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