nature genetics

Supplementary information

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

In the format provided by the authors and unedited

SUPPLEMENTARY INFORMATION

MAIN

We examined the relationship between microbial subtypes and structural variants (SVs). Intrachromosomal SVs were annotated according to size, whereas interchromosomal SVs were categorized as LINE-1–mediated or non–LINE-1 transpositions using RepeatMasker (Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*; http://www.repeatmasker.org). No significant subtype-dependent differences in SV profiles were observed (Supplementary Table 24), although a trend toward fewer LINE-1–mediated transpositions was noted in Subtype 1(Supplementary Fig. 3).

DISCUSSION

Through interpretable artificial intelligence (AI)–based analysis of the gut microbiome in colorectal cancer (CRC), we delineated four distinct microbiome-associated subtypes. The clinicopathological, genetic, and immunological characteristics of each subtype are summarized in Table 1. Although this classification deviates from the conventional consensus molecular subtype (CMS) framework, it has potential utility in predicting genetic alterations and patient prognosis[1].

The present immunogenomic analyses further highlight the interplay between the gut microbiota and the tumor immune microenvironment. Patients classified as Subtype 3 were more frequently represented in categories other than the immune-hot phenotype (TMB^{high}/GEP^{high}). Moreover, tumors exhibiting a high colibactin-associated mutational signature demonstrated decreased infiltration of immune effector cells and an increased proportion of regulatory T cells. Consistent with these observations, López et al. reported

that colibactin-producing *Escherichia coli* reduced tumor-infiltrating T cells, including cytotoxic CD8+ T cells, and attenuated the response to anti–PD-1 immunotherapy in a murine model[2]. Although the precise molecular mechanisms remain to be elucidated, these findings collectively suggest that colibactin exposure may suppress anti-tumor immunity, thereby facilitating tumor development.

Several limitations of this study warrant consideration. First, the relatively small sample size may have introduced bias and limited statistical power, particularly in subgroup analyses. Second, because fecal samples were collected at a single preoperative time point from patients already diagnosed with CRC, potential influences of preoperative clinical conditions cannot be excluded. Third, stool samples were collected from the first bowel movement after hospital arrival for colonoscopy, and bowel preparation had already been initiated in a subset of the participants. Although previous studies have suggested that bowel-cleansing agents may influence gut microbial composition[3], the samples in this study predominantly consisted of solid stool. Our prior investigation demonstrated that such specimens exhibit extremely high concordance with samples collected the day before colonoscopy, prior to bowel preparation[4]. Moreover, the microbial profiles generated using this standardized collection protocol have been independently validated across multiple large-scale, multi-continental CRC cohort studies[5,6,7,8], underscoring the robustness and reproducibility of the microbiome data obtained in this study. Finally, a recent study reported that the relative abundance of two of the four key bacterial species identified here (*P. stomatis* and *F. nucleatum*) may vary depending on confounding factors[9]. Thus, further large-scale validation will be required to confirm the robustness and generalizability of these associations.

METHODS

Rationale for Sample Size Setting

In our previously published study[10], we established a subtype classification framework based on gut microbiota composition using four international cohorts comprising 239 CRC patients and 231 healthy controls (HCs)[6,11-13] as the training dataset. This classification rule was subsequently applied and validated in an independent test cohort (Cohort-2; *n* = 199) as reported in another publication[7]. In the present study, we applied the same classification algorithm, developed by Rynazal et al.,[10] to a new test dataset (Cohort-1) comprising CRC cases recruited from the National Cancer Center Hospital and the University of Osaka Hospital. Each case was classified into one of four microbial subtypes, after which we examined bacterial community structures and evaluated reproducibility.

To further validate the robustness of the classification, data from Cohorts-1 and -2 were combined and analyzed jointly. It should be noted that while Cohort -2 was previously used as test data to assess the diagnostic performance of the classifier[10], its prognostic implications for overall survival had not been investigated; this study is the first to evaluate the prognostic relevance of the microbial subtypes. The present study was designed with sufficient statistical power to detect subtypes associated with a significantly poorer prognosis. In the prior analysis[10], Cohort-2 cases were distributed across the four subtypes as follows: Subtype 1, *n* = 128 (~ 60%); Subtype 2, *n* = 14 (~ 10%); Subtype 3, *n* = 37 (~ 20%); and Subtype 4, *n* = 20 (~ 10%). Based on these proportions, power calculations were performed assuming that one subtype exhibits a distinct prognosis.

Under scenarios where the hazard ratio (HR) between one subtype and the others was 2, 2.5, or 3, the detection power varied depending on subtype prevalence and the number of observed events (L). When Subtype 1 (~ 60% of cases) showed a poorer prognosis and the HR = 3, the statistical power exceeded 90% with L = 50 events (scenario S21). Similarly, when Subtype 3 (approximately 20% of cases) exhibited a poorer prognosis with HR of 3, the statistical power exceeded 90% at L = 75 and remained approximately 75% even at L = 50 (scenario S23). For Subtype 2 (or 4), which comprised approximately 10% of cases, the power reached around 80% at L = 100. Under a more moderate effect size (HR = 2.5), the power was lower for Subtypes 2 and 4 (~10% of cases) but remained above 75% for Subtypes 1 and 3 (scenarios S11 and S13) at L = 75.

This study aimed to obtain approximately 400 evaluable CRC cases by integrating ~ 200 new cases from the National Cancer Center Hospital and the University of Osaka Hospital (Cohort-1) with the previously analyzed 199 cases from Cohort-2. The number of events depended on follow-up duration; assuming mortality in ~ 25% of cases (L = 100), the study retained sufficient power to detect HR = 3. Even under a lower mortality rate of ~15% (L = 60), the power remained adequate for detecting prognostic differences in Subtypes 1 and 3.

	HR: Subtype (prevalence)					
Scenario	Subtype 1 (60%)	Subtype 2 (10%)	Subtype 3 (20%)	Subtype 4 (10%)	L	Power
s1	2	1	1	1	50	0.50038

					75	0.69199
					100	0.82394
s2	1	2	1	1	50	0.20532
					75	0.29504
					100	0.38545
s3	1	1	2	1	50	0.34547
					75	0.50038
					100	0.63472
s11	2.5	1	1	1	50	0.76451
					75	0.91912
					100	0.976
s12	1	2.5	1	1	50	0.33998
					75	0.49287
					100	0.62638
s13	1	1	2.5	1	50	0.57001
					75	0.76451
					100	0.88233
s21	3	1	1	1	50	0.90674
					75	0.98397
					100	0.99777
s22	1	3	1	1	50	0.47462
					75	0.66297
					100	0.79839

s23	1	1	3	1	50	0.74472
					75	0.90674
					100	0.97037

Structural Variant Calling

CallallSV (v1.0.0)[14] was used with default parameters to detect four classes of structural variation (SV): deletions, tandem duplications, inversions, and translocations. The method requires paired-end (PE) whole-genome sequencing reads from matched tumor and non-tumor samples. PE reads of 150 bp in length, with fragment sizes ranging from 450–700 bp, were generated from 138 CRC samples together with their corresponding normal tissues. Equivalent sequencing data were also obtained for ten tumor samples (CM009T, CM010T, CM011T, CM012T, CM013T, CM014T, CM015T, CM016T, CM017T, and CM018T) for which matched normal tissues were unavailable. For these ten cases, mock normal datasets were constructed by randomly sampling reads from 128 normal samples and adjusting the coverage to a read depth of 35. All PE reads were aligned to the human reference genome (hg19) using BWA-MEM[15] with the -T 0 option. SV candidates were identified from discordantly mapped PE reads and from soft-clipped single-end reads. Among SV candidates detected across all 138 tumor samples—including the ten cases analyzed with mock normal data—those present in any of the 128 normal samples were removed. Further methodological details are available on GitHub (https://github.com/ma9606/callallSV) and in reference 14.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Fig. 1 | Comparison of the proportion of non-human bacterial reads among microbiome-defined CRC subtypes.

Significant differences were observed in the ratio of non-human bacterial reads across the subtypes ($P = 3.4 \times 10^{-4}$). Subtype 2 exhibited a markedly higher proportion compared with the other subtypes, with pairwise comparisons showing significant differences: Subtype 2 vs. Subtype 1 ($P = 2.7 \times 10^{-5}$, $q = 1.6 \times 10^{-4}$), Subtype 2 vs. Subtype 3 ($P = 0.0013$, $q = 0.0038$), and Subtype 2 vs. Subtype 4 ($P = 0.028$, $q = 0.056$). Box plots: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 × the interquartile range; points, outliers. Statistical significance was determined using two-sided Wilcoxon rank-sum tests, with *P* values adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) method.

Supplementary Fig. 2 | Correlation between microbiome-defined CRC subtypes, tumor fraction, and the extent of copy-number amplification.

a, Tumor purity estimated from WGS data across CRC subtypes. Tumor purity was inferred from variant allele frequencies and copy-number profiles derived from WGS. No significant differences were observed among subtypes. 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. **b,** Number of amplified genes by CRC subtype. Tumors classified as Subtype 1 occasionally exhibited focal regions of high-level copy-number amplification; however, neither the Kruskal–Wallis test nor subsequent pairwise comparisons revealed statistically significant differences in the

overall degree of copy-number alteration among subtypes. Box plots: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 × the interquartile range; points, individual cases.

Supplementary Fig. 3 | Metagenome-based classification of CRC and structural variants (SVs).

a, Number, type, and size distribution of SVs detected in each case. Three classes of intrachromosomal SVs—deletions, tandem duplications, and inversions—were individually quantified, distinguishing small (breakpoint distance ≤ 1 kb) from large (> 1 kb) events. Breakpoint locations of translocation SVs resulting in interchromosomal rearrangements were annotated using RepeatMasker. These translocation SVs were further classified into those mediated by LINE-1 retrotransposition and other non-LINE-1–associated translocations. **b,** Distribution of the number of LINE-1 retrotransposition events per case by CRC subtype, visualized using a beeswarm boxplot. For clarity, an outlier case from Subtype 4 (CM116T) was excluded. Statistical significance was assessed using the two-sided Wilcoxon rank-sum rank-sum test. 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

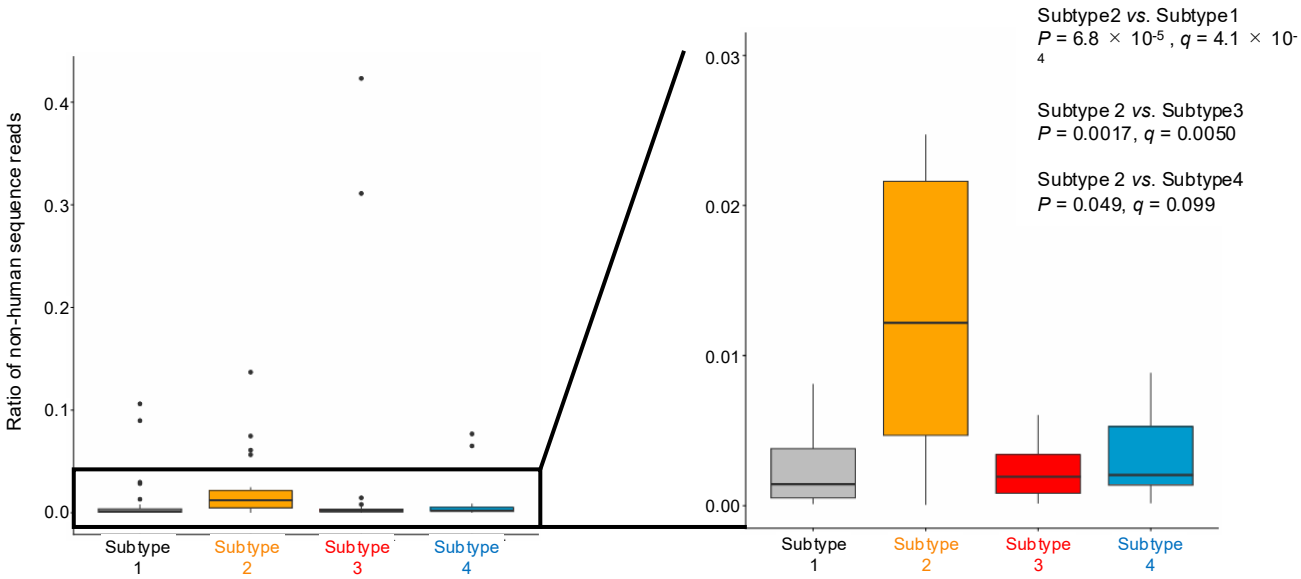
REFERENCES

1. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350-6 (2015).
2. Lopes, A. *et al.* Colibactin-positive Escherichia coli induce a procarcinogenic immune environment leading to immunotherapy resistance in colorectal cancer. *Int J Cancer* **146**, 3147-3159 (2020).
3. Nagata, N. *et al.* Effects of bowel preparation on the human gut microbiome and metabolome. *Sci Rep* **9**, 4042 (2019).
4. Nishimoto, Y. *et al.* High stability of faecal microbiome composition in guanidine thiocyanate solution at room temperature and robustness during colonoscopy. *Gut* **65**, 1574-5 (2016).
5. 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).
6. 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).
7. 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).
8. Piccinno, G. *et al.* Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. *Nat Med* **31**, 2416-2429 (2025).
9. Tito, R.Y. *et al.* Microbiome confounders and quantitative profiling challenge

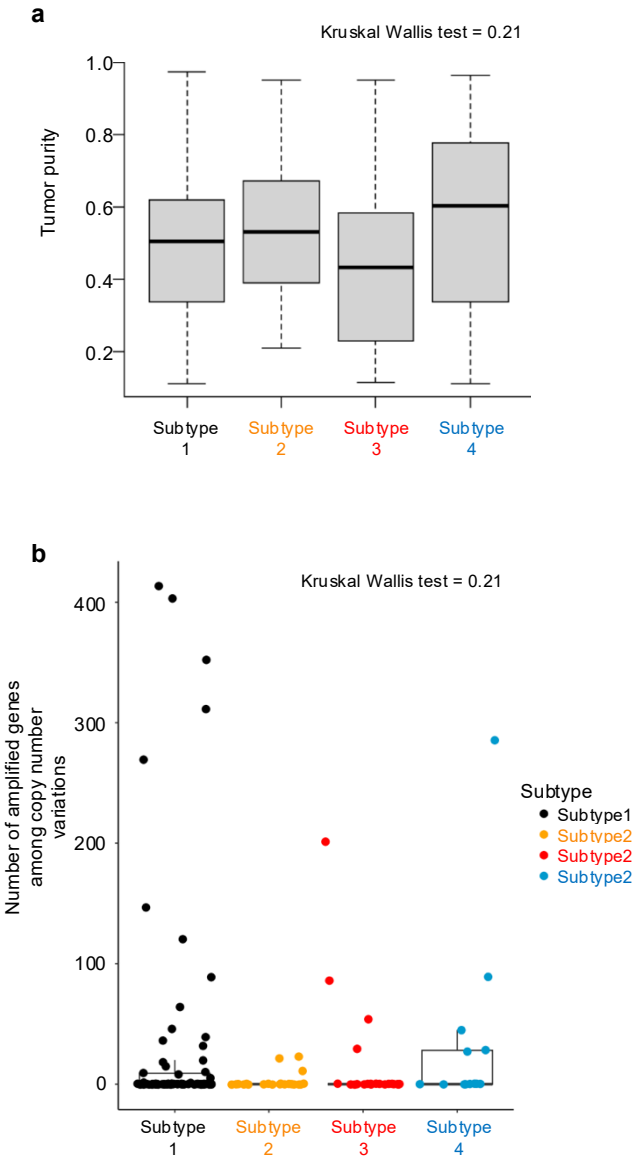
predicted microbial targets in colorectal cancer development. *Nat Med* **30**, 1339-1348 (2024).

10. Rynazal, R. *et al.* Leveraging explainable AI for gut microbiome-based colorectal cancer classification. *Genome Biol* **24**, 21 (2023).
11. 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).
12. Vogtmann, E. *et al.* Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One* **11**, e0155362 (2016).
13. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**, 766 (2014).
14. Saito-Adachi, M. *et al.* Oncogenic structural aberration landscape in gastric cancer genomes. *Nat Commun* **14**, 3688 (2023).
15. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

Supplementary Fig. 1

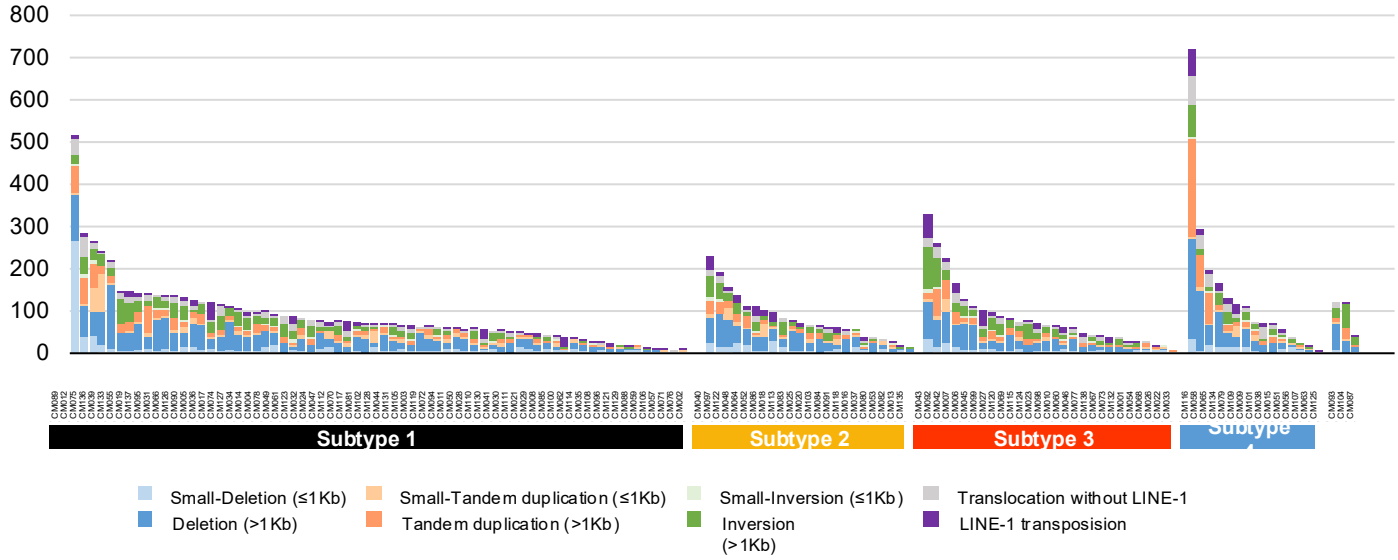


Supplementary Fig. 2



Supplementary Fig. 3

a



b