nature portfolio

Peer Review File

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

Corresponding Author: Professor Tatsuhiro Shibata

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Genetics.

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

13th Feb 2026

Dear Dr Shibata,

Your Article, "Prevalence and chronology of colibactin-associated mutational processes and their microbiome spectra in Japanese colorectal cancer" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised by Reviewers #2 and #3. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a “Response to referees” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary: https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

EXTENDED DATA FIGURES

When re-submitting your manuscript, please ensure that any supplementary figures and tables that are crucial to the manuscript's conclusions are converted into Extended Data figures and tables to increase visibility of these data. Extended Data figures and tables are online-only (present in the online PDF and full-text HTML versions of the paper), peer-reviewed display items that provide essential background to the article but are not included in the main article due to space constraints. A maximum of ten Extended Data display items (figures and tables) is permitted.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within about eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Wei Li, PhD
Senior Editor
Nature Genetics
www.nature.com/ng

Reviewers' Comments:

Reviewer #1 (Remarks to the Author):

I reviewed the new manuscript and response to reviewers. The authors addressed all my comments.

Reviewer #2 (Remarks to the Author):

The authors have addressed all my comments and I would like to congratulate them with an excellent paper. I have one minor comment:
- The y-axis title of Figure 3a (middle and lower panels) now reads: "Proportion of SBS18/ID88 signature". In the original manuscript this was "Proportion of SBS/ID signatures". I think the original titles better represented that what is shown.

Reviewer #3 (Remarks to the Author):

In this manuscript, Shiba and co-authors, investigate the interplay between gut microbiome in CRC and its mutational features, including the presence of the colibactin mutational signature in cohorts of Japanese patients with CRC. They first generate a microbiome-based interpretable AI CRC classifier, classifying CRC into 4 subtypes according to the presence (or absence in the case of Subtype 1) of major distinguishing bacterial species. Stage at diagnosis was more advanced in Subtype 2 and Subtype 3 was associated with a younger age on diagnosis and increased prevalence in those born after the 1980s. Additionally, Subtype 3 was enriched in the colibactin mutational signatures SBS88 and ID 18, which, consistently with other studies were found at higher relative prevalence in CRCs from younger patients at diagnosis. They do also show, however, that these signatures are also increased in those born after 1965/by decade of birth which is novel. They ascribe APC hotspot, TP53 and the BRAF D594G driver mutations to colibactin-induced mutagenesis and also show lack of correlation between colibactin itself (clbP) in stool (which is novel using stool WGMS) and its mutational signature in CRCs,

with more variable results when tumor clbP was assessed. Interestingly, they show that colibactin induced mutations that are early clonal events and that these tumors have an earlier occurrence of CNN-LOH as well as higher Tregs in their TME.

Comments:
-The authors state that the prevalence of SBS88 and/or ID18 was 44.8% in the Japanese cohort and highlight this as significantly higher than previously reported. However, in the seminal PMID 40267983 Nature article, from which the authors also integrated data for signature deconvolution, the prevalence of SBS88 or ID18 in countries with a high ASR of early-onset CRC (which includes Japan) is ~50-60% so actually similar, if not higher than this reported here. Also when SBS88 and ID 18 are examined by decade of diagnosis their frequencies are lower in those <40 (for example) than in the Nature WGS paper.

-E.coli (pks+ strain or any) was not identified or at least shown among the bacteria present in Subtypes 1-4 (Figure 1, Extended Figure Data 2). It would be worth commenting whether it was picked up at lower levels that are not shown by the figures or completely absent.

-In Extended Figure 6c, it is a little hard to distinguish the proportions of the individual signatures based on the current color scheme. If I interpret correctly that SBS 30 is present across cases, it is more likely that this is not due to NTHL1 germline events (these are rare) but rather to alkylating damage present in CRC (PMID: 34140290, PMID: 41416871) which could be worth discussing. The etiology of SBS30 being that of CRC alkylating damage is also supported by its decrease with younger age at diagnosis (should be the opposite for NTHL1 polyposis-related CRCs).

-It would be good to validate the CIBERSORT inferences for resting NK cells, plasma cells and mast cells with some orthogonal methods (multiplex IF, IHC), The authors do that nicely for Tregs. Otherwise, that limitation can be acknowledged.

-It would be worth specifying in the text whether all the specimens analyzed with genomics were the primary tumors versus metastatic lesions. This is relevant for the Stage IV patients. Extended Table 1 lists the primary tumor locations but it is still unclear if the primary was the sequenced specimen for Stage IV patients. In the same Table it is also not clear what “Single” vs “Multiple” Lesions refers to for Stage IV patients. Single primary tumor (vs two primaries) or single metastatic lesion (oligometastatic CRC)?

-In Figure 2c, Stage 0 (carcinoma in situ) cases are included. Typically, since these are not invasive cancers, they can be excluded from CRC analyses.

-In Extended Figure 9, specify how “colibactin sequence motifs” were identified / reference/ method? (assuming reference 13?)

-In my opinion, the authors have addressed the comments of Reviewer #4.

Version 1:

Decision Letter:

Our ref: NG-A71371R

8th Apr 2026

Dear Dr. Shibata,

Thank you for submitting your revised manuscript "Prevalence and chronology of colibactin-associated mutational processes and their microbiome spectra in Japanese colorectal cancer" (NG-A71371R). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Wei Li, PhD
Senior Editor
Nature Genetics

Reviewer #3 (Remarks to the Author):

I thank the authors for addressing my comments and their excellent work.

March 12, 2026

Dr. Wei Li
Senior Editor
Nature Genetics

Dear Dr. Li,

Please find enclosed our revised manuscript entitled **"Prevalence and chronology of colibactin-associated mutational processes and their microbiome spectra in Japanese colorectal cancer."**
We have fully addressed the referees' comments and have revised and expanded the manuscript accordingly. All requested analyses and revisions have been incorporated, and we have clarified the points highlighted by the reviewers. We have also included a manuscript file with all modifications and additions indicated in red. A detailed, point-by-point response to each comment is provided below.

Referees' comments:

Reviewer #1 (Remarks to the Author):
I reviewed the new manuscript and response to reviewers. The authors addressed all my comments.
Response: We thank Reviewer #1 for the positive evaluation of our revised manuscript. We greatly appreciate the reviewer's thoughtful comments and careful review in the previous round.

Reviewer #2 (Remarks to the Author):
The authors have addressed all my comments and I would like to congratulate them with an excellent paper. I have one minor comment:
- The y-axis title of Figure 3a (middle and lower panels) now reads: "Proportion of SBS88/ID18 signature". In the original manuscript this was "Proportion of SBS/ID signatures". I think the original titles better represented that what is shown.
Response:

We are deeply grateful to the reviewer for the generous assessment of our work and for noticing this point. We have revised the y-axis title in Figure 3a (middle and lower panels) to “Proportion of SBS/ID signatures”.

Reviewer #3 (Remarks to the Author):

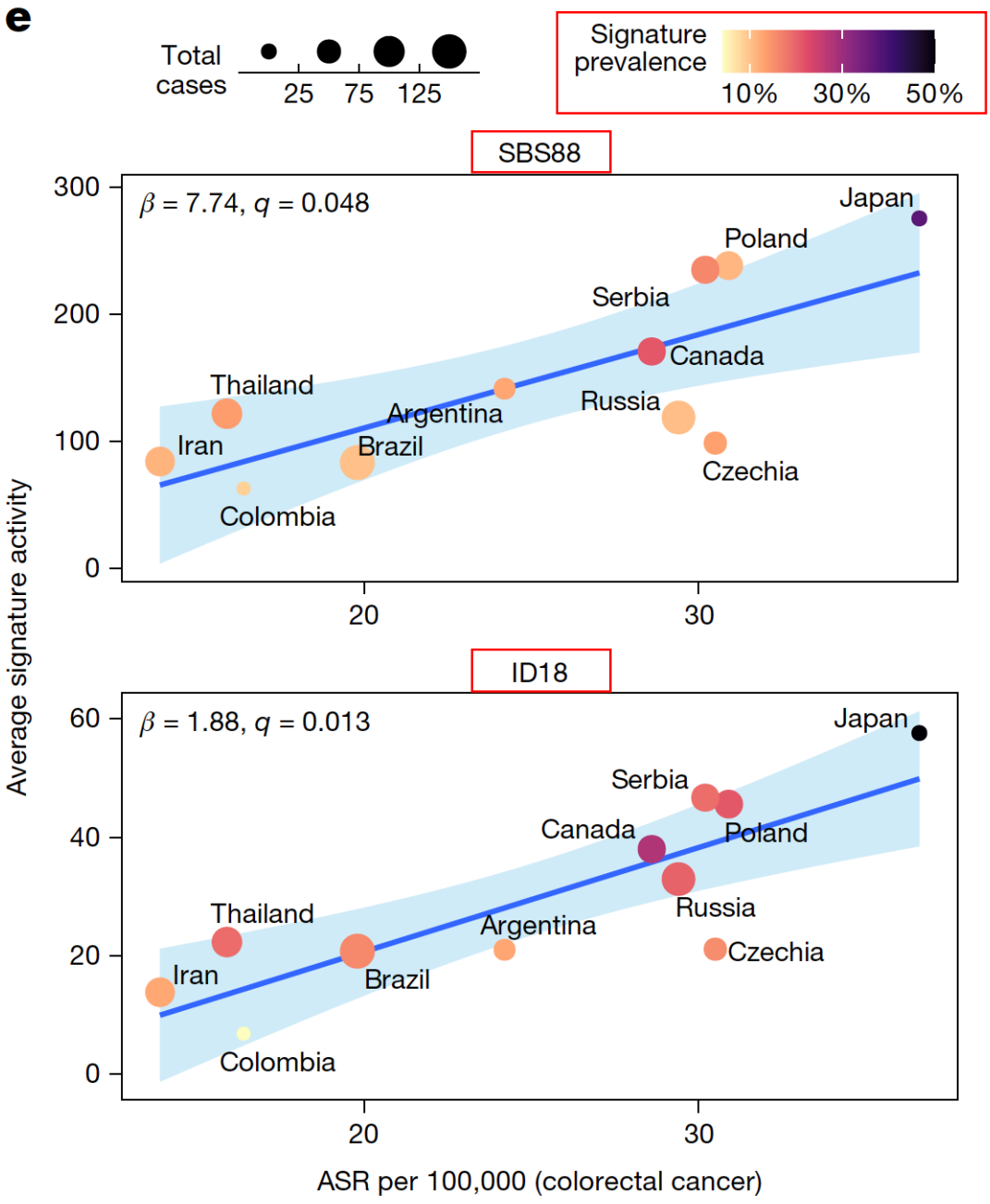
In this manuscript, Shiba and co-authors, investigate the interplay between gut microbiome in CRC and its mutational features, including the presence of the colibactin mutational signature in cohorts of Japanese patients with CRC. They first generate a microbiome-based interpretable AI CRC classifier, classifying CRC into 4 subtypes according to the presence (or absence in the case of Subtype 1) of major distinguishing bacterial species. Stage at diagnosis was more advanced in Subtype 2 and Subtype 3 was associated with a younger age on diagnosis and increased prevalence in those born after the 1980s. Additionally, Subtype 3 was enriched in the colibactin mutational signatures SBS88 and ID18, which, consistently with other studies were found at higher relative prevalence in CRCs from younger patients at diagnosis. They do also show, however, that these signatures are also increased in those born after 1965/by decade of birth which is novel. They ascribe *APC* hotspot, *TP53* and the *BRAF* D594G driver mutations to colibactin-induced mutagenesis and also show lack of correlation between colibactin itself (*clbP*) in stool (which is novel using stool WGMS) and its mutational signature in CRCs, with more variable results when tumor *clbP* was assessed. Interestingly, they show that colibactin induced mutations that are early clonal events and that these tumors have an earlier occurrence of CNN-LOH as well as higher Tregs in their TME.

Comments:

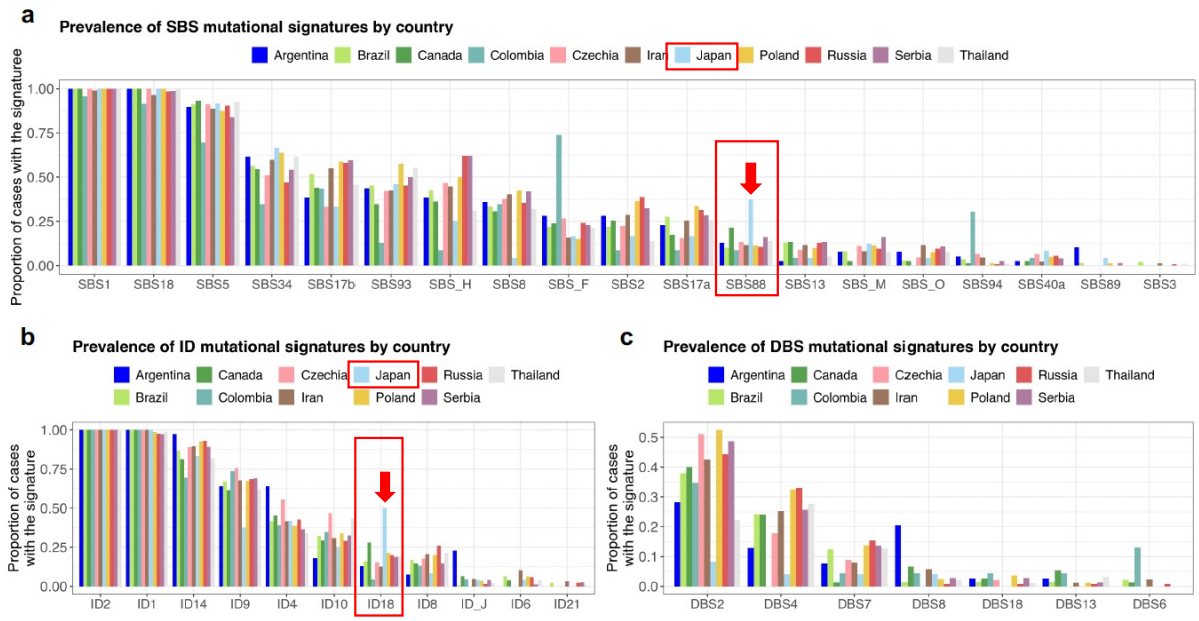
Comment 1. The authors state that the prevalence of SBS88 and/or ID18 was 44.8% in the Japanese cohort and highlight this as significantly higher than previously reported. However, in the seminal PMID 40267983 Nature article, from which the authors also integrated data for signature deconvolution, the prevalence of SBS88 or ID18 in countries with a high ASR of early-onset CRC (which includes Japan) is ~50-60% so actually similar, if not higher than this reported here. Also when SBS88 and ID18 are examined by decade of diagnosis their frequencies are lower in those <40 (for example) than in the Nature WGS paper.

Response:

We thank the reviewer for this important comment. The prevalence of SBS88 and/or ID18 in all cases is shown in Fig. 2e of the Nature study (PMID: 40267983) as reproduced below. In this figure, the circle representing Japan appears relatively small because only a limited number of Japanese cases were included; however, the prevalence of SBS88 and ID18 is visually higher in Japan compared with other countries. Although their Fig. 2e is somewhat difficult to interpret because of the colour scheme, **Supplementary Fig. 5a** and **5b** of their study present the same information as bar plots, which more clearly demonstrate that the prevalence of both SBS88 and ID18 is higher in Japan than in other countries.



Supplementary Figure 5



However, as the reviewer correctly noted, this observation was already reported in the previous study (PMID: 40267983) rather than being a novel finding of the present work. In addition, the thresholds and analytical conditions used to define SBS88 and ID18 positivity differ somewhat between the previous report and the present study, making a direct quantitative comparison difficult. For these reasons, we have removed the following sentence from the main text:

"The prevalence of these mutational signatures among Japanese CRCs is substantially higher than that reported in cohorts from other countries (approximately 5–20%)[11,14,24]."

The text has been revised as follows:

Page 13, lines 258–260:

"A previous study showed that the prevalence of these mutational signatures among Japanese CRCs is higher than that in cohorts from other countries[15]."

Reference:

15. Diaz-Gay, M. *et al.* Geographic and age variations in mutational processes in colorectal cancer. *Nature* **643**, 230-240 (2025).

The reviewer also pointed out that the prevalence of SBS88 and/or ID18 by age at diagnosis was not presented in the manuscript. We agree that this is important information and thank the reviewer for highlighting this omission. Also, please note that

in our original manuscript, Fig. 3b presents the average contribution of each mutational signature per case (i.e., the proportion of mutations attributed to each signature), rather than the prevalence of signature-positive cases. We apologize for any lack of clarity arising from this presentation. In response to the reviewer's suggestion, we have now added the prevalence of SBS88- and ID18-positive cases stratified by age at diagnosis in **Extended Data Fig. 9c**, and included the following description in the **Main** text and figure legend:

Page 13, line 258:

"This prevalence increased among younger patients (Extended Data Fig. 9c)."

Page 39, lines 812–815:

"**c**, Distributions of the prevalence of SBS88 (top panel) and ID18 (bottom panel) positive cases according to age at diagnosis. The prevalence increases in younger patients, consistent with the higher contribution of these signatures shown in Fig. 3b."

We are grateful to the reviewer for pointing out the importance of presenting the age-at-diagnosis-specific prevalence of SBS88 and ID18, and we agree that the inclusion of age-at-diagnosis–specific prevalence data strengthens the interpretation of SBS88 and ID18 signatures in our cohort.

Comment 2. *E.coli* (*pks+* strain or any) was not identified or at least shown among the bacteria present in Subtypes 1-4 (Figure 1, Extended Figure Data 2). It would be worth commenting whether it was picked up at lower levels that are not shown by the figures or completely absent.

Response:

We thank the reviewer for raising this point. In Fig. 1 and Extended Data Figs. 2b, c, only the top 10 bacterial taxa contributing most strongly to each subtype are shown. To provide a more comprehensive view, we have now included all bacterial taxa contributing to each subtype in Extended Data Table 4. *Escherichia coli* was detected in the metagenomic data and appears among the contributing taxa; however, the sequencing depth for *pks*$^{+}$ *E. coli* was insufficient to allow reliable detection of the pks genomic island or to enable robust comparisons across subtypes. Consequently, *pks*$^{+}$ *E. coli* could not be confidently quantified in the current dataset. We have clarified this point in the revised manuscript.

Page 9, lines 172–174:

"*Escherichia coli* was detected in the metagenomic data but was not among the most abundant taxa contributing to the subtype classification (Extended Data Table 4)."

Comment 3. In Extended Figure 6c, it is a little hard to distinguish the proportions of the individual signatures based on the current color scheme. If I interpret correctly that SBS 30 is present across cases, it is more likely that this is not due to *NTHL1* germline events (these are rare) but rather to alkylating damage present in CRC (PMID: 34140290, PMID: 41416871) which could be worth discussing. The etiology of SBS30 being that of CRC alkylating damage is also supported by its decrease with younger age at diagnosis (should be the opposite for *NTHL1* polyposis-related CRCs).

Response:

To improve the clarity of the figure, we have revised the color scheme used to represent the mutational signatures in Extended Data Fig. 6c and Fig. 3a, making the proportions of individual signatures more readily distinguishable. Following the reviewer's suggestion, we examined whether SBS30 in this cohort could be attributed to *NTHL1*-associated mutagenesis. However, no germline or somatic mutations in the *NTHL1* gene were identified in our dataset. This observation suggests that the SBS30 signal observed across cases is unlikely to reflect *NTHL1*-related tumorigenesis.

Recent studies have proposed that SBS30 observed in colorectal cancer more commonly reflects alkylating DNA damage rather than *NTHL1* deficiency (PMID: 34140290; PMID: 41416871). Supporting this interpretation, we observed that SBS30 was significantly less frequent in patients diagnosed at younger ages (Extended Data Table 13), which is inconsistent with the expected pattern for *NTHL1* polyposis–associated colorectal cancers. Accordingly, we have added the following text to the revised manuscript:

Page 14, lines 280-286:

"SBS30 was observed in most cases in this cohort (Fig. 3a; Extended Data Fig. 6c). 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 (Extended Data Table 13)."

References:

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).

Comment 4. It would be good to validate the CIBERSORT inferences for resting NK cells, plasma cells and mast cells with some orthogonal methods (multiplex IF, IHC), The authors do that nicely for Tregs. Otherwise, that limitation can be acknowledged.

Response:

We thank the reviewer for this suggestion. As the reviewer noted, immune cell fractions were estimated using CIBERSORTx-based digital cytometry applied to bulk RNA-seq data, which has inherent limitations in accurately resolving certain cell populations, and we agree orthogonal validation using immunohistochemistry or multiplex immunofluorescence would provide additional support for the immune cell composition inferred from our analysis.

However, while we were able to perform orthogonal validation for Tregs, we feel additional validation for plasma cells, resting NK cells, and mast cells is beyond the scope of the current study. To address this point, we have now acknowledged this limitation in the **Discussion** section of the revised manuscript. Specifically, the following sentences have been added to the limitations section:

Page 24, lines 495-500:

"In this study, 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."

We are currently pursuing further validation of these immune cell populations using single-cell transcriptomic analyses, which will be reported in a future study.

Comment 5. It would be worth specifying in the text whether all the specimens analyzed with genomics were the primary tumors versus metastatic lesions. This is relevant for the Stage IV patients. Extended Table 1 lists the primary tumor locations but it is still unclear if the primary was the sequenced specimen for Stage IV patients. In the same Table it is

also not clear what “Single” vs “Multiple” Lesions refers to for Stage IV patients. Single primary tumor (vs two primaries) or single metastatic lesion (oligometastatic CRC)?

Response:

We apologize for the lack of clarity in the original description. Whole-genome sequencing was performed on primary tumor specimens in all cases, including patients with Stage IV disease. Metastatic lesions were not used for genomic sequencing in this study. In addition, the notation “Single” versus “Multiple” lesions in Extended Data Table 1 refers to the presence of additional colorectal lesions within the colorectum, such as synchronous CRC, in addition to the index primary tumor. This classification does not refer to the number of metastatic lesions. To clarify these points, we have revised the description in the **Methods** and added more detailed information regarding concurrent colorectal lesions to **Extended Data Table 1**.

Page 1, lines 19-20 (Methods):

“Whole-genome sequencing (WGS) data were obtained from primary CRC tissues in 200 cases from Cohort-1.”

Comment 6. In Figure 2c, Stage 0 (carcinoma in situ) cases are included. Typically, since these are not invasive cancers, they can be excluded from CRC analyses.

Response:

We thank the reviewer for this helpful comment. We agree that Stage 0 lesions (carcinoma in situ) are non-invasive and that their inclusion in analyses of CRC stages may be inappropriate. Following the reviewer’s suggestion, we repeated the analysis after excluding Stage 0 cases. The results were consistent with those obtained in the original analysis, and the figure has been revised accordingly in the updated manuscript.

Comment 7. In Extended Figure 9, specify how “colibactin sequence motifs” were identified / reference/ method? (assuming reference 13?)

Response:

To clarify how the colibactin-associated sequence motifs were identified, we have added a description of the analytical method and the relevant references to the legend of Extended Data Fig. 9. The motifs were defined according to previously reported characteristics of colibactin-induced mutations (refs. 12, 13, 23). Specifically, the SBS88-associated motif was defined by the enrichment of adenines at positions −3 and −4 upstream of the T>N substitution site, whereas the ID18-associated motif was defined by

the presence of an adenine located 1–4 bases upstream of a single thymine deletion within a thymine homopolymer tract**.** The revised figure legend now reads as follows:

Page 38, lines 794–799:

"**a,** 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 proportion of adenines at positions −3 and −4 was calculated separately, and the average of these two values was used as the motif enrichment score (y-axis)[12,13]."

Page 39, lines 803–809:

"**b,** 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 1–4 bases upstream of a single thymine deletion within a 1–5 bp thymine homopolymer region. The proportion of adenine at positions −4 to −1 relative to the deletion site was calculated, and the average of these proportions was used as the motif enrichment score (y-axis) [12,23]."

References:

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).

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).

These additions clarify the method used to identify colibactin-associated sequence motifs in the revised manuscript.

Comment 8. In my opinion, the authors have addressed the comments of Reviewer #4.

Response:

We thank the reviewer for this positive assessment and we appreciate the reviewer's recognition that the previous comments from Reviewer #4 have been adequately addressed.

We are most grateful to the reviewers and the editor for their careful evaluation of our manuscript and for their constructive comments and suggestions. We hope that the revised manuscript is now suitable for publication in ***Nature Genetics*.**