of CRC. The top ten most important microbiome features were determined by calculating the mean of the absolute SHAP values (right). **b**, The integration of PCA with SHAP values, with microbial CRC trait score superimposed. The plotted figures represent case samples from cohort 1, cohort 2 and HCs from our previous study[8]. **c**, PCA was performed on SHAP values labeled according to four different subtypes. The case samples from cohorts 1 and 2 were effectively clustered into four subtypes using k-means clustering (left). The estimated microbial CRC trait scores by each subtype were calculated using the microbiome as a feature (right). Subtype 2 was identified as the most accurate in the diagnosis of CRC. The boxes represent 25th–75th percentiles, horizontal lines indicate the medians and whiskers extend to the maximum and minimum values within 1.5× the interquartile range. **d**, The top ten most important microbiome traits in each of the four subtypes. In particular, the right panel in **a** highlights four globally significant species. The x axis indicates the mean of the absolute SHAP values. **e**, The 5 year overall survival by microbial subtype. The log-rank test indicated significant differences in survival among subtypes; however, Cox proportional hazards modeling revealed no subtype-specific effect after adjustment for stage, identifying stage as a confounding factor. **f**, Progression-free survival by microbial subtype. As with overall survival, the log-rank test demonstrated significant differences in progression-free survival among subtypes. However, after adjustment for disease stage in the Cox proportional hazards model, no subtype-specific effects were observed, indicating that stage acted as a confounding factor.

insertion and deletion (ID)18, which reflect cumulative mutational processes operative during colorectal tumorigenesis[12–14].

A recent large-scale WGS study encompassing approximately 1,000 CRC cases from multiple continents revealed substantial geographic variation in mutational processes[15]. Although only 28 Japanese cases were included, Japan exhibited the highest prevalence of SBS88 and ID18 among the 11 participating countries, suggesting potential host–microbiome–environmental interactions unique to this population.

In this study, we integrated a newly established cohort (cohort 1) with our previously reported dataset (cohort 2)[8], affording a total of 600 participants (454 patients with CRC and 146 healthy controls (HCs)) (Extended Data Fig. 1a). WGS and whole-transcriptome sequencing were performed on Japanese CRC cases from cohort 1 ($n = 200$) to delineate their genomic landscape comprehensively. In parallel, whole-genome metagenomic sequencing (WGMS) was conducted on fecal samples from 395 cases and 146 HCs to characterize gut microbial profiles. An interpretable artificial intelligence framework stratified CRC cases into distinct subgroups according to microbial composition. This integrative analysis aimed to elucidate the relationship between fecal metagenomic subtypes and the genomic architecture of CRC, with particular emphasis on mutational processes underlying tumorigenesis.

Results

Participants and methods of analysis

In cohort 1, a total of 255 treatment-naive patients with histologically confirmed CRC were enrolled in this study (Supplementary Table 1 and Extended Data Fig. 1a). Cases with hereditary CRC syndromes were excluded. Fresh–frozen resected tumor specimens, matched lymphocyte-derived DNA and stool samples were collected from each participant. WGS was performed on tumor and lymphocyte DNA ($n = 200$), and WGMS was conducted on stool samples ($n = 196$).

HCs were defined as individuals who underwent colonoscopy for indications such as a positive fecal occult blood test but showed no neoplastic or other clinically significant findings. Control data (146 individuals) were derived from cohort 2 (Supplementary Table 2)[8,16]. The Brinkman Index was significantly higher in patients with CRC than in HCs ($P = 3.2 \times 10^{-4}$), and the CRC cohort was also significantly older ($P = 0.0010$) (Supplementary Table 3).

Construction of a gut microbiome-based CRC classifier

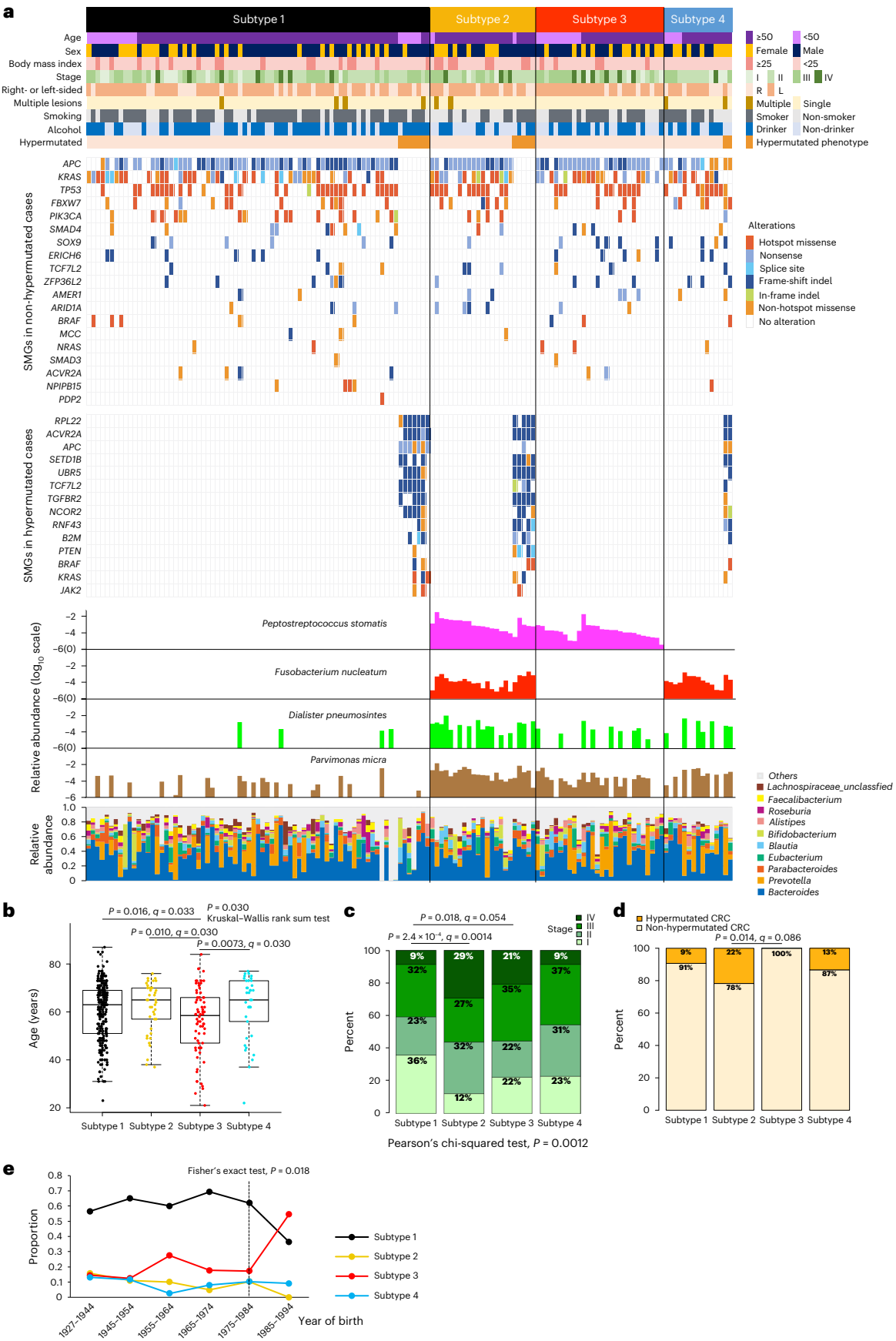
A random forest classifier was trained on four independent cohorts[7,17–19] using the curatedMetagenomicData R package[20], applying abundance and prevalence thresholds set at 1.0×10^{-5} and 0.95, respectively. Tenfold cross-validation yielded an area under the receiver operating characteristic curve (AUC) of 0.84. The trained model was subsequently applied to the cohort 2 WGMS dataset[8], which comprises species-level relative abundance profiles generated by MetaPhlAn3, to estimate the probability of CRC occurrence, reflecting the predicted CRC-associated microbial traits. The model achieved an AUC of 0.68 (Extended Data Fig. 1b). Next, application of the classifier to the current cohort of 196 CRC cases (cohort 1), together with the same HC group[8], yielded an AUC of 0.74, further supporting the reproducibility and robustness of the model. Finally, when applied to the combined CRC cohort ($n = 395$; HCs, $n = 146$) encompassing cohort 1 and cohort 2, the classifier achieved an AUC of 0.71.

Identification of CRC subtypes based on gut microbiota composition

The classifier described above was subjected to interpretable artificial intelligence analysis using the Shapley additive explanations (SHAP) framework, as previously reported[16]. *P. stomatis*, *F. nucleatum*, *P. micra* and *Dialister pneumosintes* were identified as the major bacterial species distinguishing patients with CRC from HCs (Fig. 1a and Supplementary Table 4). In Fig. 1b, individual data points on the

Fig. 2 | Clinical and genomic background of CRC subtypes based on gut microbiota. a, Clinical and molecular characteristics of CRC stratified by microbial subtype. The top row shows the clinical background of patients diagnosed with CRC. The next row shows the mutations in significantly mutated genes (SMGs). The following rows show the relative abundance of the four bacterial species contributing most strongly to the CRC-associated microbial traits, as determined by SHAP values. The bottom row presents the distribution of relative abundances for the ten most prevalent bacterial genera. **b**, Microbiota-based subtypes of CRC linked to the age of the patient. Patients classified as subtype 3 were significantly younger than those in all other subtypes in pairwise comparisons adjusted for multiple testing. **c**, Distribution of pathological stage by microbial subtype. Patients classified as subtype 2 or subtype 3 were diagnosed at significantly more advanced stages compared with those in subtype 1. **d**, Proportion of hypermutated CRC by microbial subtype. Hypermutated cases were absent in subtype 3, showing a significant difference compared with subtype 2. P values were calculated using two-sided tests, and q values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) method. The boxes represent the 25th–75th percentiles, the horizontal lines indicate the medians and the whiskers extend to the maximum and minimum values within 1.5× the interquartile range (**b**–**d**). **e**, Chronological distribution of CRC subtypes by birth cohort. The relative frequency of subtype 3 exhibited a marked increase among patients born after 1980 (the year 1980 is indicated by a black dashed line). P values were calculated using two-sided tests.

https://doi.org/10.1038/s41588-026-02692-x



https://doi.org/10.1038/s41588-026-02692-x

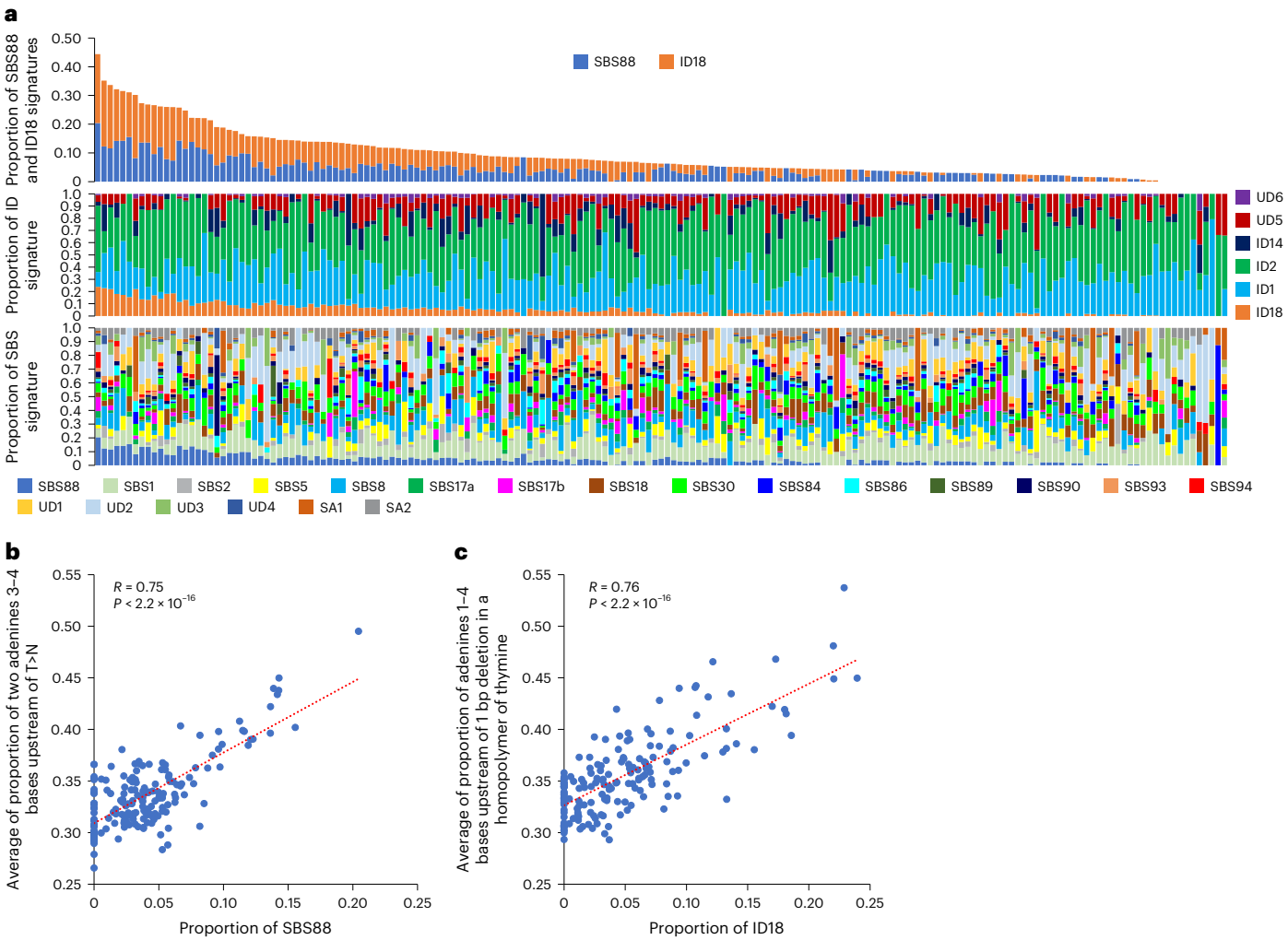


Fig. 3 | Validation of colibactin-associated mutational signatures and sequence motifs in colorectal cancer. a, Proportions of colibactin-associated mutational signatures. The upper panel shows the proportions of SBS88 (colibactin-associated SBS signature) and ID18 (colibactin-associated indel signature). Among 183 non-hypermutated CRC cases, 82 (44.8%) exhibited an SBS88 or ID18 contribution of ≥5%. The middle and lower panels depict the relative contributions of individual ID and SBS signatures, respectively. **b**, Correlation between SBS88 mutational signature activity and the colibactin-associated SBS motif. The colibactin-associated SBS motif is characterized by two adenines located at positions −3 and −4 upstream of the T > N substitution site. For each case, the proportions of adenines at positions −3 and −4 were calculated separately, and the average of these two values was used as the motif enrichment score (*y* axis)[12,13]. Statistical significance was assessed using a two-sided Pearson correlation test. **c**, Correlation between ID18 mutational signature activity and the colibactin-associated ID motif. The colibactin-associated ID motif is defined by the presence of an adenine located one to four bases upstream of a single-thymine deletion within a 1–5-bp thymine homopolymer region. The proportions of adenines at positions −4 to −1 relative to the deletion site were calculated, and the average of these proportions was used as the motif enrichment score (*y* axis)[12,23]. Pearson's correlation coefficient was calculated, and statistical significance was assessed using a two-sided Pearson correlation test. Both the SBS88 signature and the corresponding colibactin-specific SBS motif (**b**), as well as the ID18 signature and the corresponding colibactin-specific ID motif (**c**), showed significant positive correlations.

scatterplot are colored according to the predicted CRC-associated microbial traits inferred from classifier outputs. *K*-means clustering of the resulting SHAP values identified four distinct CRC subtypes, with the optimal solution determined by minimizing the within-cluster sum of squares (Fig. 1c and Extended Data Fig. 2a). These were designated as subtypes 1–4, and the microbial CRC trait scores were compared across subtypes. Subtype 2 exhibited the highest CRC-associated microbial trait scores, whereas subtype 1 displayed the lowest. *F. nucleatum* and *P. stomatis* were the most influential contributors to subtype 2, with mean SHAP values of 0.039 and 0.029, respectively (Fig. 1d and Supplementary Table 4). By contrast, subtype 1 was characterized by the absence of dominant bacterial drivers, as no species exhibited a mean SHAP value exceeding 0.01. *P. stomatis* and *F. nucleatum* were the defining taxa in subtypes 3 and 4, respectively. *E. coli* was detected in the metagenomic data but was not among the most abundant taxa contributing to the subtype classification (Supplementary Table 4). When the classifier was applied separately to cohort 1 (Extended Data Fig. 2b) and cohort 2 (Extended Data Fig. 2c), both datasets independently segregated into four similar microbiome-defined subtypes, confirming the robustness of the classification scheme and justifying the integration of the two cohorts for subsequent analyses.

Kaplan–Meier analysis demonstrated significant differences in overall survival among the four microbial subtypes (log-rank $P = 0.040$) (Fig. 1e). Univariate analysis of established prognostic variables (that is, tumor site and stage) revealed that only the proportion of stage IV cases differed significantly among subtypes, indicating an imbalance in stage distribution. In multivariate Cox modeling including subtype and stage as covariates, stage—but not subtype—remained significantly associated with prognosis (Supplementary Table 5). These results suggest

that the observed survival differences are largely attributable to disparities in stage composition, particularly the overrepresentation of stage IV disease in subtype 2, rather than intrinsic biological differences.

For progression-free survival (PFS), the analysis likewise revealed divergence among the four microbial subtypes, with Kaplan–Meier curves showing significant variation in outcomes (log-rank $P = 0.0070$) (Fig. 1f). Examination of clinical covariates indicated that only the frequency of stage IV disease differed significantly between subtypes, suggesting heterogeneity in baseline disease burden (Supplementary Table 5). In the Cox proportional hazards model adjusted for stage, stage remained a strong determinant of PFS, whereas subtype showed no independent effect (Supplementary Table 6).

Differentiation of clinical characteristics and genomic abnormalities by microbiome-based subtypes

We analyzed clinical features and genomic abnormalities using WGS and transcriptome sequencing (RNA sequencing (RNA-seq)). The RNA-seq analysis showed that subtype 2 had significantly more non-human bacterial reads than the other subtypes, indicating a higher level of microbiota in subtype 2 (Supplementary Fig. 1). Figure 2a presents detailed clinical and pathological characteristics, along with genomic alterations, for patients who underwent both WGS and WGMS ($n = 141$) (Supplementary Table 7). We compared the clinical and pathological characteristics of 395 patients with CRC undergoing WGMS, stratified by gut microbiota-defined subtypes (Supplementary Table 8). Patients with subtype 3 CRC were significantly younger than those in all other subtypes (vs subtype 1, $P = 0.016$, $q = 0.033$; vs subtype 2, $P = 0.010$, $q = 0.030$; vs subtype 4, $P = 0.0073$, $q = 0.030$) (Fig. 2b). We assessed intergroup differences across microbial subtypes ($P = 9.4 \times 10^{-4}$, Pearson's chi-squared test) and applied corrections for multiple comparisons. Statistically significant differences in stage distribution were observed between subtype 1 and subtype 2 ($P = 2.4 \times 10^{-4}$, $q = 0.0014$), as well as between subtype 1 and subtype 3 ($P = 0.018$, $q = 0.054$) (Fig. 2c). These disparities in stage distribution probably contribute to the observed differences in overall survival and PFS (Fig. 1e,f). In relation to the hypermutated phenotype (more than 500 non-synonymous single-nucleotide variants (SNVs; 15 SNVs per Mb) or 100 coding insertions or deletions (indels; three indels per Mb)), no cases were detected in subtype 3, whereas subtype 2 demonstrated a significantly higher frequency of hypermutated tumors than subtype 3 ($P = 0.014$, $q = 0.086$) (Fig. 2d and Supplementary Table 7). We analyzed the distribution of microbial subtypes by birth year. Notably, a statistically significant difference in subtype proportions was observed for individuals born after 1980 ($P = 0.018$, Fisher's exact test), characterized by a marked increase in the prevalence of subtype 3 (Fig. 2e). No significant differences in tumor purity or copy number variation profiles were observed among subgroups (Supplementary Fig. 2).

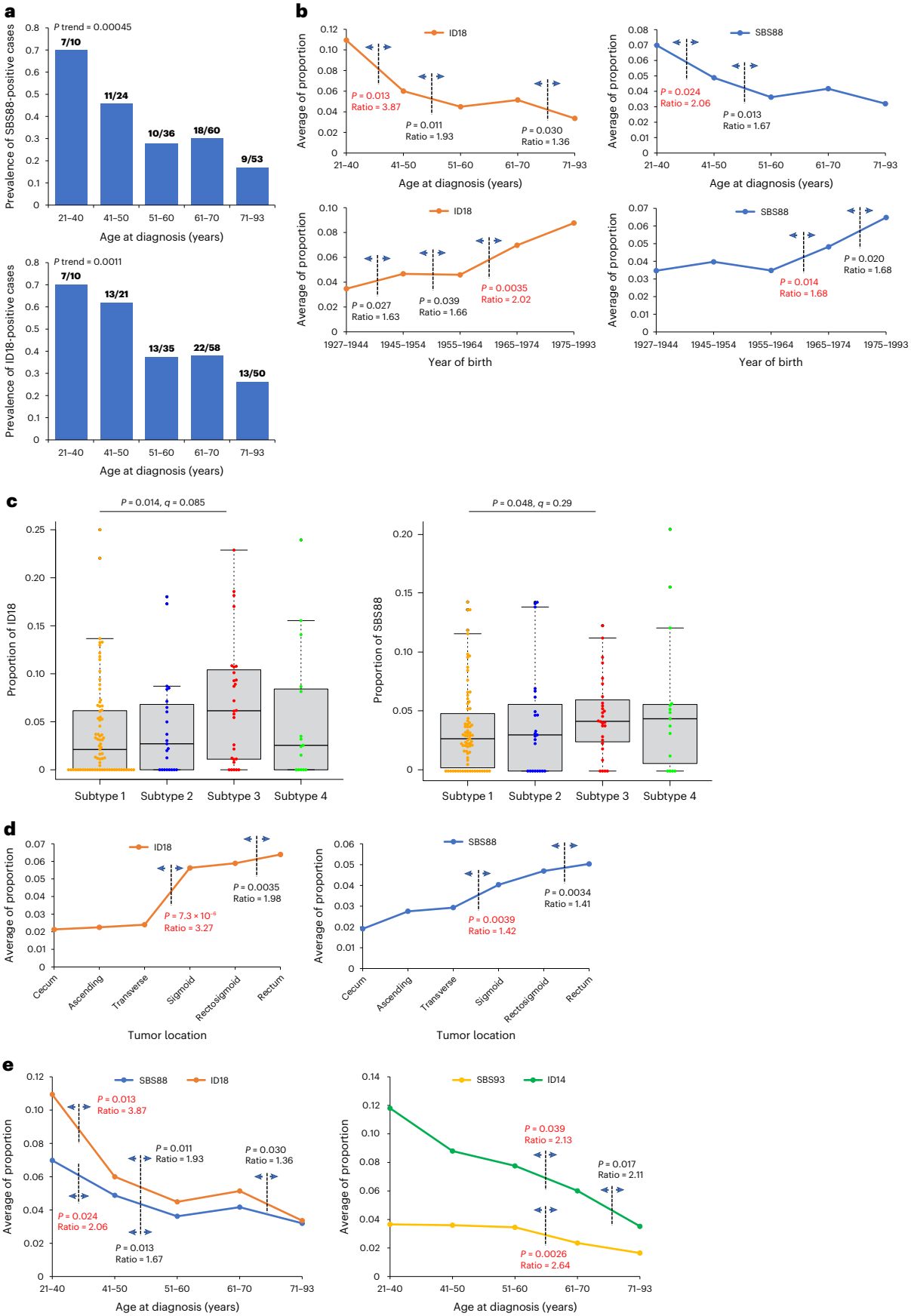
Analysis of mutational signatures and their correlation with the gut microbiome

WGS data from 200 CRC cases were subjected to mutational signature analysis. As the mutational profiles differed markedly between non-hypermutated and hypermutated tumors, these groups were analyzed separately. Significantly mutated genes in this cohort were identified based on previously reported mutation calls[21,22], comprising 19 genes in non-hypermutated CRCs and 14 genes in hypermutated cases (Supplementary Tables 9 and 10). To enhance the resolution and robustness of signature decomposition, somatic mutation data from an additional 953 CRC cases from the Mutographs project[15] were integrated, resulting in a combined cohort of 1,153 cases for comprehensive analysis. Decomposition of SBS and short ID spectra identified 21 SBS (Extended Data Fig. 3a,b) and seven ID (Extended Data Fig. 4a,b) signatures in 183 non-hypermutated cases, and 11 SBS and two ID signatures in 17 hypermutated cases (Extended Data Fig. 5). We observed a significant enrichment of mutational signatures associated with defective DNA mismatch repair, specifically SBS44 and SBS26, in subtype 2 compared with subtype 3 ($P = 0.015$, $q = 0.086$ for SBS44; $P = 0.014$, $q = 0.086$ for SBS26).

Remarkably, these signatures (SBS88 and/or ID18) exerted a pronounced impact on the mutational landscape of CRC in this Japanese cohort, being detected in 82 of 183 non-hypermutated cases (44.8%) (Fig. 3a and Extended Data Figs. 3c and 4c). The colibactin-associated mutational signatures (SBS88 and ID18) showed a significant correlation with the presence of colibactin-specific sequence motifs[12,13,23,24] (Fig. 3b,c). This prevalence increased among younger patients (Fig. 4a). A previous study showed that the prevalence of these mutational signatures among Japanese CRCs is higher than that in cohorts from other countries[15]. By contrast, no colibactin-associated signatures were observed in hypermutated CRCs (Extended Data Fig. 5 and Supplementary Table 11). Among patients

Fig. 4 | Clinicopathological and microbiome associations of colibactin-associated mutational signatures in colorectal cancer. a, Prevalence of SBS88-positive (top panel) and ID18-positive (bottom panel) cases according to age at diagnosis. The prevalence of both signatures was higher in younger patients, consistent with the greater contributions of these signatures shown in **b**. *P* values for trend across age groups were calculated using the Cochran–Armitage trend test. **b**, Association between colibactin-associated signatures (ID18 and SBS88) and age at diagnosis (upper panel) or birth year (lower panel). 'Ratio' indicates the median fold increase between two groups separated by a black dashed line. For group comparisons stratified by the years indicated by the black dashed lines, the highest ratio and corresponding *P* value are shown in red. *P* values were calculated using a two-sided Wilcoxon rank-sum test. The median fold increases in the relative contributions of SBS88 and ID18 were greatest in younger patients (≤40 years) and in individuals born after 1965. The numbers of patients in the age-at-diagnosis groups of ≤40, 41–50, 51–60, 61–70 and ≥71 years were 10, 24, 36, 60 and 53, respectively. The numbers of patients in the birth-year groups of ≤1944, 1945–1954, 1955–1964, 1965–1974 and ≥1975 were 44, 65, 37, 24 and 12, respectively. **c**, Association between colibactin-associated mutational signatures and microbiome-based subtypes. Patients classified as subtype 3 exhibited increased colibactin-associated ID18 signatures compared with other subtypes, particularly subtype 1 (left). Association between the colibactin-associated SBS88 signature and microbiome-defined CRC subtypes (right). Patients classified as subtype 3 showed higher SBS88 contributions compared with subtype 1. *P* values were calculated using a two-sided Wilcoxon rank-sum test, and *q* values were adjusted for multiple testing using the Benjamini–Hochberg FDR method. The numbers of patients assigned to subtypes 1, 2, 3 and 4 were 75, 23, 28 and 15, respectively. The boxes represent the 25th–75th percentiles, the horizontal lines indicate the medians and the whiskers extend to the maximum and minimum values within 1.5× the interquartile range. **d**, Association between colibactin-associated mutational signatures and tumor location. Ratio indicates the fold change (median ratio) between groups separated by a black dashed line. For both SBS88 and ID18, significant differences were observed in tumors arising in the sigmoid colon and rectum, indicating a higher prevalence of colibactin-associated signatures in the distal colorectum. *P* values were calculated using a two-sided Wilcoxon rank-sum test. For group comparisons stratified by the anatomical locations indicated by the solid lines, the highest ratio and corresponding *P* value are shown in red. The numbers of tumors located in the cecum, ascending colon, transverse colon, sigmoid colon, rectosigmoid junction and rectum were 10, 26, 10, 36, 30 and 61, respectively. **e**, Relationship between co-occurring mutational signatures and age at diagnosis. Ratio denotes the median fold change between the two groups separated by a black dashed line. For comparisons stratified by age at diagnosis, the ratio and corresponding *P* value for the comparison showing the greatest ratio are highlighted in red. *P* values were calculated using the Wilcoxon rank-sum test. Significant co-occurrence was observed between the colibactin-associated signatures (SBS88 and ID18) and between SBS93 and ID14. Although all four signatures tended to be more frequent in early-onset CRC, SBS88 and ID18 were most strongly enriched in patients aged ≤40 years, whereas SBS93 and ID14 were most frequent among those aged ≤60 years.

https://doi.org/10.1038/s41588-026-02692-x



https://doi.org/10.1038/s41588-026-02692-x

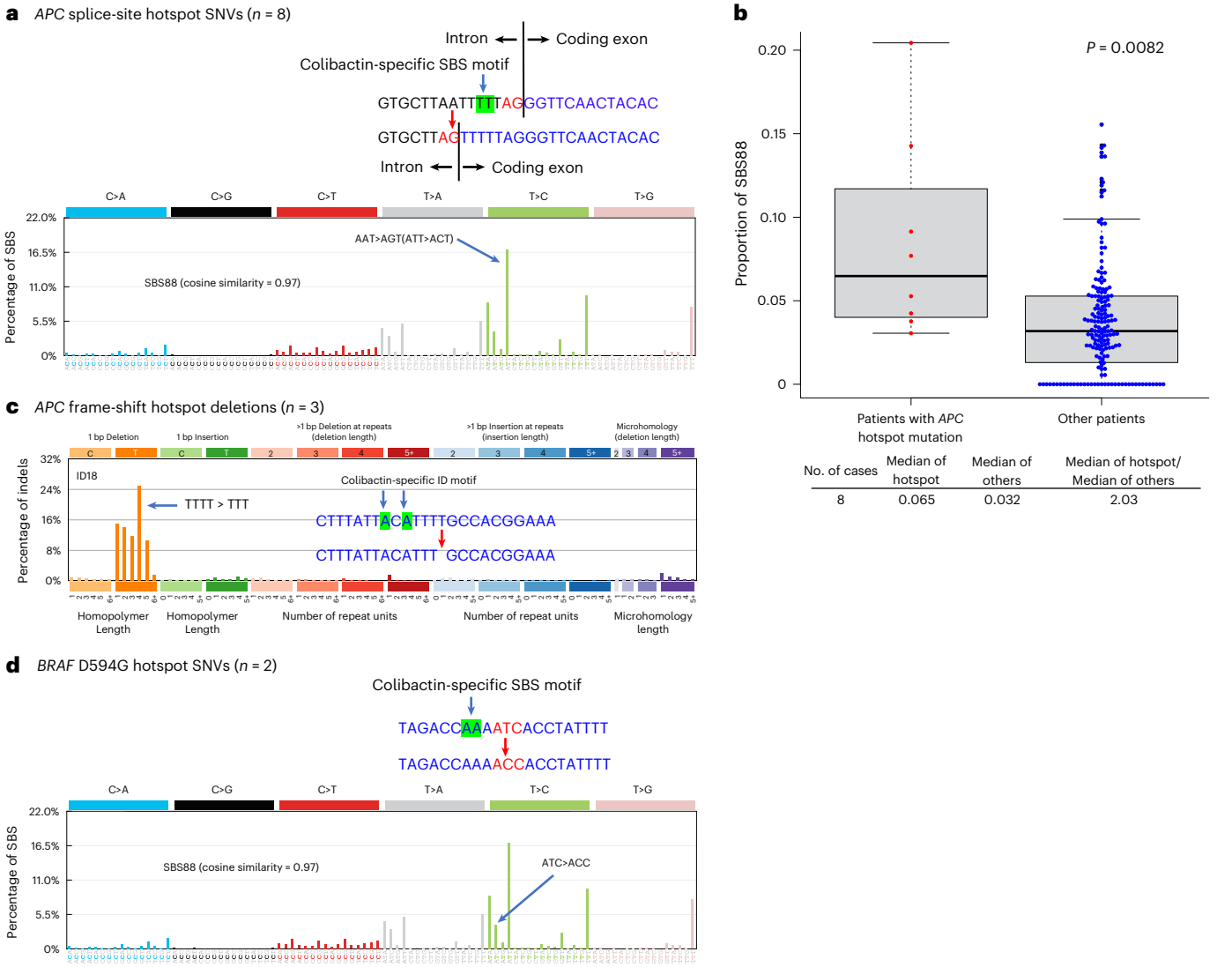


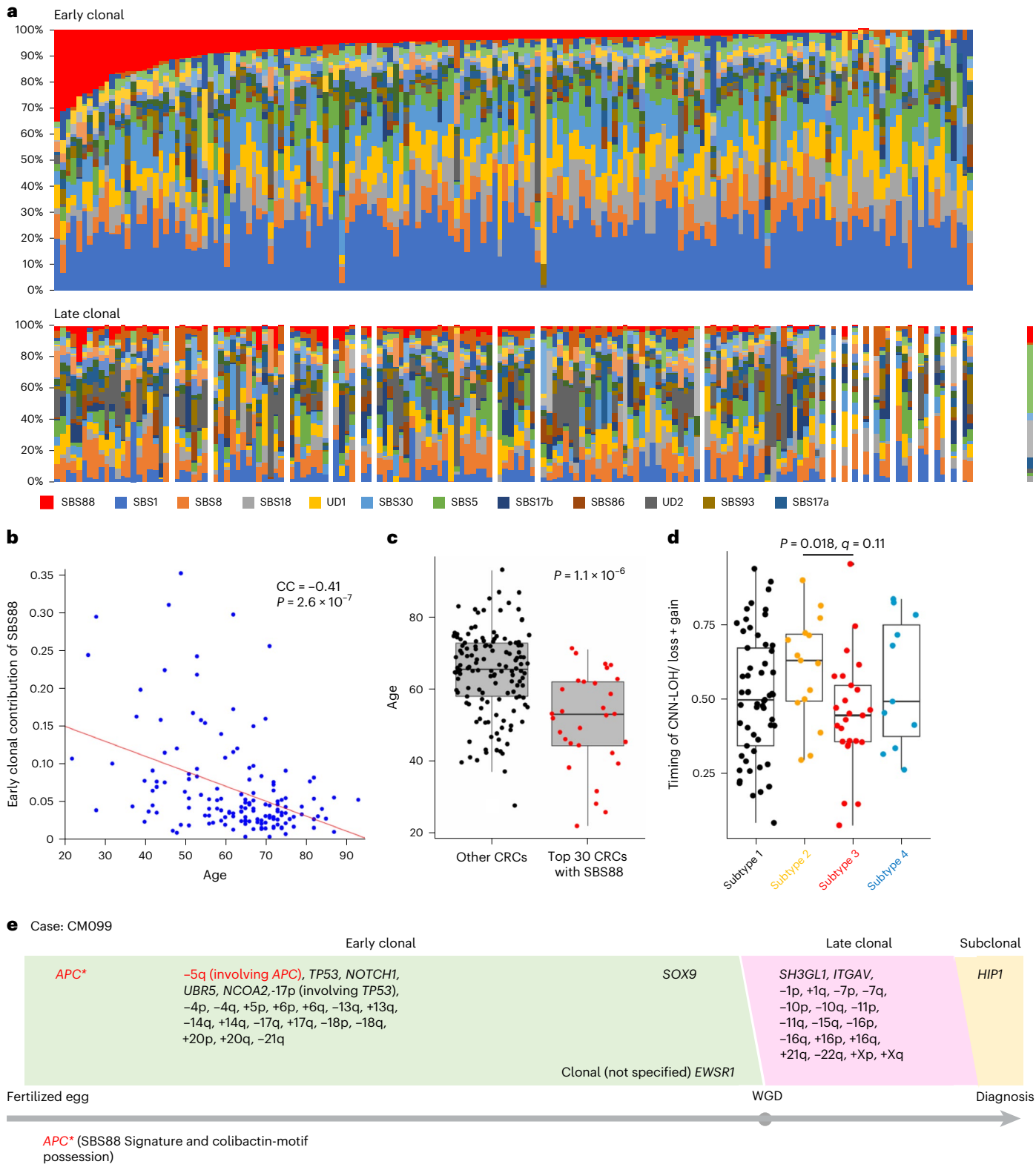
Fig. 5 | Colibactin-associated driver hotspot mutations in colorectal cancer. **a**, *APC* hotspot mutation associated with the SBS88 signature. This *APC* splice-site hotspot mutation exhibited the AAT > AGT (ATT > ACT) substitution pattern characteristic of SBS88 together with the colibactin-specific sequence motif TT(AA) located at positions −3 and −4 upstream of the mutation site. The hotspot resides within intron 8 of *APC*, creating a novel AG splice site that introduces a seven-base insertion into the coding sequence, resulting in a frameshift. **b**, Correlation between *APC* hotspot mutations and the colibactin-associated SBS88 signature. Patients harboring hotspot mutations in intron 8 of *APC* exhibited significantly higher SBS88 contributions than other patients ($P = 0.0082$, two-sided Wilcoxon rank-sum test). The boxes represent the 25th–75th percentiles, the horizontal lines indicate the medians and the whiskers extend to the most extreme data points within 1.5× the interquartile range. **c**, *APC* ID18 hotspot mutations. Among 16 *APC* frameshift deletions exhibiting the ID18 mutational pattern, seven represented recurrent hotspot deletions. The representative somatic deletion shown was identified in three independent CRC cases. This recurrent *APC* frameshift mutation (TTTT > TTT), representing a single thymine deletion (−1 T) within a 1–5-bp thymine homopolymer region, is a defining feature of the colibactin-associated indel mutational signature ID18. **d**, *BRAF* hotspot mutation. The D594G hotspot mutation exhibited both the characteristic SBS88 trinucleotide substitution pattern and the colibactin-specific sequence motif.

diagnosed with non-hypermutated CRC before the age of 50, particularly those younger than 40, the contribution of colibactin-associated mutational signatures was significantly elevated (patients younger than 40 years: ID18, $P = 0.013$; SBS88, $P = 0.024$) (Fig. 4b and Supplementary Table 12). When compared with subtype 1, subtype 3 likewise exhibited a significant increase in these signatures (ID18, $P = 0.014$, $q = 0.085$; SBS88, $P = 0.048$, $q = 0.29$) (Fig. 4c). Furthermore, the relative contribution of ID18 was markedly increased in patients with left-sided or rectal tumors and in those with stage II or higher disease ($P = 7.3 \times 10^{-6}$, 0.0035 and 0.0029, respectively) (Fig. 4d and Supplementary Tables 13–15). Analysis across birth cohorts revealed a distinct rise in the prevalence of colibactin-associated signatures among individuals born after 1965 ($P = 0.0035$ for ID18; $P = 0.014$ for SBS88) (Fig. 4b, Extended Data Fig. 6a and Supplementary Table 16). Consistently, these signatures were significantly enriched in CRCs diagnosed before age 50, with the highest prevalence observed in patients younger than 40 years (Fig. 4a).

SBS30 was observed in most cases in this cohort (Fig. 3a and Extended Data Fig. 3c). Although *NTHL1* germline mutations have historically been proposed as the underlying cause of SBS30, recent evidence suggests that this signature more commonly reflects alkylating DNA damage in CRC[25,26]. No germline or somatic mutations in the *NTHL1* gene were identified in this dataset. Consistent with this interpretation, SBS30 was significantly less frequent in younger patients at diagnosis (Supplementary Table 12).

Our analyses further identified significant co-occurrence between the colibactin-associated mutational signatures SBS88 and ID18, as well as between SBS93 and ID14 ($P = 2.5 \times 10^{-12}$ and 6.6×10^{-6}, respectively)

https://doi.org/10.1038/s41588-026-02692-x



(Supplementary Table 17). These mutational processes were more frequently observed in early-onset CRC overall[14]. Specifically, SBS88 and ID18 were significantly enriched among patients aged ≤40 years, whereas SBS93 and ID14 were more prevalent in patients aged ≤60 years (Fig. 4e, Extended Data Fig. 6b and Supplementary Table 12).

To evaluate the contribution of colibactin-associated mutational processes to recurrent driver alterations in this cohort, we estimated the expected probability of each mutational signature for individual mutations based on their trinucleotide context, following a previously published method[27] (see Methods). Among non-hypermutated CRCs, 15.3% of patients (28 out of 183) harbored mutations with a high predicted probability and/or mutational pattern consistent with the colibactin signature in common driver genes, of which 78.6% (22 out of 28) affected *APC* (Supplementary Table 18). Notably, hotspot mutations in the *APC* splice site, *BRAF* D594G and three *TP53* variants exhibited high predicted probabilities for SBS88 and characteristic

https://doi.org/10.1038/s41588-026-02692-x

Fig. 6 | A comparison of copy number aberration timing in CRC based on mutational signature and gut microbiota-based subtype. a, Proportion of SBS signatures in early and late clonal mutations. The distribution of SBS signatures was assessed in both early and late clonal mutations. The upper panel displays CRC cases with a high contribution of SBS88 among early clonal mutations, arranged from left to right in descending order of SBS88 proportion. The lower panel shows the same set of patients in the identical left-to-right order to enable direct comparison between early and late clonal phases. Patients with fewer than 50 SNVs in either phase were excluded from the analysis. **b**, The relationship between age and the proportion of SBS88. A significant positive association was observed between younger age and a higher proportion of SBS88. The *P* value was calculated using a two-sided Pearson correlation test. **c**, Comparison of patient age between cases with high and low SBS88 contribution. This box-and-whisker plot compares the ages of the top 30 CRC cases with the highest SBS88 contribution to those of the remaining cohort. Patients with the highest SBS88 contributions were significantly younger than the remaining cases ($P = 1.1 \times 10^{-6}$, one-sided Wilcoxon rank-sum test). Boxplots: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× the interquartile range; points, individual cases. **d**, Correlation between gut microbiome-based CRC subtypes and the sequential acquisition of copy number gains. The temporal order of three classes of copy number gain events was compared across gut microbiome-defined CRC subtypes: (1) copy number neutral loss of heterozygosity (CNN-LOH) or loss followed by gain (N:0); (2) monoallelic gain (N:1); and (3) biallelic gain (N:M). Among these categories, subtype 3 exhibited a significantly earlier occurrence of CNN-LOH/loss + gain events compared with subtype 2. Boxplots: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× the interquartile range; points, individual cases. Statistical significance was determined using a two-sided Wilcoxon rank-sum test, with the *P* value adjusted for multiple testing using the Benjamini–Hochberg FDR method. **e**, Timeline of somatic evolution in a representative CRC case (CM099T) characterized by *APC* hotspot mutations and colibactin-associated mutational signatures and sequence motifs. This schematic illustrates the temporal sequence of mutational and genomic events observed in case CM099T. Notably, it shows biallelic (two-hit) inactivation of key driver genes, including *APC* and *TP53*, together with chromosomal-level alterations that occurred during the early clonal phase preceding whole-genome duplication (WGD). '−' denotes copy number loss, and '+' denotes copy number gain.

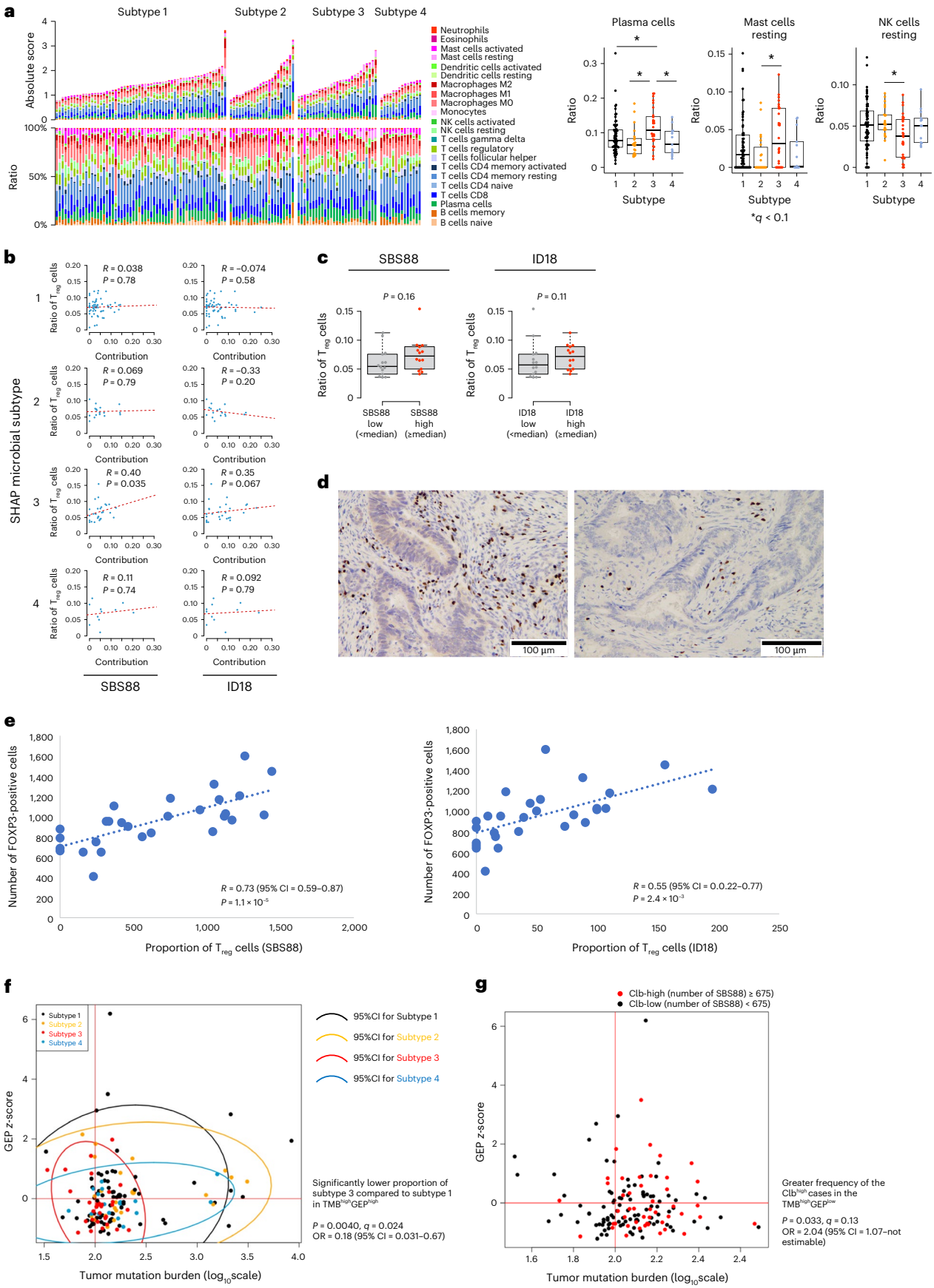
trinucleotide contexts enriched for colibactin-associated sequence motifs (Supplementary Table 18). The *APC* splice-site mutation, located in intron 8, generates a novel AG acceptor site, resulting in a 7 bp insertion and frameshift (Fig. 5a). This recurrent alteration, previously reported[13], was identified in eight out of 183 non-hypermutated CRCs (4.4%), seven of which also harbored a second inactivating alteration or loss of heterozygosity (LOH) of *APC* (Supplementary Table 19). These patients exhibited a significantly higher contribution of SBS88 compared with other cases ($P = 0.0082$) (Fig. 5b), supporting a causal role of colibactin in generating this hotspot.

In addition, 14 patients exhibited 16 *APC* frameshift deletions attributable to ID18 (Supplementary Table 18), seven of which contained the canonical colibactin-associated ID18 motif, and seven represented recurrent hotspot deletions (Fig. 5c). Two cases carried the *BRAF* D594G hotspot mutation (Fig. 5d)—previously associated with younger age, left-sided or rectal localization and non-hypermutated CRC[28]—consistent with the clinical and mutational features observed in these patients. Overall, the expected probability of SBS88 was significantly enriched among non-synonymous substitutions in cancer driver genes within the Japanese cohort ($P = 4.3 \times 10^{-7}$). Although ID18 was not globally enriched across driver events, it was specifically overrepresented in *APC* ($P = 0.014$, odds ratio 2.2, 95% CI 1.1–3.9).

Using WGMS, we detected the presence of the *clbP* gene, which encodes the atypical peptidase ClbP essential for colibactin biosynthesis[29], in 16 (11.9%) of the fecal samples (Extended Data Fig. 7a). Quantitative PCR (qPCR) confirmed a strong correlation between *clbP* copy number in stool and that estimated by WGMS (Extended Data Fig. 7b). By contrast, no correlation was observed

Fig. 7 | Immune environment in CRCs based on mutational signature and gut microbiota-based subtype. a, Immune cell composition inferred by CIBERSORTx across CRC subtypes (left). CIBERSORTx was used to estimate the relative proportions of immune cell populations in all CRC cases with available RNA-seq data. The analysis revealed no significant differences in overall immune cell composition among the CRC subtypes. Comparative analyses of inflammatory cell fractions by subtype are shown in Extended Data Fig. 10a,b. Immune cell types showing significant differences among CRC subtypes (right). Across the 22 immune and inflammatory cell fractions inferred by CIBERSORTx, plasma cells, resting mast cells and resting natural killer (NK) cells showed significant differences among subtypes, as assessed using a one-sided Wilcoxon rank-sum test. The boxes represent the 25th–75th percentiles, the horizontal lines indicate the medians and the whiskers extend to the maximum and minimum values within 1.5× the interquartile range. **b**, Correlation between colibactin-associated mutational signatures (SBS88 and ID18) and the proportion of T_{reg} cells across microbiome-defined CRC subtypes. Among the four subtypes, a positive correlation was observed exclusively in subtype 3, as assessed using Spearman's rank correlation test. **c**, Comparison of T_{reg} cell ratios in subtype 3 stratified by the median contributions of SBS88 and ID18. Within subtype 3, cases were divided into two groups based on whether their SBS88 and ID18 mutational signature contributions were above or below the respective medians. Boxplots depict the distribution of T_{reg} cell ratios in each group. Although the differences did not reach statistical significance using a one-sided Wilcoxon rank-sum test, cases with higher SBS88 and ID18 contributions exhibited a trend toward higher proportions of T_{reg} cells. The boxes represent the 25th–75th percentiles, the horizontal lines indicate the medians, and the whiskers extend to the most extreme data points within 1.5× the interquartile range. **d**, Representative immunohistochemical staining for FOXP3, a transcription factor marker of T_{reg} cells. Shown are representative images from two CRC cases: CM040, exhibiting pronounced SBS88 and ID18 mutational signatures (left), and CM053, displaying minimal colibactin-associated signatures (right). The CM040 case demonstrates a higher density of FOXP3-positive lymphocytes. **e**, Quantitative immunohistochemical analysis of FOXP3 expression in 28 CRC cases classified as subtype 3. For each case, five representative regions containing tumor-infiltrating lymphocytes were randomly selected and imaged at ×200 magnification. FOXP3-positive lymphocytes were manually enumerated within each field of view. Quantitative assessment revealed a significant positive correlation between the number of FOXP3-positive cells and the proportion of T_{reg} cells estimated by CIBERSORTx, for both the SBS88 mutational signature ($R = 0.73$, 95% CI 0.59–0.87, $P = 1.1 \times 10^{-5}$; two-sided Pearson correlation test) and the ID18 mutational signature ($R = 0.55$, 95% CI 0.22–0.77, $P = 2.4 \times 10^{-3}$; two-sided Pearson correlation test). **f**, Correlation between microbiome-defined subtypes, tumor mutational burden (TMB) and T cell-associated inflammatory gene expression profile (GEP) signatures. Tumors classified as $TMB^{high}GEP^{high}$—indicative of an immune-responsive phenotype—were significantly less frequent in subtype 3 than in subtypes 1 and 2 (subtype 3 vs subtype 1: $P = 0.0040$, $q = 0.024$, odds ratio (OR) 0.18, 95% CI 0.031–0.67; subtype 3 vs subtype 2: $P = 0.042$, $q = 0.13$, OR 0.22, 95% CI 0.032–1.09). *P* values were calculated using a two-sided Fisher's exact test, and *q* values were adjusted for multiple testing using the Benjamini–Hochberg FDR method. These results also suggest that subtype 3, characterized by a high contribution of the SBS88 signature, represents an immune 'cold' tumor phenotype. **g**, Correlation between colibactin-associated mutational signatures, TMB and GEP signatures. After excluding hypermutated CRC cases, patients with a high SBS88 mutation count (≥675; mean value) exhibited a significantly higher proportion of $TMB^{high}GEP^{low}$ tumors than those with a low SBS88 mutation count ($P = 0.033$; one-sided Fisher's exact test). After adjustment for multiple testing using the Benjamini–Hochberg FDR method, the association remained suggestive ($q = 0.13$, OR 2.04, 95% CI 1.07–not estimable). When applying the mean ID18 mutation count (51) as the threshold, no significant association was observed ($P = 0.20$, OR 1.42, 95% CI 0.76–not estimable). Detailed analyses of the relationship between TMB and GEP for ID18 are presented in Extended Data Fig. 10c.

https://doi.org/10.1038/s41588-026-02692-x



https://doi.org/10.1038/s41588-026-02692-x

Table 1 | Clinical data and genomic abnormalities in colorectal cancer subtypes categorized based on gut microbiota

	Classification based on gut microbiota			
	Subtype 1	**Subtype 2**	**Subtype 3**	**Subtype 4**
Gut microbial composition		*F. nucleatum* ↑ *P. stomatis* ↑	*P. stomatis* ↑	*F. nucleatum* ↑
Age			Younger	
Stage	Early	Advanced	Advanced	
Hypermutated phenotype			Less	
Mutational signatures		MMR-related signatures	Colibactin-related signatures	
LOH timing		Late	Early	
Microenvironment			Immunosuppressive	

MMR, mismatch repair

between *clbP* abundance in stool and the contribution of colibactin-associated mutational signatures (Extended Data Fig. 7c). The presence of *clbP* was further examined in CRC tissues using qPCR[30–32], which revealed *clbP* detection in all tumor samples as well as in all 12 matched non-cancerous colonic mucosa specimens obtained from regions distant from the primary CRC lesion (Supplementary Table 20). However, the *clbP* copy number in stool showed no correlation with that in matched tumor tissues (Extended Data Fig. 8a,b). Among tumors, the three cases with the highest *clbP* abundance exhibited a marked contribution of colibactin-associated signatures (Extended Data Fig. 8c). Nevertheless, most tumors with prominent colibactin signatures displayed low *clbP* copy numbers (Extended Data Fig. 8d), indicating that the prevalence of colibactin-producing genes in a single sample does not necessarily reflect the cumulative mutational burden induced by colibactin.

Timing of colibactin signatures and their association with early-onset CRC

We analyzed the distribution of early and late clonal and subclonal mutations associated with specific mutational signatures by using previously outlined methodology[33]. By comparing the contributions of early and late clonal mutations as well as clonal and subclonal mutations across different mutational signatures, the analysis revealed that colibactin-related mutational signatures demonstrate a greater prevalence of early clonal mutations than late clonal and subclonal mutations (Fig. 6a and Extended Data Fig. 9a). Among the 149 cases with higher SBS88 in the clonal early phase than in the clonal late phase, younger cases acquired a significantly greater SBS88 contribution in the clonal early phase (correlation coefficient = −0.41, $P = 2.6 \times 10^{-7}$) (Fig. 6a). The 30 patients with the highest burden of early clonal SBS88 mutations were significantly younger than the remaining cohort ($P = 1.1 \times 10^{-6}$) (Fig. 6c). This observation is consistent with previous reports indicating that individuals exposed to colibactin develop CRC at a younger age[15,34].

Using somatic mutations and copy number variations, we were able to estimate the relative timing of three types of copy number gains: (1) copy number neutral loss of heterozygosity (CNN-LOH) / loss + gain (N:0); (2) monoallelic gains (N:1); and (3) biallelic gains (N:M)[33]. Subtype 3 demonstrated an earlier occurrence of CNN-LOH / loss + gain than subtype 2 ($P = 0.018$, $q = 0.11$) irrespective of the patient's age (Fig. 6d). Monoallelic gains showed no significant association, and although biallelic gains were nominally significant between subtype 2 and subtype 3 ($P = 0.044$), these associations did not remain significant after correction for multiple testing (Extended Data Fig. 9b,c).

An in-depth case study of a CRC patient in subtype 3 highlighted the significant involvement of colibactin signature SBS88 (approximately 10%) in the patient's disease progression. Figure 6e displays the findings of CM099, a 54-year-old female with stage II left-sided CRC, a non-smoker, with a non-hypermutated phenotype. This patient showed an *APC* hotspot splice-site mutation that matched the SBS88 signature and the colibactin-specific motif annotated as early clonal. It has been previously reported[35] that cases with a high SBS88 contribution exhibited mutational activity before the age of 10 years. In this case as well, it was presumed that the *APC* mutation associated with the SBS88 signature occurred during the first decade of life. Further high-resolution analyses, such as single-cell sequencing, will be required to precisely delineate the temporal onset of this mutational event[36]. Nearly simultaneously, the case experienced *TP53* (nonsense), *NOTCH1* (missense) and chr17p loss, followed by whole-genome duplication with a mean relative timing of 0.76 (95% CI 0.75–0.76) (around 41 years old). Additionally, the case exhibited various chromosomal gains and losses (Supplementary Tables 21 and 22).

Immune environment in colibactin-associated CRC

The gut microbiota has an impact on both the local gut immune environment and the broader systemic immune system. We investigated the correlation between microbiome-based subtypes and immune cell composition as determined by CIBERSORTx (Fig. 7a and Extended Data Fig. 10a,b). Analysis of immune cell composition demonstrated that plasma cells were significantly enriched in subtype 3 relative to subtype 1 ($P = 0.011$, $q = 0.032$), subtype 2 ($P = 0.0012$, $q = 0.0073$) and subtype 4 ($P = 0.029$, $q = 0.057$). Resting mast cells were also significantly more abundant in subtype 3 than in subtype 2 ($P = 0.012$, $q = 0.074$). By contrast, resting natural killer cells were significantly depleted in subtype 3 relative to subtype 2 ($P = 0.016$, $q = 0.095$) (Fig. 7a and Supplementary Table 23).

We next investigated the relationship between microbial subtypes and regulatory T cells (T_{reg} cells). A positive correlation between the proportion of T_{reg} cells and the contributions of the SBS88 and ID18 mutational signatures was observed only in subtype 3 (Fig. 7b). Furthermore, within subtype 3, stratification by the median values of SBS88 and ID18 revealed that tumors with higher contributions of these signatures tended to have an increased proportion of T_{reg} cells (Fig. 7c). Immunohistochemical staining for FOXP3 was conducted on 28 CRC cases classified as subtype 3. For each case, five representative tumor-infiltrated regions were randomly selected and imaged at ×400 magnification (Fig. 7d). FOXP3-positive lymphocytes were enumerated within these fields. Quantitative assessment demonstrated that the density of FOXP3-positive cells exhibited a significant positive association with the mutational signature SBS88 ($R = 0.73$, 95% CI 0.59–0.87, $P = 1.1 \times 10^{-5}$) and the indel signature ID18 ($R = 0.55$, 95% CI 0.22–0.77, $P = 2.4 \times 10^{-5}$) (Fig. 7e).

The integration of tumor mutation burden and T cell-inflamed gene expression profile (GEP) score[37] is useful for classifying tumor immune environments. In the $TMB^{high}GEP^{high}$ (immune-hot) category, the proportion of subtype 3 patients was significantly lower than that of subtype 1 after adjustment for multiple testing ($P = 0.0040$, $q = 0.024$, odds ratio 0.18, 95% CI 0.031–0.67) (Fig. 7f). After excluding hypermutated CRC cases, patients with a high SBS88 mutation

count (≥675, mean value) exhibited a significantly higher proportion of $TMB^{high}GEP^{low}$ tumors compared with patients with a low SBS88 mutation count ($P = 0.033$) (Fig. 7g). After adjustment for multiple testing, the association remained suggestive ($q = 0.13$, odds ratio 2.04, 95% CI 1.07–not estimable). No significant association was observed for ID18 mutations when applying a mean cutoff of 51 ($P = 0.20$, odds ratio 1.42, 95% CI 0.76–not estimable) (Extended Data Fig. 10c).

Discussion

To our knowledge, this study represents the first integrative analysis combining stool-derived metagenomic profiling with WGS of CRC tissues. Through this approach, patients were stratified into four microbiome-based subtypes. Subtype 3, characterized by a low contribution of *F. nucleatum* in stool as determined by WGMS, was paradoxically associated with younger age and a higher prevalence of colibactin-related mutational signatures. By contrast, subtype 2, characterized by a high relative abundance of *F. nucleatum* and *P. stomatis*, was associated with advanced stages of CRC. These findings highlight that distinct microbial consortia shape CRC pathogenesis through heterogeneous mechanisms, and that the genomic alterations underlying carcinogenesis and CRC progression vary among subtypes (Table 1).

The colibactin-associated mutational signature was observed in ~45% of non-hypermutated CRCs in Japan, with a particularly high prevalence among younger patients. The frequency of this signature exceeded that reported in other countries[12,14,15], and birth cohort analyses demonstrated a striking increase in individuals born after 1965 (Fig. 4b and Extended Data Fig. 6a). In addition to validating previously reported *APC* mutations harboring the colibactin motif[13], this study newly identified *BRAF* mutations with the exact imprint.

Our analysis revealed significant co-occurrence between SBS88 and ID18, as well as between SBS93 and ID14. Although previous studies have found that SBS93 and ID14 occur more frequently in early-onset CRC[14], our findings demonstrate that SBS88 and ID18 are significantly enriched in patients younger than 40 years, whereas SBS93 and ID14 are more commonly observed in those younger than 60 years. Collectively, these results suggest that the etiological mechanisms driving SBS93/ID14-associated CRCs are probably distinct from those underlying SBS88/ID18-associated cases.

The disproportionately high burden of colibactin-associated mutations in Japanese CRCs, particularly among those born after 1965, remains unexplained. Early postnatal microbiota colonization is critical for immune development, and perturbations during this window may have long-lasting effects[38]. In the United States, carriage of *pks*$^+$ *E. coli* among breastfed term infants peaked at 38% between 6 months and 12 months but declined thereafter[39]. By contrast, preterm infants exposed to neonatal intensive care and antibiotics exhibited higher and more persistent carriage. Pediatric antibiotic use in Japan has historically been extensive, often for minor infections. Since the late 1970s, second-generation cephalosporins and macrolides have been widely prescribed at a frequency exceeding that of many Western countries[40]. Although speculative, such early-life exposures may have promoted both the persistence of *pks*$^+$ *E. coli* and the accumulation of colibactin-associated mutations, thereby contributing to the rising incidence of early-onset CRC in Japan. The temporal increase in SBS88 emerged after 1965, whereas the rise in subtype 3 became evident after 1985, indicating an approximate 20 year lag. Although this chronological discrepancy cannot be ascribed to a single causal factor, the marked escalation in antibiotic consumption in Japan beginning in the late 1970s may have further altered the intestinal milieu exposed to colibactin-producing bacteria, thereby potentially amplifying colibactin-mediated carcinogenic processes. Nonetheless, other environmental factors, including dietary and lifestyle changes, may act in concert to drive this epidemiological trend.

Interestingly, subtype 3 showed significant T_{reg} cell infiltration, suggesting an immunosuppressive tumor microenvironment distinct from other microbial subtypes. The predominance of T_{reg} cells may reflect immune modulation driven by *pks*$^+$ *E. coli*, which produces the genotoxin colibactin that induces DNA alkylation damage and senescence-associated inflammatory signaling, thereby promoting T_{reg} cell differentiation through IL-6, IL-10 and TGF-β pathways[12,41,42]. Such microbiota-driven T_{reg} cell expansion may suppress cytotoxic T cell activity and contribute to the 'cold' immune phenotype characteristic of this subtype. However, most supporting evidence derives from correlative analyses and murine models, and direct causal links between microbial exposure, T_{reg} cell accumulation and immune-cold phenotypes in human tumors remain limited[43]. Further integrative studies combining microbial genomics and spatial immune profiling will be required to clarify these mechanisms in CRC.

No significant correlation was observed between *clbP* expression in tumor tissues and the colibactin-associated mutational signatures SBS88 and ID18. This probably reflects a temporal discordance between the two measurements: mutational signatures capture cumulative, historical exposure, whereas *clbP* abundance represents the bacterial load at the time of sampling[15]. Although mutational signatures provide a durable record of past exposure to colibactin-producing bacteria, the intestinal microbiota is inherently dynamic and subject to temporal fluctuations driven by host and environmental factors. Consistent with this interpretation, a previous publication[44] reported no significant differences in *clbB* prevalence among patients with CRC, those with adenomas and healthy individuals.

This study has limitations. First, the average sequencing depth for tumor samples (56×) was modest compared to the >100× depths increasingly used in recent studies[14]. Second, differences in plasma cells, resting mast cells and resting natural killer cells were observed between gut microbiota-based subtypes using CIBERSORTx-based inference of immune cell composition. However, these findings should be interpreted with caution and will require validation using orthogonal approaches such as single-cell or spatial transcriptomic analyses.

In summary, this study identifies a potential mechanism underlying the elevated incidence of CRC in Japan and its increasing frequency among younger individuals. These findings suggest that preventive strategies aimed at limiting colonization by *pks*$^+$ bacteria and suppressing colibactin production[45] may represent viable approaches to reducing the global burden of CRC.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-026-02692-x.

References

1. Morgan, E. et al. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut* **72**, 338–344 (2023).
2. Sung, H. et al. Colorectal cancer incidence trends in younger versus older adults: an analysis of population-based cancer registry data. *Lancet Oncol.* **26**, 51–63 (2025).
3. Akimoto, N. et al. Rising incidence of early-onset colorectal cancer—a call to action. *Nat. Rev. Clin. Oncol.* **18**, 230–243 (2021).
4. Sinicrope, F. A. Increasing incidence of early-onset colorectal cancer. *N. Engl. J. Med.* **386**, 1547–1558 (2022).
5. White, M. T. & Sears, C. L. The microbial landscape of colorectal cancer. *Nat. Rev. Microbiol.* **22**, 240–254 (2024).
6. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).

7. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
8. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
9. Bossuet-Greif, N. et al. The colibactin genotoxin generates DNA interstrand cross-links in infected cells. *mBio* **9**, https://doi.org/10.1128/mbio.02393-17 (2018).
10. Iftekhar, A. et al. Genomic aberrations after short-term exposure to colibactin-producing *E. coli* transform primary colon epithelial cells. *Nat. Commun.* **12**, 1003 (2021).
11. Thomas, C. E. et al. Epidemiologic factors in relation to colorectal cancer risk and survival by genotoxic colibactin mutational signature. *Cancer Epidemiol. Biomarkers Prev.* **33**, 534–546 (2024).
12. Pleguezuelos-Manzano, C. et al. Mutational signature in colorectal cancer caused by genotoxic *pks*$^+$ *E. coli*. *Nature* **580**, 269–273 (2020).
13. Rosendahl Huber, A. et al. Improved detection of colibactin-induced mutations by genotoxic *E. coli* in organoids and colorectal cancer. *Cancer Cell* **42**, 487–496.e6 (2024).
14. Cornish, A. J. The genomic landscape of 2,023 colorectal cancers. *Nature* **633**, 127–139 (2024).
15. Diaz-Gay, M. et al. Geographic and age variations in mutational processes in colorectal cancer. *Nature* **643**, 230–240 (2025).
16. Rynazal, R. et al. Leveraging explainable AI for gut microbiome-based colorectal cancer classification. *Genome Biol.* **24**, 21 (2023).
17. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
18. Vogtmann, E. et al. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLoS ONE* **11**, e0155362 (2016).
19. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
20. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
21. Totoki, Y. et al. Multiancestry genomic and transcriptomic analysis of gastric cancer. *Nat. Genet.* **55**, 581–594 (2023).
22. Yachida, S. et al. Comprehensive genomic profiling of neuroendocrine carcinomas of the gastrointestinal system. *Cancer Discov.* **12**, 692–711 (2022).
23. Boot, A. et al. Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types. *Genome Res.* **30**, 803–813 (2020).
24. Georgeson, P. et al. Genotoxic colibactin mutational signature in colorectal cancer is associated with clinicopathological features, specific genomic alterations and better survival. Preprint at *medRxiv* https://doi.org/10.1101/2023.03.10.23287127 (2024).
25. Gurjao, C. et al. Discovery and features of an alkylating signature in colorectal cancer. *Cancer Discov.* **11**, 2446–2455 (2021).
26. Gurjao, C. et al. Germline predisposition to oncogenic alkylating damage in colorectal cancer. *Cancer Epidemiol. Biomarkers Prev.* **35**, 420–428 (2026).
27. Chen, B. et al. Contribution of *pks*$^+$ *E. coli* mutations to colorectal carcinogenesis. *Nat. Commun.* **14**, 7827 (2023).
28. Sahin, I. H. & Klostergaard, J. BRAF mutations as actionable targets: a paradigm shift in the management of colorectal cancer and novel avenues. *JCO Oncol. Pract.* **17**, 723–730 (2021).
29. Zhou, T. et al. Screening method toward ClbP-specific inhibitors. *Genes Environ.* **45**, 8 (2023).
30. Kawanishi, M. et al. In vitro genotoxicity analyses of colibactin-producing *E. coli* isolated from a Japanese colorectal cancer patient. *J. Toxicol. Sci.* **44**, 871–876 (2019).
31. Kawanishi, M. et al. Genotyping of a gene cluster for production of colibactin and in vitro genotoxicity analysis of *Escherichia coli* strains obtained from the Japan Collection of Microorganisms. *Genes Environ.* **42**, 12 (2020).
32. Hirayama, Y. et al. Activity-based probe for screening of high-colibactin producers from clinical samples. *Org. Lett.* **21**, 4490–4494 (2019).
33. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
34. Severino, A. & Ianiro, G. Early-life colibactin exposure is associated with early-onset colorectal cancer occurrence. *Gastroenterology* **169**, 1305 (2025).
35. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
36. Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).
37. Ayers, M. et al. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127**, 2930–2940 (2017).
38. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539–544 (2016).
39. Levy, S. et al. Colibactin genes are highly prevalent in the developing infant gut microbiome. *Gut Microbes* **18**, 2604874 (2025).
40. Hsia, Y. et al. Consumption of oral antibiotic formulations for young children according to the WHO Access, Watch, Reserve (AWaRe) antibiotic groups: an analysis of sales data from 70 middle-income and high-income countries. *Lancet Infect. Dis.* **19**, 67–75 (2019).
41. Cougnoux, A. et al. Small-molecule inhibitors prevent the genotoxic and protumoural effects induced by colibactin-producing bacteria. *Gut* **65**, 278–285 (2016).
42. Arthur, J. C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
43. Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The microbiome, cancer, and cancer therapy. *Nat. Med.* **25**, 377–388 (2019).
44. Shimpoh, T. et al. Prevalence of *pks*-positive *Escherichia coli* in Japanese patients with or without colorectal cancer. *Gut Pathog.* **9**, 35 (2017).
45. Volpe, M. R. et al. A small molecule inhibitor prevents gut bacterial genotoxin production. *Nat. Chem. Biol.* **19**, 159–167 (2023).

[1]Department of Cancer Genome Informatics, Graduate School of Medicine, The University of Osaka, Osaka, Japan. [2]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. [3]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives (OTRI), The University of Osaka, Osaka, Japan. [4]Japan Agency for Medical Research and Development–Core Research for Evolutional Science and Technology (AMED–CREST), Tokyo, Japan. [5]School of Life Science and Technology, Institute of Science Tokyo, Tokyo, Japan. [6]Innovative Microbiome Therapy Research Center, Juntendo University Graduate School of Medicine, Tokyo, Japan. [7]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [8]Department of Gastroenterological Surgery, Graduate School of Medicine, The University of Osaka, Osaka, Japan. [9]Department of Gastroenterology and Hepatology, Graduate School of Medicine, The University of Osaka, Osaka, Japan. [10]Endoscopy Division, National Cancer Center Hospital, Tokyo, Japan. [11]Department of Pathology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [12]Department of Respiratory Medicine and Clinical Immunology, Graduate School of Medicine, The University of Osaka, Osaka, Japan. [13]Department of Colorectal Surgery, National Cancer Center Hospital, Tokyo, Japan. [14]Department of Biomedical Statistics, Graduate School of Medicine, The University of Osaka, Osaka, Japan. [15]These authors contributed equally: Satoshi Shiba, Shinichi Yachida, Sayaka Mizutani, Yasushi Totoki, Hiromi Nakamura. ✉e-mail: syachida@cgi.med.osaka-u.ac.jp; takuji@bio.titech.ac.jp; tshibata@ims.u-tokyo.ac.jp

Methods

Patients and samples

This study was approved by the Research Ethics Committees of the National Cancer Center, the University of Osaka and Institute of Science Tokyo as meeting the ethical guidelines for medical and health research involving human individuals (National Cancer Center Institutional Review Board, 2013-244; Research Ethics Committee, the University of Osaka, 20064-2: Human Subjects Research Ethics Review Committee, Institute of Science Tokyo, 2014018). Written informed consent was obtained from all participants before enrollment. All data were de-identified before analysis, and a double-anonymization procedure was applied to all participant identifiers.

This study enrolled a total of 600 individuals, comprising 454 patients with histologically confirmed CRC and 146 HCs, who were recruited at the National Cancer Center Hospital (Tokyo, Japan) and the University of Osaka Hospital (Osaka, Japan) between 2014 and 2025. Among these, cohort 1 consisted of cases prospectively collected specifically for the present study, whereas cohort 2 comprised cases previously described in our published work[8] (Extended Data Fig. 1a).

These studies were designed to be observational and included participants for whom the biological specimens required for comprehensive analyses—such as stool samples, tumor tissues and matched normal DNA—were available between 2014 and 2025. Individuals with hereditary CRC syndromes, including familial adenomatous polyposis and hereditary non-polyposis CRC, or with inflammatory bowel disease, were excluded from the analysis.

WGS data were obtained from primary CRC tissues in 200 cases from cohort 1. WGMS data derived from stool samples were available for a total of 395 patients with CRC and 146 HCs across both cohorts. Detailed information on participants' lifestyle and dietary habits was collected using a comprehensive self-administered questionnaire comprising 475 items across 25 pages. The questionnaire was developed based on the framework of the Japan Public Health Center-based Next Generation Study (JPHC-NEXT)[46].

Stool sample collection and processing

Individuals classified as HC were those who presented with a positive fecal occult blood test during a hospital visit and subsequently underwent colonoscopy for secondary screening, with no clinically significant findings. Stool samples were obtained from HCs and CRC patients before the initiation of any treatment, including surgical resection or chemotherapy. The first stool passed at the hospital on the day of colonoscopy was collected for analysis[47,48]. In most cases, the stool samples collected on the day of colonoscopy were solid. Consistent with our previous observations[47], the microbial taxonomic composition showed a high degree of concordance with that of standard frozen stool samples obtained without bowel preparation (Pearson's correlation coefficient = 0.91, $P < 0.01$). All stool samples were immediately frozen on dry ice and subsequently stored at −80 °C until DNA extraction. Participants were instructed to consume a low-residue diet on the day preceding colonoscopy. On the day of examination, all participants received a bowel-cleansing agent followed by colonoscopic evaluation.

Tumor and matched normal sample collection

Tumor tissues were obtained from surgically resected specimens of patients with CRC. Matched non-tumor DNA was primarily extracted from peripheral blood lymphocytes; in selected cases, non-neoplastic colonic mucosa collected from regions anatomically distant from the primary tumor was also used as a normal control.

Clinical endpoint definitions

Overall survival was defined as the interval from the initiation of treatment to death from any cause, with censoring at the date of last confirmed follow-up for surviving patients (Supplementary Table 1). PFS was defined as the interval between the initiation of treatment and the first documented evidence of disease progression, as determined by serial radiological examinations (for example, computed tomography or magnetic resonance imaging) or by clinical confirmation of tumor recurrence or metastasis.

Microbial subtype classification and reproducibility assessment

For the combined test dataset (cohorts 1 + 2), comprising 199 CRC cases from our previous publication[8] and 196 CRC cases newly collected from the National Cancer Center Hospital and the University of Osaka Hospital, we applied a previously published microbial subtype classification algorithm[16]. Cases were categorized into four microbial subtypes, and their bacterial community compositions were compared to assess reproducibility.

Baseline clinicopathological characteristics were summarized across subtypes, and overall survival was analyzed using the Kaplan–Meier method with comparisons performed with the log-rank test. These univariate comparisons assessed differences in survival that may reflect both subtype-related and background clinical variations. To determine whether microbial subtypes independently influence prognosis, we further applied multivariate Cox proportional hazards models adjusted for established prognostic covariates.

WGMS of fecal samples

Genomic DNA was extracted from frozen stool samples using a bead-beating protocol[49] in conjunction with the GNOME DNA Isolation Kit (MP Biomedicals). DNA integrity and concentration were evaluated using the Agilent 4200 TapeStation system (Agilent Technologies). Following ethanol precipitation, the purified DNA was resuspended in TE buffer and stored at −80 °C until library preparation. Sequencing libraries were constructed from fecal DNA using the Nextera XT DNA Sample Prep Kit (Illumina) according to the manufacturer's instructions. WGMS was performed on the NovaSeq 6000 platform (Illumina) with paired-end (PE) reads (2 × 150 bp), targeting an average sequencing depth of 5.0 Gb per sample.

Metagenomic sequencing quality control and taxonomic profiling

In total, 9,875,823,911 PE reads (51,843,293 on average) from cohort 1 and 10,156,584,408 PE reads (51,038,112 on average) from cohort 2 were generated from 150-bp sequencing and subjected to stringent quality control. Raw reads containing undetermined bases ('N') were removed. Reads containing bacteriophage *phiX* sequences were identified and filtered by alignment to the reference genome using Bowtie2 (v.2.2.9) with the '–fast-local' preset. Adaptor and primer sequences were trimmed using Cutadapt (v.1.9.1) with the following parameters: for forward reads, -a CTGTCTCTTATACACATCTCCGAGCCCACGAGAC -O 33 -q 17; for reverse reads, -a CTGTCTCTTATACACATCTGACGCTGCCGACGA -O 32 -q 17. Within Cutadapt, reads with consecutive quality scores of ≤17 were trimmed at the 3′ end, and reads shorter than 50 bp after trimming were discarded. Reads with an average Phred quality score of ≤25 were also excluded. High-quality reads were aligned to the human reference genome (GRCh38; gi 568336000–568336023) using Bowtie2 (v.2.2.9), and all reads mapping to the human genome were removed. Unpaired reads were subsequently discarded. After these filtering steps, a total of 8,777,929,480 PE reads (46,899,434 on average) from cohort 1 and 8,896,915,982 PE reads (44,708,121 on average) from cohort 2 were retained for downstream analyses (hereafter referred to as high-quality reads). Quality-controlled reads were processed using MetaPhlAn3 with default parameters to generate species-level and genus-level taxonomic profiles[50].

Identification of CRC subtypes based on metagenome-derived SHAP values

Construction of a CRC classifier using a random forest model. A random forest model was constructed to estimate the likelihood of CRC, hereafter referred to as the 'microbial CRC trait score', which was calculated for each sample. The model was trained on four independent metagenomic cohorts[7,17–19] curated through the curatedMetagenomicData R package[20], applying filtering thresholds of 1×10^{-5} for taxonomic abundance and 0.95 for prevalence. Model training and evaluation were performed using the scikit-learn library (v.1.2.2). Model performance was assessed by tenfold cross-validation, yielding an AUC of 0.84. The trained classifier was subsequently applied to MetaPhlAn3-derived species-level relative abundance profiles from 196 CRC samples (cohort 1) and 146 HCs previously reported[8]. This evaluation yielded an AUC of 0.74, confirming consistent discriminatory performance in an independent dataset (Extended Data Fig. 1b).

SHAP analysis and subtype identification. Interpretable artificial intelligence frameworks such as SHAP and LIME (local interpretable model-agnostic explanations) have been developed to elucidate the internal decision-making mechanisms of complex predictive models[51]. Among these, SHAP provides a mathematically rigorous and reproducible framework for feature attribution, quantifying the contribution of each input variable to the model's output. SHAP analysis is grounded in cooperative game theory, in which Shapley values represent the fair distribution of contributions among all predictors. In this study, SHAP was used to interpret the internal decision process of the random forest model and to identify distinct microbial features associated with CRC. To further characterize CRC heterogeneity, unsupervised clustering of the SHAP values was performed using the *k*-means algorithm. The optimal number of CRC subtypes was determined by minimizing the within-cluster sum of squares (Extended Data Fig. 2a). Given that SHAP does not inherently account for potential confounding factors such as age, tumor stage or primary tumor site, we systematically examined these variables across SHAP-derived clusters to evaluate potential confounding influences. The reproducibility of the SHAP-based subtype classification was validated in an independent cohort of 199 CRC cases (cohort 2) (Extended Data Fig. 2b,c). All analyses were performed using the shapmat Python package (https://github.com/ryzary/shapmat).

WGS analysis of CRC

Genomic DNA was extracted from primary tumor tissues and matched peripheral blood samples. WGS libraries with an average insert size of 550 bp were prepared from 2 µg of genomic DNA using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina). Sequencing was performed on the NovaSeq 6000 platform (Illumina) with 150 bp PE reads. The median sequencing depth was 52× (mean, 56×) for tumor samples and 32× (mean, 34×) for matched normal samples. The median estimated tumor purity was 0.52 (range, 0.11–0.97), and all cases met the inclusion criterion of ≥10% tumor purity.

Mutation calling

PE reads were aligned to the human reference genome (GRCh37) using BWA-MEM[52]. PCR duplicates were removed by eliminating paired reads that mapped to identical genomic coordinates, and pileup files were subsequently generated using SAMtools[53]. Somatic point mutations, including SNVs and short indels, were identified according to the following eight criteria: (1) mapping quality of ≥20 and (2) base quality of ≥10. Candidate somatic mutations were further filtered by applying the following conditions: (3) in each tumor sample, variants were required to be supported by at least four reads when the tumor variant allele frequency (TVAF) was ≥0.15, or by at least eight reads when 0.15 > TVAF ≥ 0.05, with at least one supporting read having a base quality ≥30; (4) the variant allele frequency (VAF) of the matched non-tumor sample had to be <0.03, with a minimum read depth of eight. To reduce context-dependent sequencing errors, reads from all non-tumor samples were aggregated to identify recurrent false-positive sites. For each genomic position with a non-tumor sequence depth of ≥10 and VAF < 0.2, the aggregate non-tumor VAF (NVAF) was calculated. Variants were retained only if (5) NVAF < 0.03 for sites with TVAF ≥ 0.15 or NVAF < 0.01 for sites with 0.15 > TVAF ≥ 0.05, and (6) the ratio of TVAF to NVAF was ≥20. To exclude germline single-nucleotide polymorphisms, (7) the proportion of non-tumor samples with a VAF ≥ 0.1 was required to be <0.002. Finally, (8) variants showing extreme strand bias (>95% of reads derived from a single strand) were excluded.

Copy number analysis

Copy number ratios, tumor purity and ploidy were estimated using Battenberg (v.2.2.10)[54] with default parameters. The inferred copy number ratios were subsequently adjusted according to the estimated tumor purity for each sample to obtain purity-corrected copy number profiles.

Significantly mutated gene analysis

Significantly mutated genes were identified using three complementary approaches: the inactivation bias method[22], the activation bias method and dNdScv[55]. For the inactivation bias test, the number of samples harboring inactivating mutations (nonsense, read-through, splice-site or frameshift variants) was compared with the number carrying alternative mutations using Fisher's exact test. Conversely, for the activation bias test, the number of activating (hotspot missense) mutations was compared with the number of other mutations using Fisher's exact test. Genomic positions exhibiting at least two identical mutations were defined as mutational hotspots. Mutations occurring within 5 bp of a hotspot were also classified as hotspot mutations. As hotspot missense mutations are generally activating but may occasionally be inactivating or functionally ambiguous, genes with $P < 0.05$ by the inactivation bias method were designated as putative tumor suppressor genes, and their activation bias *P* values were fixed at 1. Multiple hypothesis testing was corrected using the Benjamini–Hochberg false discovery rate procedure, and adjusted *P* values are reported as *q* values. Genes were considered significantly mutated if any of the three tests yielded a *q* value of <0.1.

Mutational signature analysis

De novo decomposition of SBS and small ID mutational signatures was performed using SigProfilerExtractor on WGS data from 1,153 CRC cases, comprising 200 Japanese CRC cases from the present study and additional non-Japanese CRC cases reported previously[15]. Given that mutational spectra differ substantially between hypermutated and non-hypermutated tumors, de novo decomposition was conducted separately for these two groups, following the approach described in a previous study[15]. The number of decomposed signatures was determined based on the optimal solution estimated by SigProfilerExtractor. For each decomposed signature, cosine similarity was calculated relative to the COSMIC reference signature set. Signatures with cosine similarity of ≥0.75 were assigned to known COSMIC signatures, whereas those with cosine similarity of <0.75 were interpreted as novel or undecomposed composite signatures and classified as undetermined.

For the 200 CRC cases analyzed in this study, reconstruction accuracy was evaluated by comparing the reconstructed and observed mutational spectra. SBS signatures demonstrated consistently high reconstruction fidelity (cosine similarity of ≥0.97 in all cases). By contrast, two cases exhibited ID signature cosine similarity values of <0.9, both of which contained a very small number of indels (44 and 68, respectively), indicating insufficient data for reliable signature decomposition. These two cases were therefore excluded from subsequent ID signature analyses.

Definition of colibactin-associated mutational signatures

The presence of a colibactin-associated mutational signature was defined as a relative contribution of SBS88 of ≥5% or ID18 of ≥5%. This threshold was adopted because ten out of 82 colibactin-exposed cases exhibited only one of the two signatures (SBS88 or ID18). Moreover, a 5% cutoff has been widely applied in prior investigations, including the first report characterizing the colibactin-associated mutational process[12]. Given that mutational signature decomposition algorithms may artifactually assign low-frequency components, contributions below 5% are generally regarded as unreliable and are typically truncated to zero, whereas contributions ≥5% are considered robust signals.

Expected probability of mutational signatures

The expected probability of each mutational signature for a given trinucleotide context in an individual patient was calculated using the following equations:

$$N_{ijk} = S_{ik} \times C_{ij}$$

$$P_{ijk} = N_{ijk} / \sum_{i} N_{ijk}$$

where S_{ik} denotes the number of mutations attributed to mutational signature i in patient k, C_{ij} represents the proportion of trinucleotide context j within mutational signature i, N_{ijk} is the estimated number of mutations corresponding to mutational signature i and trinucleotide context j in patient k, and P_{ijk} is the expected probability of observing mutational signature i in trinucleotide context j for patient k (ref. 27).

Clonal analysis of mutational signatures

Mutations were categorized as clonal or subclonal using MutationTimeR (v.1.00.2)[33]. Mutations that occurred before copy number gains were annotated as clonal (early); those that arose following copy number gains were annotated as clonal (late); and mutations for which timing could not be determined were designated as clonal (NA). For each mutation group, mutational signature decomposition was performed using the deconstructSigs R package (v.1.9.0)[56]. The parameter signature 'signature.cutoff' was set to zero, and 'signatures.ref' was assigned to the SBS and ID signatures extracted by SigProfilerExtractor, as described above. Clonal (early), clonal (late) and clonal (NA) categories were collectively defined as clonal mutations. Cases with fewer than 50 mutations in any group were excluded from the analysis to ensure robust signature estimation.

Timing analysis of copy number gains

The relative timing of three classes of copy number gains was inferred from (1) CNN-LOH / loss + gain (N:0), (2) monoallelic gains (N:1) and (3) biallelic gains (N:M). Analyses were performed using MutationTimeR (v.1.00.2)[33], which integrates somatic SNVs and copy number variation data derived from WGS. For each tumor, copy number timing estimates were generated using genomic segments containing at least 100 somatic mutations to ensure reliable inference. Weighted averages of mutation counts were computed across segments for each copy number gain category. Comparative analyses among the four CRC subtypes were performed after exclusion of hypermutated cases. To further reconstruct the temporal sequence of genomic events within individual tumors, PhylogicNDT SinglePatientTiming (v.1.0; https://doi.org/10.1101/508127) was applied. The inferred relative timing of events was scaled from zero, representing the zygotic (fertilization) stage, to one, corresponding to the time of clinical diagnosis.

Transcriptome sequencing (RNA-seq)

Total RNA was extracted from fresh–frozen tumor tissues using the miRNeasy Mini Kit (Qiagen) according to the manufacturer's protocol. RNA integrity was assessed using the BioAnalyzer system (Agilent Technologies), and samples with an RNA integrity number of >6.0 were selected for sequencing. RNA-seq libraries were prepared using the SureSelect Strand-Specific RNA Library Preparation Kit (Agilent Technologies) with 300 ng of total RNA as input. Sequencing was performed on the HiSeq 2500 platform (Illumina) to generate 101 bp PE reads. Reads were aligned to the human reference genome (GRCh37/hg19) using STAR[57], and gene-level read counts were quantified against the UCSC human transcriptome reference using HTSeq[58]. Expression levels were calculated as fragments per kilobase of exon per million mapped fragments, incorporating strand-specific information. Differential gene expression was assessed using the Wilcoxon rank-sum test.

Estimation of non-human bacterial reads in tumor tissues

We used the kneaddata tool (v.0.12.0)[59] to eliminate human reads from human rRNA-depleted total RNA sequencing data of tumor tissues, encompassing human and bacterial transcripts. Relative abundance of bacterial reads was calculated as the ratio of non-human reads to all reads in each sample. The total RNA was extracted from fresh–frozen tumor tissues using the miRNAeasy kit (Qiagen). A total of 250 ng of total RNA was used to generate libraries for total RNA sequencing. The Universal Plus Total RNA-seq kit with human rRNA depletion (Tecan) was used for library preparation, followed by sequencing on a NovaSeq 6000 in PE 150 base mode.

Decomposition of tumor immune environments

Gene expression profiles, annotated using GENCODE v43lift37, were deconvoluted to estimate tumor immune cell composition. The proportions of 22 immune cell types were inferred using CIBERSORTx based on the LM22 reference signature matrix, which includes naive and memory B cells, plasma cells, $CD8^+$ T cells, naive and memory $CD4^+$ T cells (resting and activated) and follicular helper T cells. Additional immune cell subsets analyzed comprised T_{reg} cells, γδ T cells, resting and activated natural killer cells, monocytes, M0, M1 and M2 macrophages, resting and activated dendritic cells, resting and activated mast cells, eosinophils and neutrophils. CIBERSORTx was executed in 'B-mode' with the 'disable quantile normalization' option enabled, as recommended by the developers for RNA-seq data. Only samples with a CIBERSORTx output P value of <0.05 were retained for downstream analyses, ensuring robust estimation of immune cell fractions.

T cell-inflamed GEP analysis

The T cell-inflamed GEP score[60] was calculated as the weighted sum of the normalized expression values of 18 genes, each multiplied by its corresponding coefficient: CCL5 (0.008346), CD27 (0.072293), CD274 (0.042853), CD276 (−0.023900), CD8A (0.031021), CMKLR1 (0.151253), CXCL9 (0.074135), CXCR6 (0.004313), HLA-DQA1 (0.020091), HLA-DRB1 (0.058806), HLA-E (0.071750), IDO1 (0.060679), LAG3 (0.123895), NKG7 (0.075524), PDCD1LG2 (0.003734), PSMB10 (0.032999), STAT1 (0.250229) and TIGIT (0.084767). The resulting composite score reflects the degree of T cell-mediated inflammation within the tumor microenvironment.

Quantification of the *clbP* gene by qPCR

Quantification of the *clbP* gene (copies per ml) in genomic DNA extracted from tumor tissues or fecal samples was performed by Adenoprevent[32]. qPCR was conducted using the KAPA SYBR Fast qPCR Kit (Roche Molecular Systems) on the Thermal Cycler Dice Real-Time System (Takara Bio). The copy number of the *clbP* gene was determined from cycle threshold values using a standard calibration curve generated from *clbP* gene standards (Adenoprevent)[30,31].

Immunohistochemistry

Formalin-fixed, paraffin-embedded tissue sections (4 µm thick) were deparaffinized and treated with 3% hydrogen peroxide for 15 min to

quench endogenous peroxidase activity. Antigen retrieval was performed using an autoclave in 10 mM citrate buffer (pH 6.0) for 10 min. Slides were incubated with a mouse monoclonal anti-FOXP3 antibody (clone 236 A/E7; dilution 1:100; ab20034, Abcam) for 1 h at room temperature (20–25 °C), followed by detection using the EnVision system (Dako) with mouse linker (Dako), according to the manufacturer's protocol. Signal visualization was achieved using the Chem-Mate EnVision detection method (Dako). Immunohistochemical staining for FOXP3 was performed in 28 CRC cases classified as subtype 3. For each case, five representative tumor regions exhibiting lymphocytic infiltration were randomly selected and imaged at ×200 magnification. FOXP3-positive lymphocytes were manually counted within each field.

Statistical analyses

When multiple comparison groups were present, overall differences were assessed using the Kruskal–Wallis test. Pairwise comparisons between individual groups were subsequently performed using the Wilcoxon rank-sum test or Fisher's exact test, as appropriate. Correction for multiple hypothesis testing was performed using the Benjamini–Hochberg false discovery rate procedure, and adjusted *P* values are reported as *q* values.

Reporting summary

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

Data availability

The raw sequencing data reported in this study have been deposited in the DNA Data Bank of Japan (DDBJ), Shizuoka, Japan, under the following accession numbers: PRJDB4176, PRJDB18241, PRJDB18242, PRJDB38070, JGAS000696 and JGAS000872. Access to these datasets is controlled because they contain potentially identifiable human genomic information. Qualified researchers may obtain access by submitting an application to the National Bioscience Database Center (NBDC) through the DDBJ portal, subject to review and approval in accordance with applicable ethical and legal requirements. Detailed application procedures are available at https://humandbs.dbcls.jp/en/data-use.

Code availability

All analyses were conducted using publicly available software. The callallSV tool is freely accessible for non-commercial use at https://github.com/ma9606/callallSV under the GPL-3.0 license. Details of all other software used in this study are provided in Methods and the accompanying Reporting Summary.

References

46. Tsugane, S. & Sawada, N. The JPHC study: design and some findings on the typical Japanese diet. *Jpn J. Clin. Oncol.* **44**, 777–782 (2014).
47. Nishimoto, Y. et al. High stability of faecal microbiome composition in guanidine thiocyanate solution at room temperature and robustness during colonoscopy. *Gut* **65**, 1574–1575 (2016).
48. Erawijantari, P. P. et al. Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. *Gut* **69**, 1404–1415 (2020).
49. Furet, J. P. et al. Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. *FEMS Microbiol. Ecol.* **68**, 351–362 (2009).
50. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).
51. Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F. & Simsekler, M. C. E. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. *BMC Med. Inform. Decis. Mak.* **25**, 110 (2025).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
55. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
56. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
57. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
58. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
59. McIver, L. J. et al. bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235–1237 (2018).
60. Cristescu, R. et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* **362**, eaar3593 (2018).

Acknowledgements

We express our deepest gratitude to all patients and their families for their participation in this study. We are also indebted to N. Sezawa, A. Ohira, H. Sekiguchi, R. Itsuki (National Cancer Center Hospital), A. Hara, H. Chikuta, A. Matsumoto, W. Mukai, T. Urushidate (National Cancer Center Research Institute) and A. Yanase (The University of Osaka) for their dedicated contributions to sample collection, clinical data acquisition, and specialized technical support. We further acknowledge the pathological expertise of T. Hashimoto and T. Haga (National Cancer Center Hospital), the bioinformatics expertise of F. Salim and K. Hirota and the assistance of the NCC Biobank for specimen management and storage. We are also grateful to M. Díaz-Gay and L.B. Alexandrov (University of California, San Diego) for generously providing data from the Mutographs project. Computations were performed on the SHIROKANE supercomputer at the Institute of Medical Science, University of Tokyo.

Author contributions

S.Y. and T.S. conceived and led the study. S.Y., Y.T., S.H. and T.S. designed the research framework. S.M., Y.T., H.N., N.H., N.M., Y.A., M.S.-A., H.K., Y.H., H.T., K.T., R.H., H.R., S.I. and Y.K. performed integrative genomic and microbiome analyses. S.Y., S.M., Y.T., Y.H., T.Y. and T.S. interpreted the data. S.S., S.Y., S.M., Y.T., H.N., N.H., M.S.-A. and T.S. drafted the paper. Y.D., H.E. and Y.S. provided overall supervision. All authors critically reviewed and approved the final paper.

Funding

This work was supported by grants from Practical Research for Innovative Cancer Control from the Japan Agency for Medical Research and Development (AMED) (JP25ck0106800 to S.Y. and T.S.; JP22ck0106546 to S.Y., Y.S. and T.Y.; JP25ck0106799 to S.Y., Y.S. and T.Y.; JP26ck0106874 to S.Y. and T.S.; JP26ck0106162 to S.Y. and T.S.); Project for Cancer Research and Therapeutic Evolution (P-CREATE) from AMED (JP21cm0106477 to S.S., S.Y. and T.Y.); Project for Cancer Research Endowment and Advanced Tertiary Execution (P-PROMOTE) from AMED (JP25ama221430 to S.Y., Y.S. and T.Y.); Core Research for Evolutional Science and Technology (AMED–CREST) from AMED (JP25gm2010009 to S.Y., Y.S. and T.Y.); AMED-ASPIRE form AMED (JP26jf0126022 to S.Y. and S.M.); United States–Japan Cooperative Medical Science Program from AMED (JP23jk0210009 to S.Y. and T.S.); JSPS KAKENHI (JP16H06279 to S.S.; JP18K15800 to S.S.; JP 22K16336

to S.S.; JP 20H03662 to S.Y.; JP 23H02892 to S.Y.; JP 25K21771 to S.Y.); JST-AIP Acceleration Research (JPMKCR19U3 to S.Y. and T.Y.); the National Cancer Center Research and Development Fund (2023-A-05 to T.Y. and T.S.; 2020-A-4 to S.Y. and Y.S.; 2023-A-6 to Y.S.; 2023-A-7 to T.Y. and T.S.; 2026-A-09 to S.Y., H.T. and Y.S.); Joint Research Project of the Institute of Medical Science, the University of Tokyo (to S.Y. and T.S.); the Takeda Science Foundation (to S.Y.); the Yasuda Memorial Medical Foundation (to S.Y. and T.S.); the Mitsubishi Foundation (to S.Y.); the Princess Takamatsu Cancer Research Fund (to S.Y.); Yakult Bio-Science Foundation (to S.Y.); Project Mirai Cancer Research Grants (to S.Y.); and PHarma-Link between Academia and Shionogi (FLASH 2019, to S.Y.).

Competing interests

T.Y. is a co-founder of Metagen, Metagen Therapeutics and digzyme and holds equity interests in these companies. These entities had no role in the design of the study; the collection, analysis or interpretation of the data; or the writing of the paper and decision to publish. In accordance with its conflict-of-interest policies, the Institute of Science Tokyo manages the terms of these arrangements. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-026-02692-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-026-02692-x.

Correspondence and requests for materials should be addressed to Shinichi Yachida, Takuji Yamada or Tatsuhiro Shibata.

Peer review information *Nature Genetics* thanks Ruben van Boxtel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-026-02692-x

a

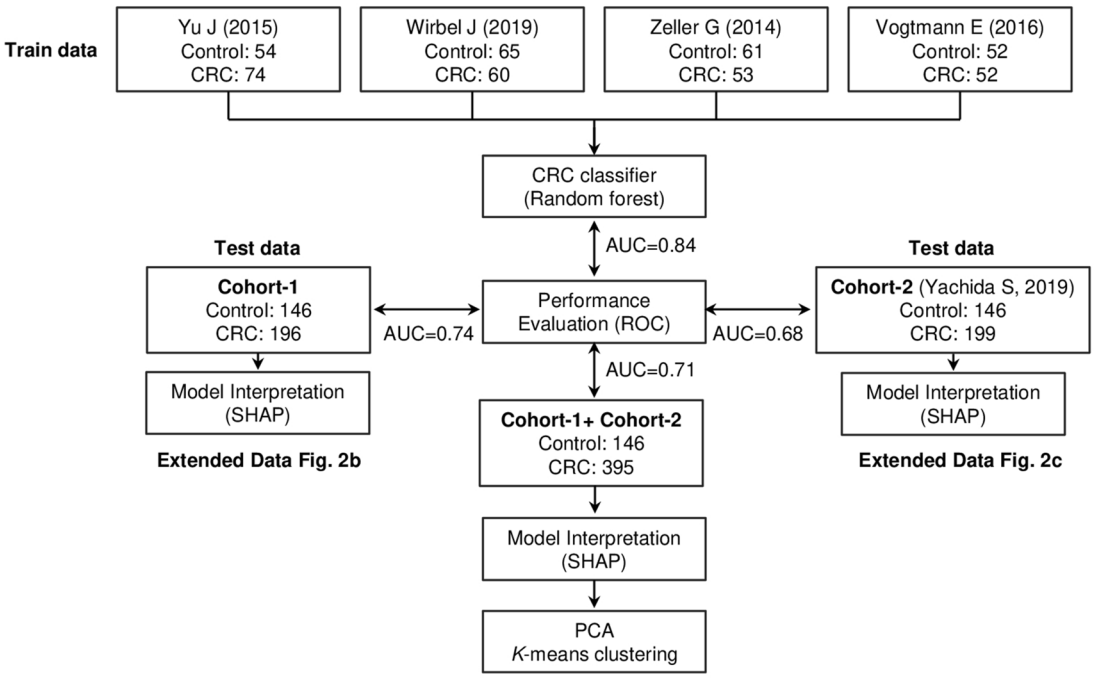
Cohort-1 (CRC only)

	East Japan (NCCH) (n = 138)	West Japan (UOsaka) (n = 117)		Total
WGS	138	62	0	200
WGMS	135	6	55	196
WGS or WGMS	138	117		255

Cohort-2 (CRC and HC; Yachida S et al. 2019)

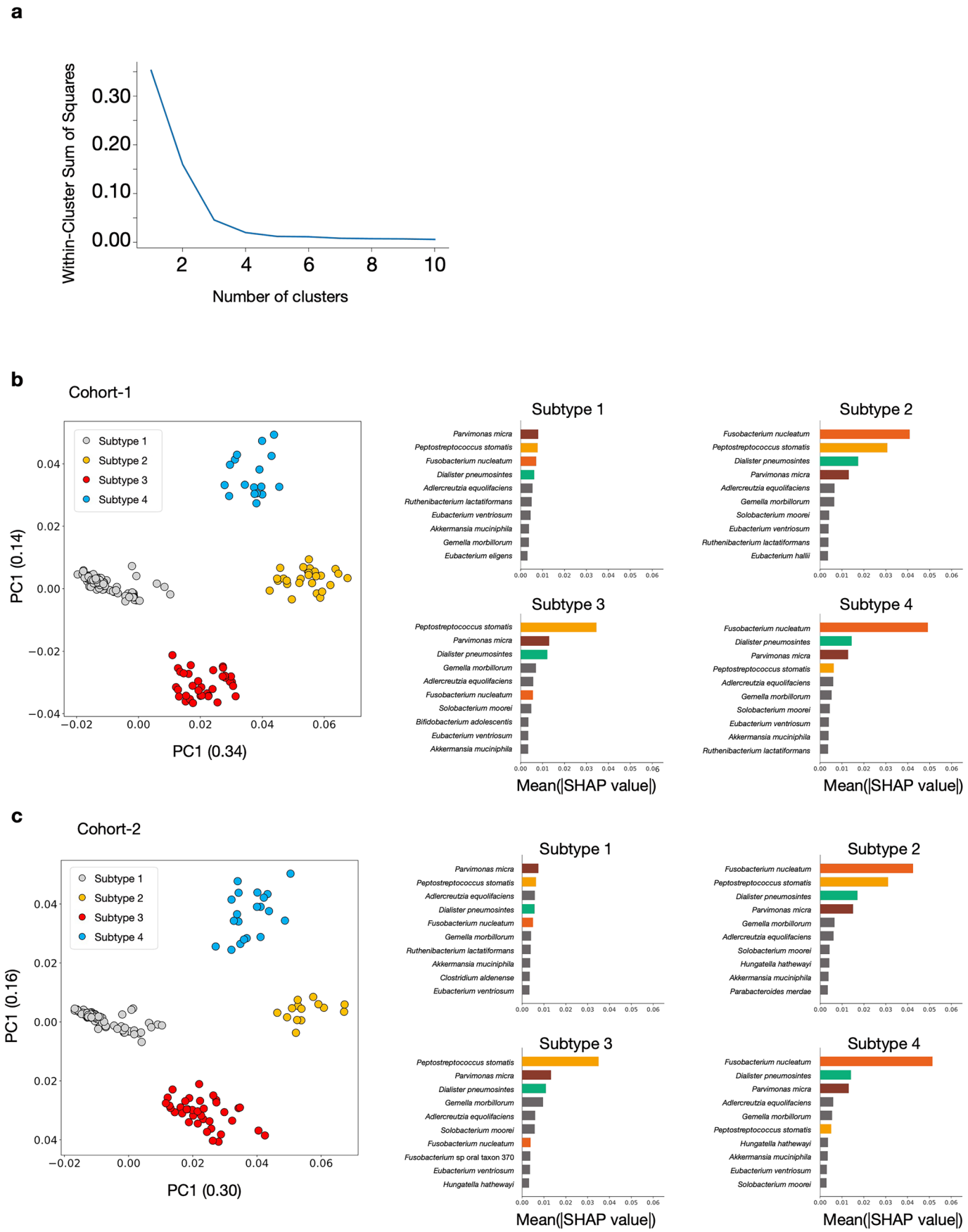
	East Japan (NCCH) (n = 345)		Total
	HC	CRC	
WGS	0	0	0
WGMS	146	199	345

b



Extended Data Fig. 1 | Summary of case numbers analyzed by whole-genome sequencing (WGS), whole-genome metagenomic sequencing (WGMS), and corresponding AI algorithms based on WGMS data. a, Numbers and types of biospecimens collected across the two participating institutions are shown. CRC, colorectal cancer; HC, healthy control; NCCH, National Cancer Center Hospital; UOsaka, The University of Osaka Hospital. **b**, Machine learning analysis was conducted using a random forest classifier trained on four previously reported cohorts. Application of the present cohort (Cohort-1) to the trained model yielded an area under the receiver operating characteristic curve (AUC) of 0.74. Subsequently, SHAP (SHapley Additive exPlanations) values were computed to evaluate feature contributions, and microbial subtypes were defined based on principal component analysis (PCA) followed by *K*-means clustering.

https://doi.org/10.1038/s41588-026-02692-x

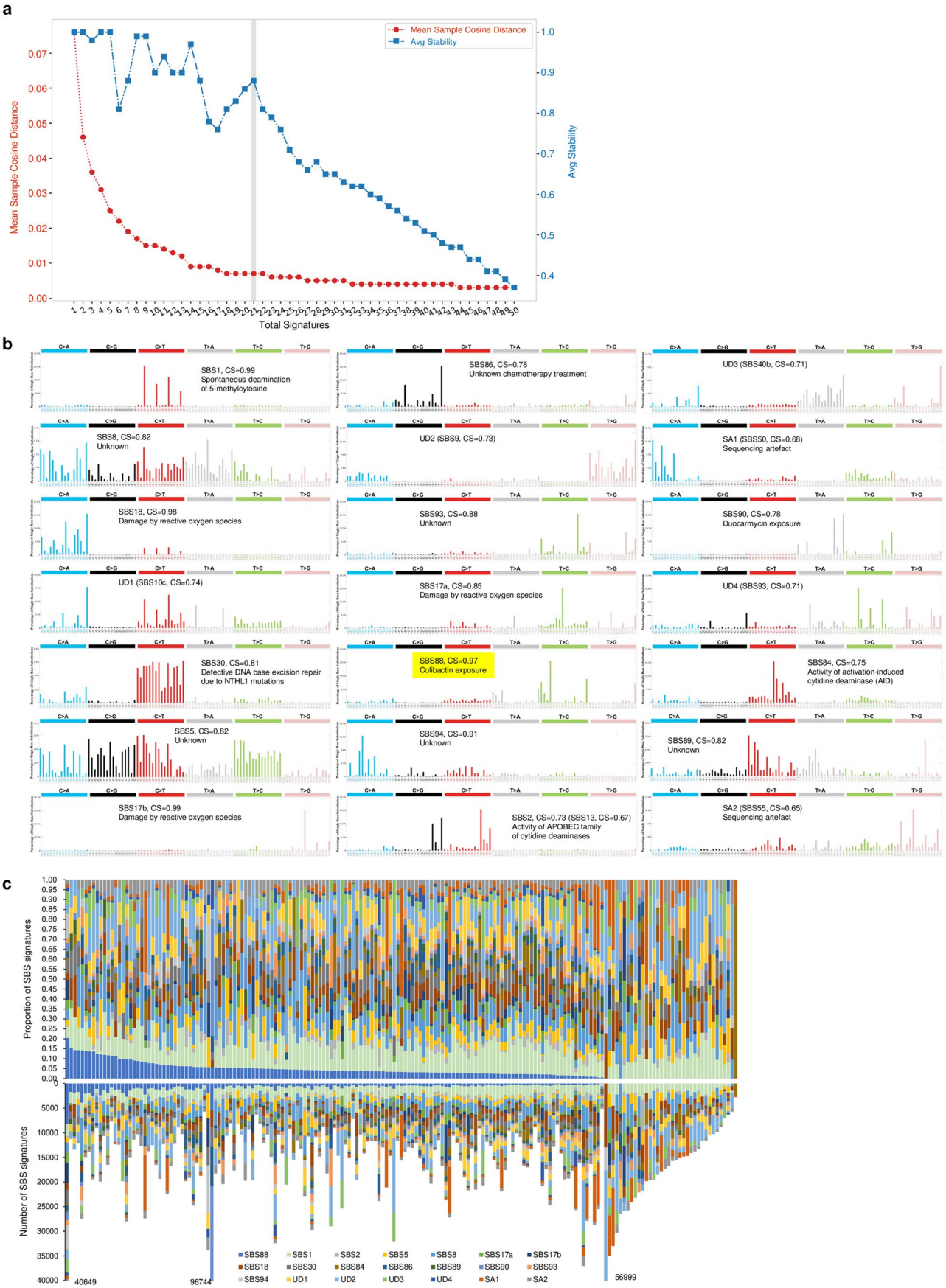


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41588-026-02692-x

Extended Data Fig. 2 | Subclassification of colorectal cancer (CRC) based on the gut microbiome Cohort-1 and Cohort-2. a, Number of clusters for *K*-means clustering determined by the within-cluster sum of squares (WCSS) in the current cohort (Cohort-1). *K*-Means clustering based on WCSS indicates that the gut microbiota of CRC patients can be classified into four distinct subtypes. **b, c**, Subtype classification based on the gut microbiome in independent cohorts. Microbial species contributing most strongly to CRC classification were identified using case and healthy control (HC) samples from Cohort-1 (**b**; CRC $n = 196$, HC $n = 146$) and Cohort-2 (**c**; CRC $n = 199$, HC $n = 146$). SHAP analysis was subsequently applied to evaluate the relative importance of microbial features. Principal component analysis (PCA) demonstrated that, in both cohorts, CRC cases were robustly stratified into four subtypes by *K*-means clustering, recapitulating the subtype architecture observed in Fig. 1b, c. The ten most influential microbial taxa identified for each subtype were largely concordant with those in Fig. 1d. The *x*-axis denotes the mean absolute SHAP value. When analyzed separately by cohort, the results reproduced the same subtype structure and feature importance patterns as in Fig. 1, underscoring the robustness and reproducibility of the findings.

https://doi.org/10.1038/s41588-026-02692-x

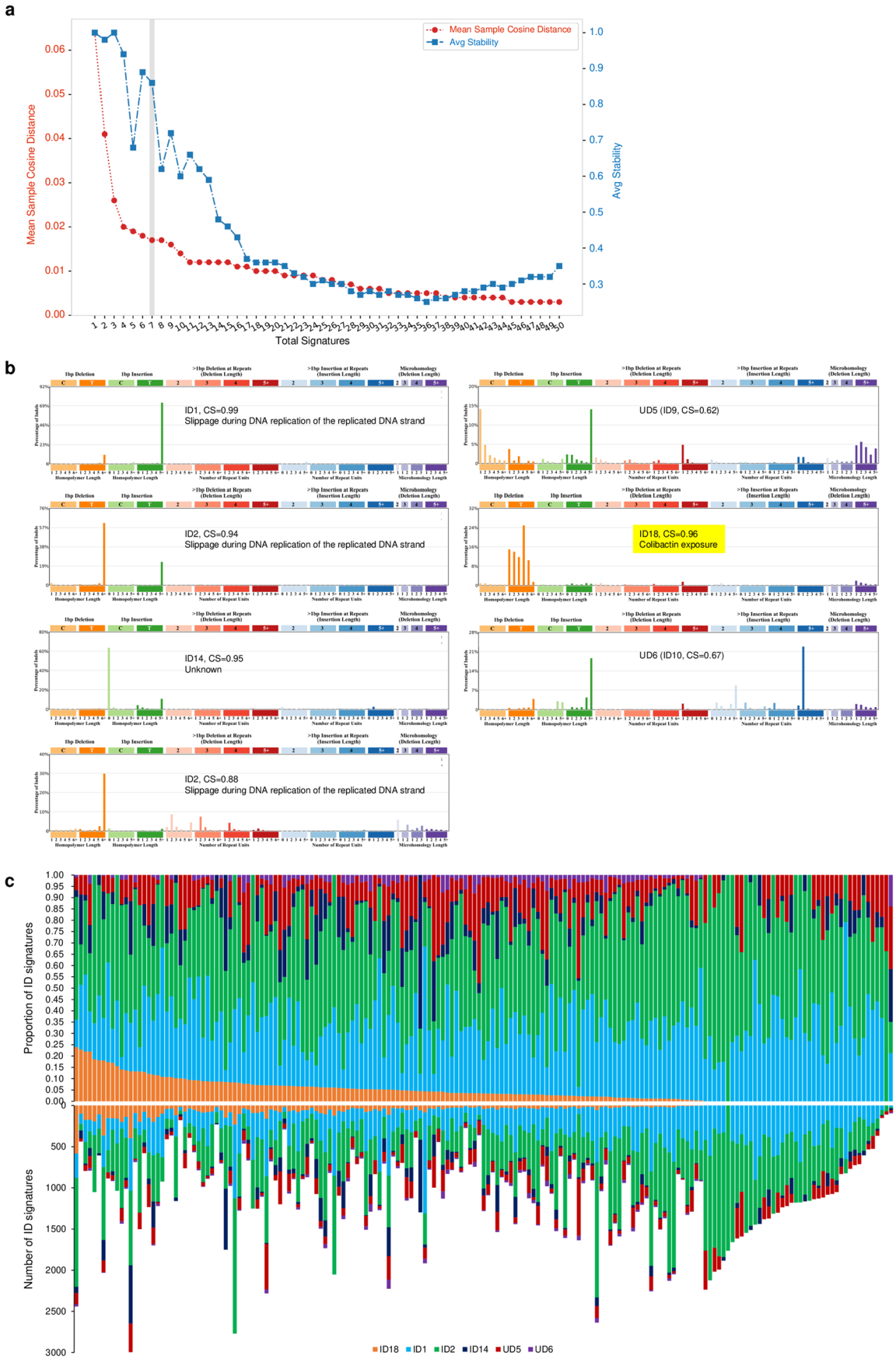


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41588-026-02692-x

Extended Data Fig. 3 | Decomposition of single base substitution (SBS) mutational signatures in 183 non-hypermutated CRCs from Japanese cohorts. **a**, Non-negative matrix factorization (NMF) with automated rank selection was applied to decompose SBS mutational signatures. The grey bar denotes the optimal solution, representing the best trade-off between average stability and mean sample cosine distance. **b**, Cosine similarity (CS) between each decomposed signature and COSMIC reference SBS signatures. Signatures with CS of ≥ 0.75 were assigned to the corresponding COSMIC SBS signatures, whereas those with CS of < 0.75 were interpreted as novel or undecomposed composite signatures and categorized as undetermined (UD). SA, sequence artifact signature. **c**, Distribution of SBS signatures among non-hypermutated CRCs. The upper panel shows the relative contributions of individual SBS signatures, while the lower panel indicates the total number of SBS mutations per case. Cases are ordered from left to right according to the relative contribution of SBS88.

https://doi.org/10.1038/s41588-026-02692-x

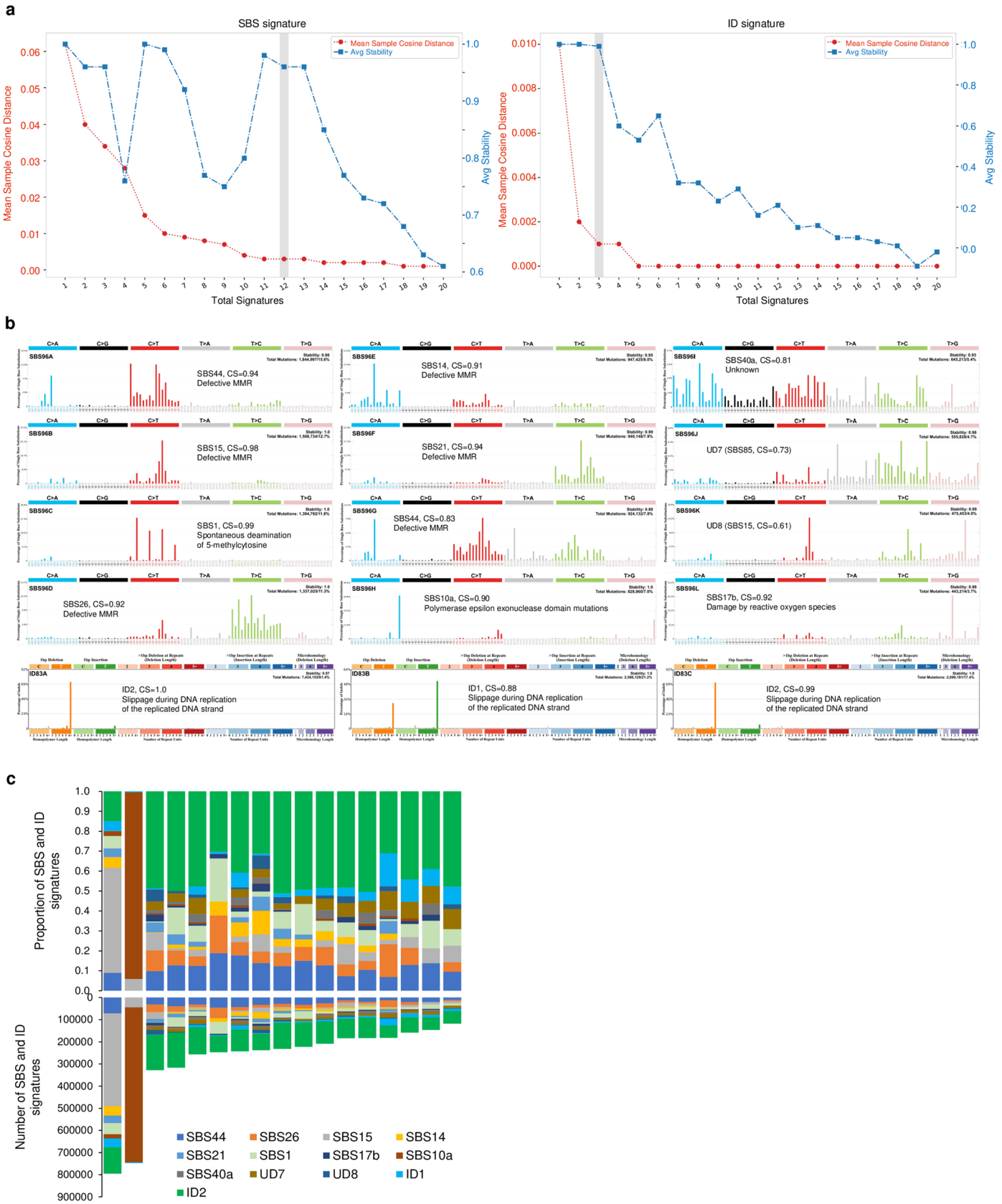


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41588-026-02692-x

Extended Data Fig. 4 | Identification of short insertion and deletion (ID) mutational signatures in 183 non-hypermutated CRCs from Japanese cohorts. **a**, NMF with automated rank selection was applied to identify ID mutational signatures. The grey bar indicates the optimal solution, representing the best trade-off between average stability and mean sample cosine distance. **b**, CS between each decomposed signature and COSMIC reference ID signatures. Signatures with CS of ≥ 0.75 were assigned to their corresponding COSMIC ID signatures, whereas those with CS of < 0.75 were interpreted as novel or undecomposed composite signatures and categorized as undetermined signatures (UD). During decomposition, ID2 was separated into two sub-signatures, which were subsequently merged for downstream analyses. **c**, Distribution of ID signatures among non-hypermutated CRCs. The upper panel depicts the relative contributions of COSMIC ID signatures across individual patients, while the lower panel shows the number of ID signatures detected per case. Patients are ordered from left to right according to the relative contribution of ID18.

https://doi.org/10.1038/s41588-026-02692-x

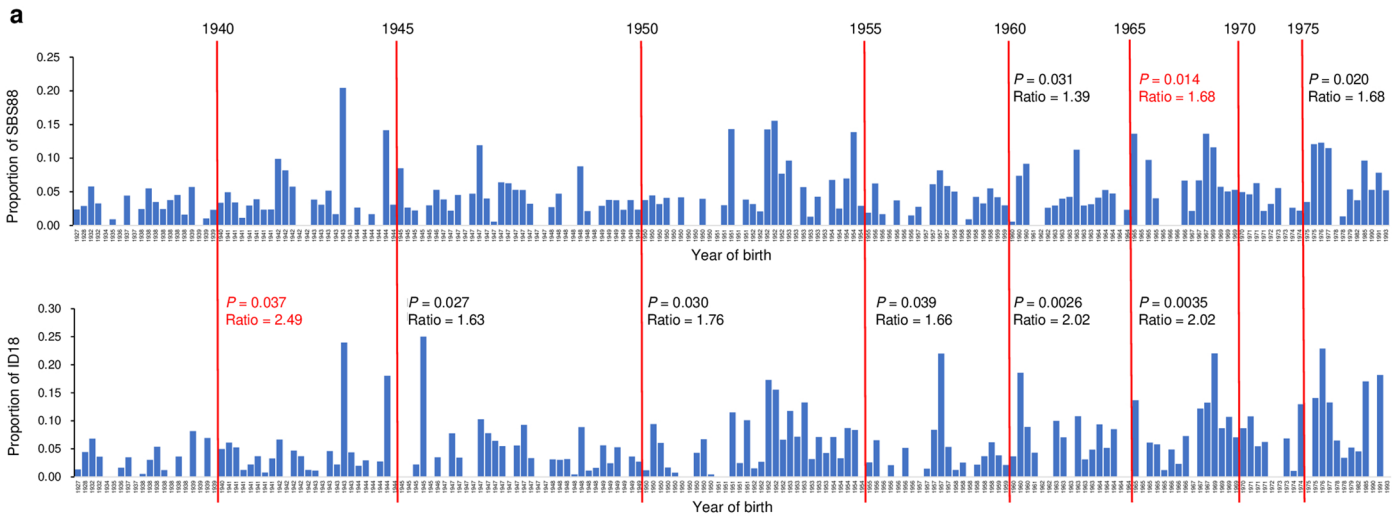


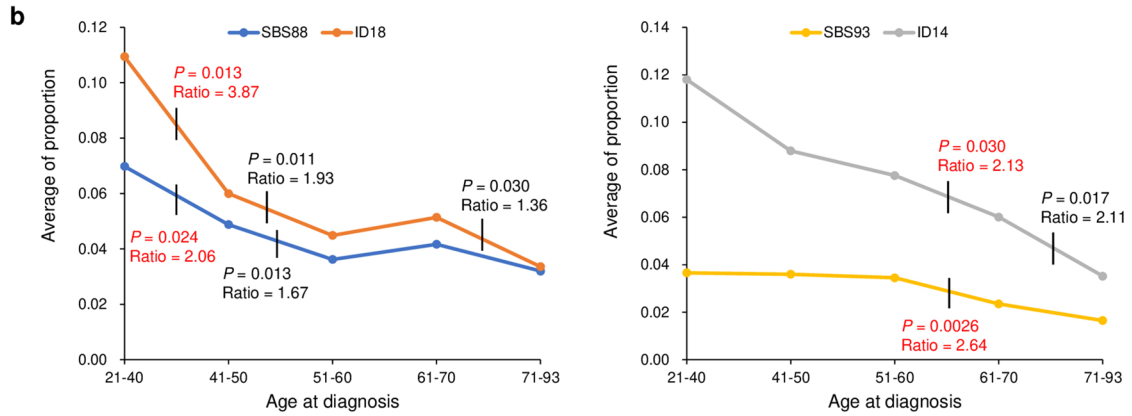
Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41588-026-02692-x

Extended Data Fig. 5 | Identification of SBS and ID mutational signatures in 17 hypermutated CRCs from Japanese cohorts. a, NMF with automated rank selection was applied to decompose SBS and ID mutational signatures. The grey bar indicates the optimal solution, representing the best trade-off between average stability and mean sample cosine distance. **b**, CS between each decomposed signature and COSMIC reference signatures. Signatures with CS of ≥ 0.75 were assigned to the corresponding COSMIC signatures, whereas those with CS of < 0.75 were interpreted as novel or undecomposed composite signatures and categorized as undetermined signatures (UD). During decomposition, ID2 and SBS44 were each partitioned into two sub-signatures, which were subsequently merged for downstream analyses. **c**. Distribution of SBS and ID signatures among hypermutated CRCs. The upper panel depicts the relative contributions of SBS and ID signatures in individual patients, whereas the lower panel shows the number of SBS and ID signatures identified per case. Patients are ordered from left to right according to the total number of SBS and ID signatures detected.

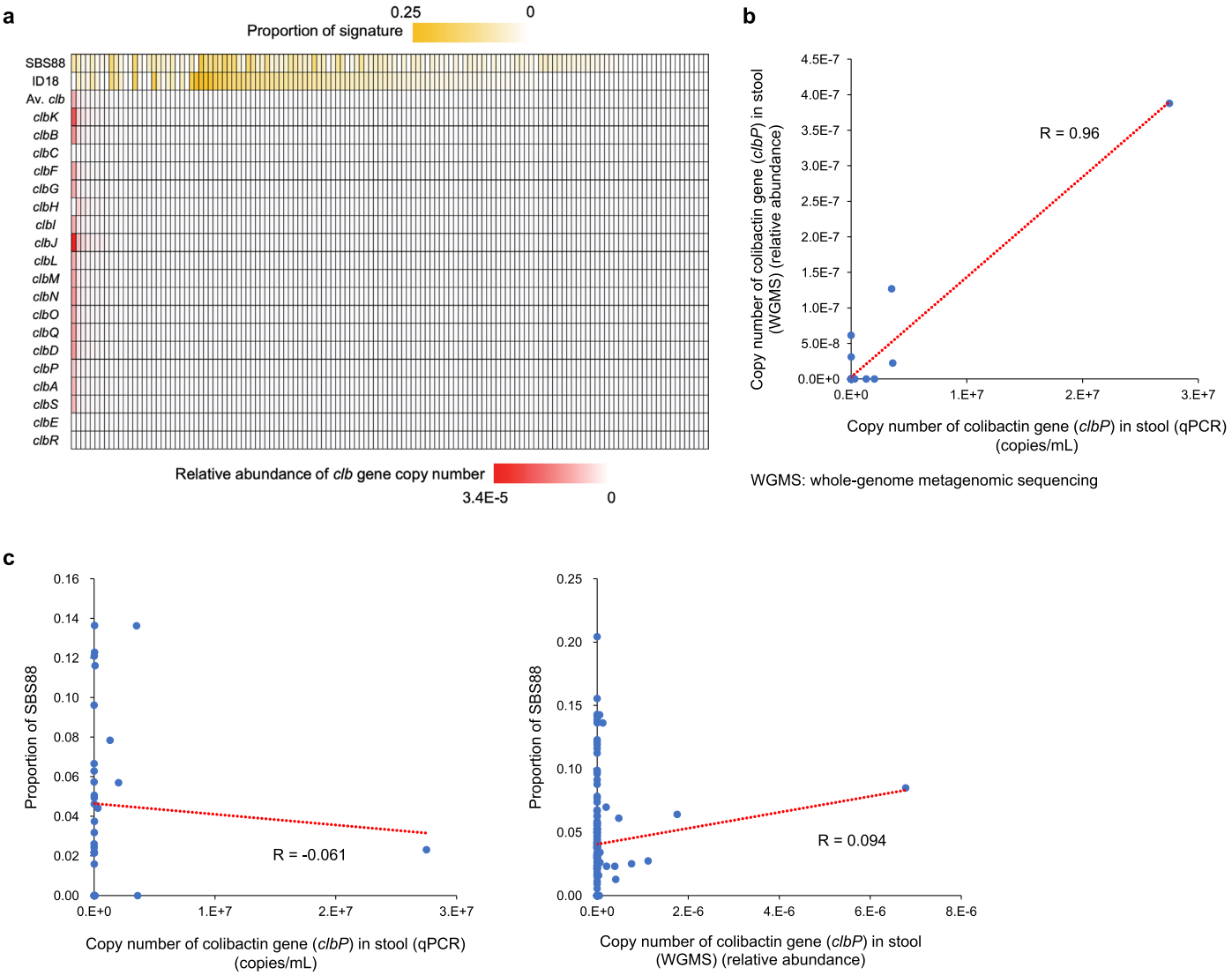
https://doi.org/10.1038/s41588-026-02692-x





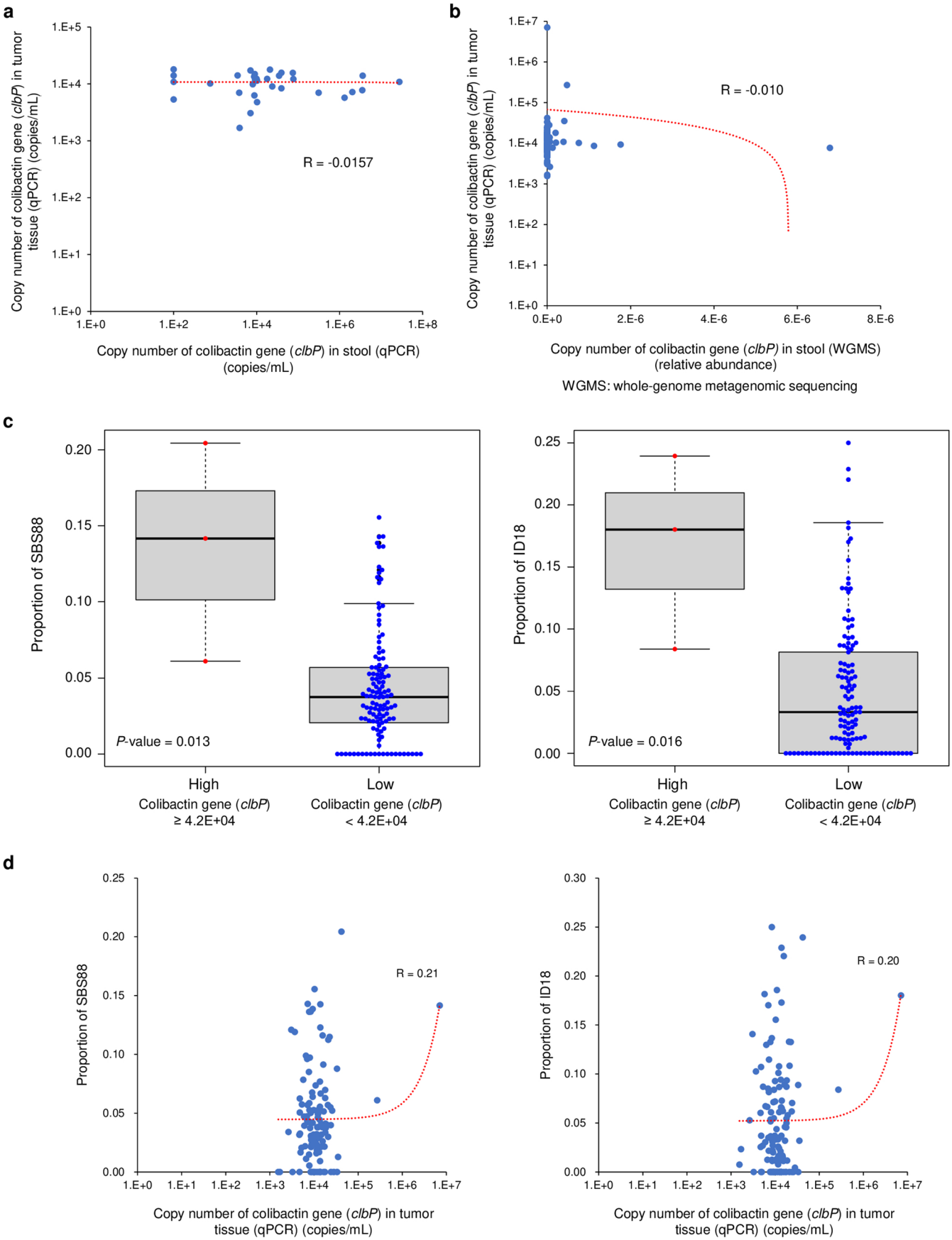
Extended Data Fig. 6 | Birth cohort-associated enrichment of colibactin-related mutational signatures in colorectal cancer (CRC). **a**, Relationship between colibactin-associated mutational signatures and patient birth year. The relative contributions of SBS88 and ID18 were analyzed across CRC cases, ordered from left to right by year of birth. *P*-values were determined using the two-sided Wilcoxon rank-sum test. '*Ratio*' denotes the fold change (ratio of medians) between the two groups separated by the year indicated by a solid red line. For group comparisons stratified by birth year, the highest *Ratio* and corresponding *P*-value are highlighted in *red*. The increase in SBS88 contribution was most prominent among patients born after 1965 compared with those born before 1965, whereas the ID18 contribution peaked in patients born after 1940 relative to those born before 1940, with a secondary elevation observed among those born after 1965. **b**, Relationship between co-occurring mutational signatures and age at diagnosis. '*Ratio*' denotes the median fold change between the two groups separated by a solid line. For comparisons stratified by age at diagnosis, the *Ratio* and corresponding *P*-value for the comparison showing the greatest *Ratio* are highlighted in *red*. *P*-values were calculated using the two-sided Wilcoxon rank-sum test. Significant co-occurrence was observed between the colibactin-associated signatures (SBS88 and ID18) and between SBS93 and ID14. Although all four signatures tended to be more frequent in early-onset CRC, SBS88 and ID18 were most strongly enriched in patients aged of ≤ 40 years, whereas SBS93 and ID14 were most frequent among those aged of ≤ 60 years. The numbers of patients in the age-at-diagnosis groups of ≤ 40, 41–50, 51–60, 61–70, and ≥ 71 years were 10, 24, 36, 60, and 53, respectively.

https://doi.org/10.1038/s41588-026-02692-x



Extended Data Fig. 7 | Correlation between the copy number of the colibactin gene cluster in stool and colibactin-associated mutational signatures in CRC. **a**, Relationship between the colibactin-associated mutational signatures SBS88 and ID18 detected in tumor tissue and the *clb* (colibactin) gene cluster identified by whole-genome metagenomic sequencing (WGMS) of stool samples from each CRC patient. The top two rows (*yellow*) indicate the proportions of SBS88 and ID18 in tumor samples, the third row (*red*) represents the average relative abundance of *clb* gene copy numbers and the subsequent rows show the relative abundances of individual *clb* genes. The detection frequency of *clb* genes in stool was determined by WGMS. Patients are arranged from left to right according to total *clb* gene cluster abundance. **b**, Correlation between *clbP* gene copy numbers determined by WGMS and those quantified by qPCR in fecal samples. A strong correlation ($R = 0.96$) between WGMS- and qPCR-derived copy numbers supports the robustness and reliability of the analytical approach. **c**, Relationship between *clbP* gene copy number and the proportion of the SBS88 signature. The left panel shows the correlation between *clbP* copy number measured by qPCR and SBS88 proportion ($R = -0.061$), whereas the right panel shows the correlation between *clbP* copy number measured by WGMS and SBS88 proportion ($R = 0.094$).

https://doi.org/10.1038/s41588-026-02692-x

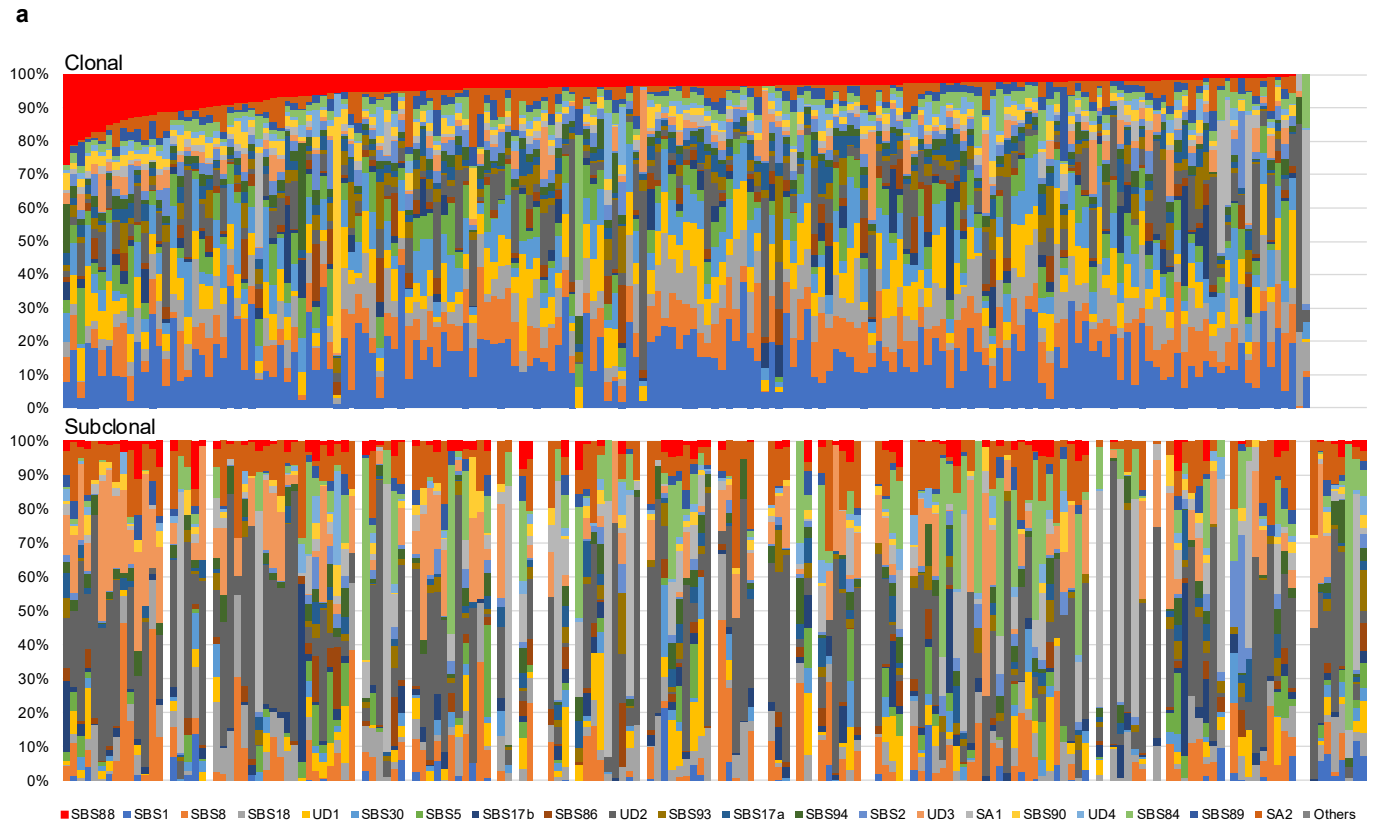


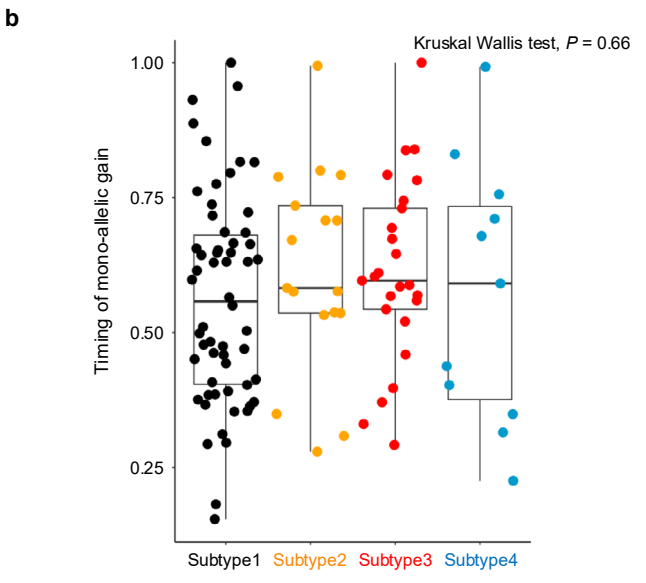
Extended Data Fig. 8 | See next page for caption.

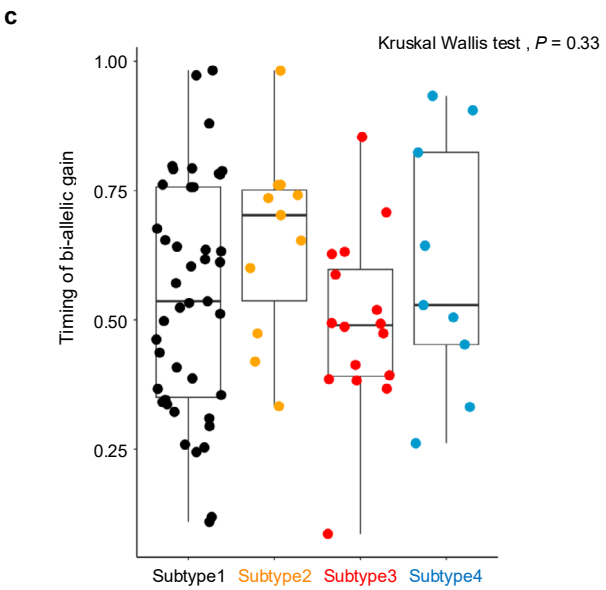
https://doi.org/10.1038/s41588-026-02692-x

Extended Data Fig. 8 | Correlation between copy number of the colibactin gene (*clbP*) in tumor tissue and the proportion of colibactin-associated mutational signatures (SBS88 and ID18). a, Relationship between *clbP* gene copy numbers measured in tumor tissue and fecal samples by qPCR. The correlation coefficient between tissue and fecal *clbP* copy numbers was −0.0157. **b**, Relationship between *clbP* gene copy numbers measured in tumor tissue by qPCR and in fecal samples by WGMS. The correlation coefficient between WGMS-derived fecal and qPCR-derived tissue *clbP* copy numbers was −0.010, indicating no significant correlation between tissue and stool measurements obtained by either qPCR or WGMS. **c**, Relationship between tissue *clbP* copy number and colibactin-associated mutational signatures. Cases with a high *clbP* copy number in tumor tissue ($> 4.2 \times 10^4$) exhibited higher proportions of SBS88 (left panel) and ID18 (right panel) compared with those with a low *clbP* copy number ($P = 0.013$ and $P = 0.016$, respectively; two-sided Wilcoxon rank-sum test). The numbers of patients with high and low *clbP* copy numbers were 3 and 122, respectively. The boxes represent the 25th–75th percentiles, the horizontal lines indicate the medians, and the whiskers extend to the most extreme values within 1.5× the interquartile range. **d**, Correlation between tissue *clbP* copy number (qPCR) and the proportions of SBS88 (left panel) and ID18 (right panel). The correlation coefficients were 0.21 and 0.20, respectively, indicating no statistically significant association.

https://doi.org/10.1038/s41588-026-02692-x

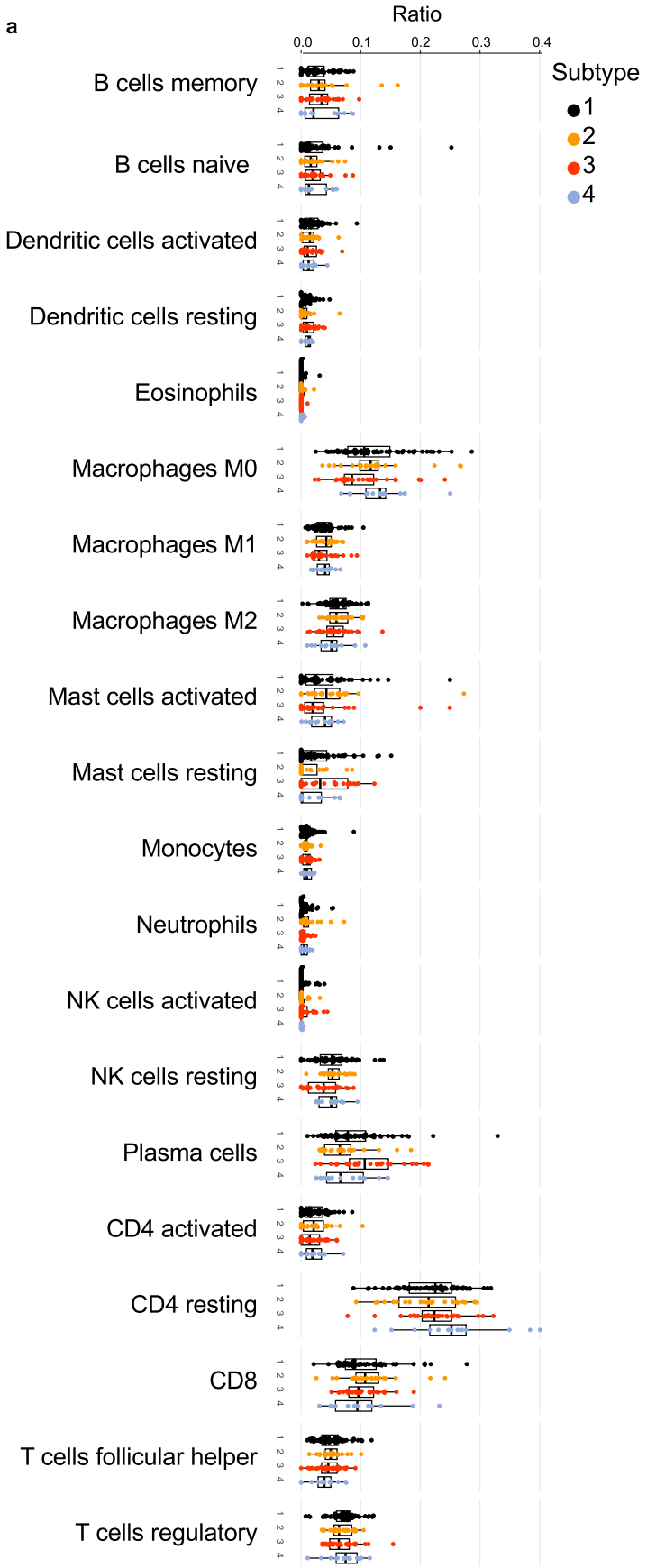


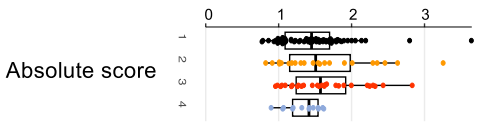


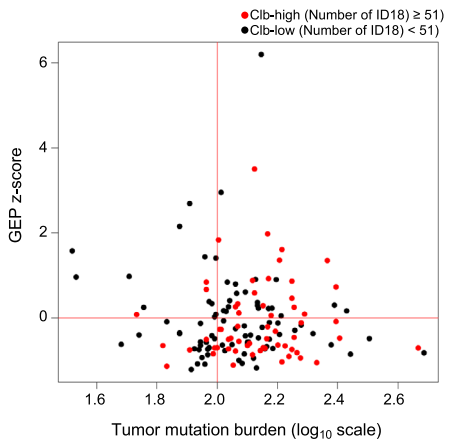


Extended Data Fig. 9 | Correlation between mutational signatures, clonal evolution, and copy-number alteration timing across microbiome-defined CRC subtypes. a, Proportions of SBS mutational signatures calculated separately for clonal and subclonal mutations in each case. Both clonal and subclonal mutations were analyzed to determine the distribution of SBS signatures. Patients with a high proportion of SBS88 are displayed in the top row, arranged from left to right in descending order of SBS88 contribution. The cases in the bottom row correspond to the same patients, shown in an identical left-to-right order for comparison. Patients with fewer than 50 single-nucleotide variants (SNVs) in either phase were excluded from the analysis. **b**, **c**, Comparison of copy-number gain timing across CRC cases based on microbiome-defined subtypes. The correlation between gut microbiome–based CRC subtypes and the sequential acquisition of three classes of copy-number gains was examined: (1) copy number–neutral loss of heterozygosity (CNN-LOH) or loss followed by gain (N:0); (2) monoallelic gains (N:1); and (3) biallelic gains (N:M). No significant differences were observed among microbiome-defined subtypes for either monoallelic (N:1) (**b**) or biallelic (N:M) gains (**c**). Box plots: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× the interquartile range; points, individual cases.

https://doi.org/10.1038/s41588-026-02692-x







In the analysis of the colibactin-associated indel signature ID18, cases with high ID18 levels (above the median value of 51) were not significantly enriched in the TMB^{high}/GEP^{low} subgroup (P = 0.20, OR = 1.42 [95%CI: 0.76 – Inf]).

Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s41588-026-02692-x

Extended Data Fig. 10 | Correlation between gut microbiome–defined CRC subtypes, immune cell composition and ID18-associated mutational and immunogenomic features. a, Immune cell composition across microbiome-defined CRC subtypes. The immune landscape of tumor tissues was inferred from bulk transcriptome data using CIBERSORTx. The boxes represent the 25th–75th percentiles, the horizontal lines indicate the medians, and the whiskers extend to the maximum and minimum values within 1.5× the interquartile range. **b**, Absolute immune score across microbiome-defined CRC subtypes. No significant differences were observed in absolute immune infiltration among subtypes. Boxes, 25th–75th percentiles; center lines, medians; whiskers, 1.5× interquartile range. **c**, Correlation between the ID18 mutational signature, tumor mutational burden (TMB), and T-cell–associated gene expression (GEP) signatures. Non-hypermutated CRC cases were stratified into ID18-high and ID18-low groups using the mean ID18 mutation count (51) as the threshold. The ID18 signature was not significantly enriched in tumors classified as TMB^{high}/GEP^{low} ($P = 0.20$). *P* values were calculated using one-sided Fisher's exact tests. *q* values were obtained by adjusting for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) method.

nature portfolio

Corresponding author(s): Tatsuhiro Shibata

Last updated by author(s): Jun 23, 2026

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No commercial software was used for data collection. All software tools employed in this study are explicitly described within the Main text or the Supplementary Information.

Data analysis	STAR (version 2.4.2a): Ultrafast universal RNA-seq aligner. deconstructSigs (version 1.9.0): Quantifies the contribution of individual mutational signatures to each tumour genome. dndscv (version 0.0.1.0): Detects positively selected driver genes by evaluating recurrent somatic mutations. Kallisto (version 0.43.1): Pseudoalignment-based quantification of transcript abundances from bulk and single-cell RNA-seq data. CallallSV (version 1.0.0): Identifies somatic structural variants from whole-genome paired-end sequencing data. CIBERSORTx (version 1.0): Estimates the relative abundances of constituent cell types within bulk gene-expression profiles. HTSeq (version 0.12.4): Counts RNA-seq reads for gene-expression quantification. kneaddata (version 0.12.0): Removes human-derived reads from rRNA-depleted total RNA-seq data. SigProfilerExtractor (version 1.1.25): Performs de novo mutational-signature extraction, determining the number, activities, and probabilistic contributions of operative signatures. Bowtie2 (version 2.2.9): Aligns metagenomic reads to reference sequences, including bacteriophage phiX and the human genome. cutadapt (version 1.9.1): Trims adapter and primer sequences from metagenomic reads. MetaPhlAn3 (version 3.0.7): Profiles microbial community composition from metagenomic reads using clade-specific marker genes. curatedMetagenomicData (R package): Provides standardized MetaPhlAn-derived taxonomic profiles from public CRC datasets. scikit-learn (Python library): Implements machine-learning models, including random forest classifiers and cross-validation–based performance assessment. shapmat (Release 1): Computes SHAP values, performs principal component analysis and k-means clustering, and generates associated visualizations

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

Data

Policy information about availability of data

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

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

All data supporting the findings of this study are available within the Main text, the Supplementary Information, or from the designated public repositories. Whole-genome sequencing (n = 200) and RNA-seq data have been deposited in the National Bioscience Database Center (NBDC), Tokyo, Japan, under the accession numbers JGAS000696 and JGAD000829. Whole-genome metagenomic sequencing data (n = 541) are available from the NBDC under the accession numbers PRJDB18241, PRJDB18242, and PRJDB38070.
Because these datasets contain personally identifiable genomic information, access is controlled in accordance with Japan's Act on the Protection of Personal Information. Researchers seeking to use these data must obtain approval from the NBDC Human Data Review Board and comply with the prescribed data use application procedures (see https://humandbs.biosciencedbc.jp/en/data-use).

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	The sample size was determined based on our previously published findings derived from stool metagenomic analyses, using overall survival as the primary endpoint. For detailed methodology and statistical justification, please refer to the referenced publication.
Data exclusions	Samples not passing quality-control have been excluded. Details of the exclusion criteria are reported in the paper.
Replication	The estimated copy number of the colibactin gene (clbP) in stool samples demonstrated a high degree of correlation between quantitative PCR (qPCR)–based estimation and whole-genome metagenomic sequencing (WGMS), validating the reproducibility of the measurement across independent analytical methods.
Randomization	Not applicable for this observational study
Blinding	Not applicable for this observational 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

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

Methods

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

Antibodies

Antibodies used	Mouse monoclonal anti-FOXP3 antibody (clone 236A/E7; dilution 1:100; ab20034, Abcam)
Validation	Validation was performed using formalin-fixed, paraffin-embedded (FFPE) human tonsil tissue as a positive control.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Colorectal cancer (CRC) cases included 200 patients who underwent whole-genome sequencing (WGS) of tumor tissues and 196 patients who underwent whole-genome metagenomic sequencing (WGMS) of stool samples.
Recruitment	Patients were recruited, with informed consent, from the National Cancer Center Hospital (Tokyo, Japan) and the University of Osaka Hospital (Osaka, Japan).
Ethics oversight	National Cancer Center(Tokyo, Japan), The University of Tokyo (Tokyo, Japan), the University of Osaka (Osaka, Japan), and Institute of Science Tokyo (Tokyo, Japan)

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	NA
Study protocol	NA
Data collection	NA
Outcomes	NA