



OPEN

Colorectal cancer-associated bacteria are broadly distributed in global microbiomes and drivers of precancerous change

Samuel S. Minot^{1,2}, Naisi Li³, Harini Srinivasan⁴, Jessica L. Ayers⁵, Ming Yu³, Sean T. Koester³, Mary M. Stangis⁶, Jason A. Dominitz^{7,8}, Richard B. Halberg⁶, William M. Grady^{3,7} & Neelendu Dey^{2,3,5,7}✉

The gut microbiome is implicated in the pathogenesis of colorectal cancer (CRC), but the full scope of this dialogue is unknown. Here we aimed to define the scale and membership of the body of CRC- and health-associated gut bacteria in global populations. We performed a microbiome-CRC correlation analysis of published ultra-deep shotgun metagenomic sequencing data from global microbiome surveys, utilizing a *de novo* (reference-agnostic) gene-level clustering approach to identify protein-coding co-abundant gene (CAGs) clusters. We link an unprecedented ~23–40% of gut bacteria to CRC or health, split nearly evenly as CRC- or health-associated. These microbes encode 2319 CAGs encompassing 427,261 bacterial genes significantly enriched or depleted in CRC. We identified many microbes that had not previously been linked to CRC, thus expanding the scope of “known unknowns” of CRC-associated microbes. We performed an agnostic CAG-based screen of bacterial isolates and validated predicted effects of previously unimplicated bacteria in preclinical models, in which we observed differential induction of precancerous adenomas and field effects. Single-cell RNA sequencing disclosed microbiome-induced senescence-associated gene expression signatures in discrete colonic populations including fibroblasts. In organoid co-cultures, primary colon fibroblasts from mice with microbiomes promoted significantly greater growth than fibroblasts from microbiome-depleted mice. These results offer proof-of-principle for gene-level metagenomic analysis enabling discovery of microbiome links to health and demonstrate that the microbiome can drive precancer states, thereby potentially revealing novel cancer prevention opportunities.

Abbreviations

CRC	Colorectal cancer
CAGs	Co-abundant genes
SEM	Standard error of the mean
GF	Germ-free
SPF	Specific pathogen-free
PERMANOVA	Permutational multivariate analysis of variance
SASP	Senescence-associated secretory phenotype
FIT	Fecal immunochemical testing

Colorectal cancer (CRC) is a common cause of cancer death and suffering worldwide¹. Epidemiological and disease cohort studies implicate the gut microbiome as a critical determinant of CRC risk^{2,3}. However, while

¹Data Core, Shared Resources, Fred Hutchinson Cancer Center, 1100 Fairview Ave North, Seattle 98109, WA, USA. ²Microbiome Research Initiative, Fred Hutchinson Cancer Center, Seattle, WA, USA. ³Translational Science and Therapeutics Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. ⁴Genomics & Bioinformatics Core, Shared Resources, Fred Hutchinson Cancer Center, Seattle, WA, USA. ⁵Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA. ⁶University of Wisconsin Carbone Cancer Center, University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA. ⁷Division of Gastroenterology, Department of Medicine, University of Washington, Seattle, WA, USA. ⁸National Gastroenterology and Hepatology Program, U.S. Department of Veterans Affairs, Seattle, WA, USA. ✉email: ndey@fredhutch.org

several carcinogenic gut microbes have been identified (e.g. enterotoxigenic *Bacteroides fragilis*, colibactin-producing *Escherichia coli*, and *Fusobacterium nucleatum*), each appears to account for a minority of CRC cases, and estimated effect sizes are modest^{3–7}. Further, recent metagenomic studies of CRC microbiomes implicate increasing numbers of CRC-associated bacteria and diverse bacterial communities^{3,8–12}, with tremendous inter-study variation complicating interpretation. Collectively, the literature appears to support the Anna Karenina Principle^{13,14}, i.e., that disease-associated dysbioses are more varied than health-associated microbiome configurations, a phenomenon inspired by the opening line of Tolstoy’s famous book¹⁵. We were therefore motivated to better understand this variation and to define the scope of CRC- and health-associated microbes. We postulated microbial associations more numerous and diverse than currently recognized.

To identify gut bacterial links to CRC across global cohorts, we performed a meta-analysis of all published microbiome surveys with ultra-deep shotgun metagenomic sequencing data. Our aggregate dataset represented 5 countries spanning North America, Europe, and Asia (Table 1)^{9–12}. Acknowledging strain-level variability in gut bacterial effects on host physiology (with potential for vastly different effects on tumorigenesis¹⁶) and the consequent dampening of signal-to-noise with traditional taxonomy-based approaches (notably, the typical lack of sub-species resolution in predictive classifiers built on metagenomic data), we performed a microbiome-CRC correlation analysis with gene-level resolution, utilizing an updated version of our previously reported de novo taxonomy-free reference-agnostic gene clustering approach to estimate relative abundances of protein-coding co-abundant gene (CAGs) clusters in ultra-deep shotgun metagenomic sequencing datasets^{17,18}. CAGs are non-overlapping collections of microbial genes observed at tightly correlated abundances across multiple microbiome samples, representing discrete genetic elements shared across multiple strains (i.e., a sub-species grouping)¹⁹. CAGs offer the opportunity to gain insights unconstrained by a more traditional taxonomic lexicon (e.g., species-level associations, in which biologically meaningful correlations may be obscured by strain-to-strain variability, or strain-level associations, in which statistical power may be compromised). As a result, the breadth of our findings using this approach far exceeds the scope of microbiome-CRC associations reported to date.

Materials and methods
Metagenomic meta-analysis

We identified all published studies with ultra-deep shotgun metagenomic sequencing data (“ultra-deep” defined here as > 10⁷ reads/sample) and comparable metadata for use in our meta-analysis. Data from a total of 992 people (5.5 terabases of sequencing data) were analyzed. Shotgun metagenomic sequencing data were downloaded from the NCBI Sequence Read Archive (SRA) repository (training dataset: BioProject IDs PRJEB6070, PRJEB7774, and PRJEB10878; test/validation dataset: BioProject ID PRJDB4176). Raw shotgun sequencing data was processed by the *geneshot* pipeline (revision v0.9, commit 6d2c08e, <https://github.com/Golob-Minot/geneshot/>)¹⁸. The *geneshot* analysis encompassed human sequence removal (bwa mem); de novo assembly of contigs independently within each specimen (megahit); prediction of protein-coding gene sequences from those contigs (prodigal); de-duplication of amino acid sequences across all samples to form a gene catalog (mmseqs2); alignment of raw shotgun sequencing data against that gene catalog (DIAMOND); probabilistically assigning multi-mapping reads (FAMLI); and clustering co-abundant genes into CAGs using a maximum cosine distance threshold of 0.2¹⁷. The gene-level metagenomic analysis we employed here was updated from our initial proof-of-concept¹⁷ by (i) expanding the collection of input datasets used to train the CAG model and (ii) updating the analytical approach such that the metagenomic space could be coalesced into 40% fewer CAGs, which permits more statistically efficient discovery of differences in abundances between groups. The coefficient of association of each CAG with CRC status was estimated by beta-binomial regression with corncob²⁰, controlling for systematic differences in microbial abundances across studies. Taxonomic annotation of the gene catalog was performed by amino acid alignment to the NCBI RefSeq database (downloaded January 27, 2021), computing the lowest common ancestor of all organisms within 5% of the top-scoring alignment (DIAMOND). For benchmarking, taxonomic classification of raw reads was also performed by the *geneshot* pipeline using MetaPhlAn (v2.8). The coefficient of association for each MetaPhlAn-reported species was estimated by beta-binomial regression with corncob. For validation testing, raw shotgun sequencing reads from the test dataset were aligned against the gene catalog generated from the training dataset to estimate the relative abundance of the CAGs in that independent cohort. The relative abundance of organisms in the validation dataset was not used to generate the gene catalog, CAG groupings, or CRC-associations reported in this study.

Microbial genome annotation

The presence of genes identified from metagenomic analysis in individual microbial genomes was performed by amino acid alignment of the six-frame conceptual genome translation (using the bacterial NCBI genetic code

Dataset role	NCBI BioProject ID	References	CRC cases	Adenoma cases*	Healthy controls	Median sequencing depth
Training data	PRJEB10878	Yu et al. ¹¹	75	0	53	56.2 M reads/sample
Training data	PRJEB6070	Zeller et al. ⁹	91	42	66	58.0 M reads/sample
Training data	PRJEB7774	Feng et al. ¹⁰	46	47	63	52.7 M reads/sample
Test data (validation)	PRJDB4176	Yachida et al. ¹²	258	0	251	46.3 M reads/sample

Table 1. Published microbiome datasets analyzed. *Adenoma metagenomes were not used for training the CAG-based model.

11), and subsequent filtering of any lower-scoring alignments which overlap higher-scoring alignments by more than 50%. Every gene-genome alignment was performed with the *gig-map* suite of bioinformatics tools which is freely accessible at <https://github.com/FredHutch/gig-map/wiki> (commit cfe0056). The RefSeq database used for alignment of CRC-associated microbial genes was downloaded on June 6, 2022 ($n = 113,938$). The analysis of *Enterobacter* genomes in RefSeq included 90 genomes, with ≥ 1 representative from each *Enterobacter* species in the RefSeq database, including all genomes available for the specific species mentioned.

CRC-association scores

Aggregate CRC-association scores were calculated for microbial genomes as the average CRC-association Wald metric for each of the genes from the metagenomic gene catalog which is found within the genome at 90% amino acid identity and 90% gene coverage. Similarly, the CRC-association score was computed for each metagenome by weighting each detected gene by its relative abundance (based on the length-adjusted proportion of metagenomic reads aligning to genes in CAGs).

Human subjects

To validate metagenome-based CRC-association scores in a local cohort, informed consent was obtained from all study participants, who were recruited via Fred Hutchinson Cancer Center Institutional Review Board (IRB) protocols 10337 and 10084. Study participants were asked to collect a fecal sample at home using a sterile stool collection kit ("Feces catcher" from Zymo Research and standard polypropylene sterile specimen collection container), store it immediately in their home freezer, and then mail it overnight on ice the next day using pre-paid shipping methods. Upon receipt, fecal samples were immediately stored at -80°C until use. Fecal samples from the CRC cohort were collected prior to initiation of treatment. The 'history of advanced adenoma' cohort was defined as individuals who had been previously found to have high-risk precancerous adenomas (based on larger size and/or villous histology) that had been resected via colonoscopy. From this cohort, 1–2 fecal samples were collected per person for a total of 19 samples. From the healthy control cohort, 1–2 fecal samples were collected per person for a total of 5 fecal samples. Fecal samples were subjected to metagenomic sequencing.

Preclinical mouse experiments

All experimental protocols were approved by the Institutional Animal Care and Use Committees of Fred Hutchinson Cancer Center and the University of Washington, and all methods were carried out in accordance with institutional guidelines and regulations. Experiments testing effects of different bacterial consortia on tumorigenesis and colonic gene expression were performed on male and female *Apc*^{Min/+} mice (an established CRC model), offspring of *Apc*^{Min/+} and *Il10*^{-/-} mice (a reported microbiome-responsive CRC model²¹), and wild-type C57BL/6 mice that were 6–8 weeks old at the start of experiments. Mice were randomly assigned to treatment groups. Gnotobiotic animal breeding and husbandry was performed as previously described²². Mice were fed an irradiated diet with moderately high fat content (9%) ad libitum. To deplete the resident mouse microbiota of conventionally raised specific pathogen-free mice, an antibiotic cocktail (ampicillin 1 g/L, neomycin 500 mg/L and metronidazole 1 g/L) in sucrose-containing water (20 g/L) was administered for 10 days. Anaerobic preparation of bacterial consortia and uncultured intact human fecal microbiota was performed as previously described²³. Microbiota transplantation was performed via oral gavage (5 times over 10 days). *Apc*^{Min/+} mice were euthanized 10 weeks after initial gavage, and offspring of *Apc*^{Min/+} and *Il10*^{-/-} mice were euthanized 8 weeks after initial gavage (earlier than planned due to institutional guidance early in the COVID pandemic). Small intestines and colons were harvested, measured, and filleted open longitudinally for visual inspection and tactile examination to identify, quantify, and measure tumors. Fresh fecal pellets were collected from mice throughout experiments in sterile tubes, snap-frozen in liquid nitrogen, and stored at -80°C until use. All methods are reported in accordance with ARRIVE guidelines.

Histologic analysis

Small intestinal and colonic tissues harvested from mice were subjected to hematoxylin and eosin and Ki-67 staining. Inflammation was scored by a pathologist blinded to group using a modification of an established scoring criteria²⁴. Specifically, scoring schemes 4 and 5 were used for colon and small intestine, respectively. Scores range from 0 to 5, with a score of 1 being mild changes and 5 representing marked pathological changes. These schemes were modified to include only the 'severity', 'extent', and 'mucosal architecture' parameters; the 'epithelial changes' category was excluded, as all mice had adenomatous polyps and thus would automatically be categorized as marked hyperplasia, category 5.

Multiplexed shotgun metagenomic sequencing of fecal samples

Genomic DNA was extracted from fecal samples using a phenol/chloroform-based protocol. Genomic DNA was sheared using a Covaris LE220 ultrasonicator (Covaris, Woburn, MA). Sample-specific barcoded adapters were ligated to end-repaired DNA fragments, after which libraries were prepared for sequencing on the Illumina MiniSeq platform (75 nt single-end reads). Reads mapping to the mouse genome (UCSC mm10; Bowtie2, version 2.2.5²⁵) or estimated to have > 1 error (USEARCH, version 11²⁶) were removed; and taxonomic classifications were performed with MetaPhlAn2²⁷ using default parameters. CRC-association scores were calculated as described above.

Cell viability assay

Caco-2 cells were purchased from American Type Culture Collection (ATCC, Manassas, VA, USA). To test for growth-promoting secreted bacterial products of the CRC-associated consortium and health-associated consortium, we cultured Caco-2 cells in media containing fecal filtrates from gnotobiotic mice colonized with these consortia and quantified cell viability as previously described²⁸.

Bulk RNA sequencing of the colon

RNA was isolated from mouse colons and run on NanoString nCounter Tumor Signaling 360 panels, with data analysis performed as previously described²². The gene *Gdf15* was quantified in the same samples using previously described qRT-PCR primers and methods²⁸.

Single-cell RNA sequencing and analysis

Mouse colonic tissue samples were disassociated and processed for single-cell sequencing by the Fred Hutchinson Cancer Center Genomics Core using the 10X Chromium Single Cell 3' Reagent Kits User Guide (v3.1 Chemistry) on the Illumina NovaSeq platform (S2 flow cell). Differential relative gene expression was tested within each cell subset independently, comparing the level of expression of each gene between the two groups of mice. Statistical significance of differences in gene expression were calculated with DESeq2²⁹ v1.34.0 using the Nextflow workflow FredHutch/pw-differential-expression (<https://github.com/FredHutch/pw-differential-expression>, commit 1e2d036). Gene set enrichment analysis³⁰ was performed with fgsea³¹ v1.20.0 using the Nextflow workflow FredHutch/pw-gene-set-enrichment (<https://github.com/FredHutch/pw-gene-set-enrichment>, commit 3686e70). Hallmark gene pathways v7.5.1 were downloaded from MSigDB³². Gene expression-based classification of cells as senescent was performed using enrichment scores based on the SenMayo panel of 125 genes³³ was calculated for each cell. As prescribed in the original paper,³³ we considered the top 10% of cells to be senescent.

Assessment of microbiome-induced senescent fibroblast growth effects in organoids

Microbiota depletion was performed in a subset of mice as described above prior to harvesting primary colonic fibroblasts and generating organoids. Mouse colon organoids were generated using a modified version of an established protocol³⁴. Briefly, following euthanasia, mouse colons were harvested, fecal matter was removed, and colons were washed repeatedly with ice-cold phosphate-buffered saline (PBS; Gibco) until no visible debris remained. Colons were cut into ~5 mm fragments and shaken 15 × in ice-cold 5 mM ethylenediaminetetraacetic acid (EDTA)-PBS buffer, then allowed to settle by gravity on ice. The 5 mM EDTA-PBS buffer was replaced, and fragments were incubated for 30 min on a benchtop roller at 4 °C. Supernatant was removed and replaced with 10 mL ice-cold PBS. Colon fragments were vigorously shaken for 10 s, and the PBS, now containing intestinal crypts, was collected. The PBS crypt collection was repeated a total of 3 times, and an equal volume of Basal Media containing 5% fetal bovine serum (FBS; Sigma) was added to each fraction. The fractions were then filtered through a 100 µm and 70 µm strainer. An aliquot of each collection fraction was examined under a microscope for the presence of intact crypts. Once the presence of intact crypts was confirmed, fractions were pooled. Intestinal crypts were pelleted at 4 °C, washed with Basal Media containing 5% FBS, pelleted again and supernatant removed. The intestinal crypt pellet was suspended in ice cold Matrigel (Corning) and 30 µL droplets were plated onto a 24-well plate. The Matrigel was allowed to harden for 15 min at 37 °C, and 500 µL of IntestiCult Mouse Organoid Growth Medium (StemCell Technologies) plus 10 µM Rho Kinase Inhibitor Y-27632 (Fisher Scientific) was added to each well. Intestinal organoids were grown under standard tissue culture conditions. Media was changed every 2–3 days, and organoids were passaged weekly. Mouse primary colon fibroblasts were isolated and cultured as previously described³⁵. Briefly, following euthanasia, mouse colons were harvested, fecal matter was removed, and colons were washed repeatedly with ice-cold PBS until no visible debris remained. Colons were then added to a 50 mL conical tube containing 25 mL of room temperature 5 mM EDTA-Hanks' Balanced Salt Solution (HBSS) (Gibco) and incubated in a shaking air bath at 37 °C for 15 min. Supernatants were removed and replaced with fresh 5 mM EDTA-HBSS, followed again by an incubation period; this cycle was repeated a total of 5 times. Colons were washed twice with ice-cold PBS and 20 mL of "RPMI-5 buffer" (recipe for 500 mL: 454.5 mL of Roswell Park Memorial Institute 1640 buffer [RPMI], 25 mL of FBS, 5 mL of 200 mM L-glutamine, 5 mL of 1 M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid [HEPES] at pH 7.4, 5 mL of 100 mM sodium pyruvate, 5 mL of 100x Penicillin-Streptomycin, and 500 mL of 50 mM beta-mercaptoethanol in PBS; Sigma) containing 10 U of dispase (Sigma), and 2000 U of collagenase D (Sigma) was added. Tissues were incubated in a shaking air bath at 37 °C for 30 min and then pelleted. Supernatants were removed, and filter-sterilized room temperature Ammonium-Chloride-Potassium (ACK) lysing buffer (Gibco) was used to resuspend the tissue pellet. After resuspension, tissues were once again pelleted at 4 °C and resuspended in RPMI-5 and passed through a 70 µm filter. Fibroblast suspensions in media were plated in T75 tissue culture flasks and grown at 37 °C. After an overnight incubation, cells were washed gently with RPMI-5 media, allowing only adherent cells to remain. Media was changed every 2–3 days and fibroblasts were passaged after 5 days. Mouse colon organoid and fibroblast co-cultures were performed using a modified version of our published protocol²⁸. Briefly, mouse colon organoids were dissociated into a single cell suspension utilizing TrypLE express enzyme (Thermo Fisher). Once organoids were dissociated, an equal volume of basal organoid media plus 10 µM Rho Kinase Inhibitor Y-27632 was added to the suspension, and cells were pelleted at 4 °C. A live cell count was taken, and 2.5×10^4 cells were plated in a 30 µL droplet of ice-cold Matrigel per well in a 24-well deep well TC treated plate (Corning), with six replicates per organoid line plated. Four independent organoid lines were used (2 isolated from mice with a gut microbiome and 2 from mice treated with antibiotics for microbiota depletion). The Matrigel was allowed to harden at 37 °C for 15 min, and 500 µL of IntestiCult Mouse Organoid Growth Medium plus 10 µM Rho Kinase Inhibitor Y-27632 was added to each well. Corning HTS Transwell inserts were placed into

each well. Two mouse fibroblast lines were trypsinized, collected, and pelleted (one isolated from a mouse with a gut microbiome and the other from a mouse treated with antibiotics for microbiota depletion). Each line was resuspended in RPMI-5 media. Live cell counts were taken, and 6.25×10^5 cells were plated onto each insert in 500 μ L of media. Co-cultures were seeded in duplicates, and ‘no co-culture’ controls were performed. The co-culture was grown under standard conditions for 4 days. After 4 days, the inserts containing fibroblasts were removed, and the organoid media was gently aspirated. 400 μ L of fresh Basal Media and 400 μ L of CellTiter-Glo 3D Cell Viability Assay (Promega) were added to each well. Manufacturer’s instructions were followed. The luminescence signal was allowed to stabilize for 30 min in the dark. Each well was aliquoted into three technical replicates in an opaque well 96-well plate. The luminescence signal was read on a Varioskan LUX microplate reader (Thermo Fisher), and total cell metabolic activity, as represented by the luminescence readout, was recorded.

Mock fecal immunochemical testing (FIT) kit assays

A FIT sampling bottle contains 2 mL of HEPES buffer (Polymedco), and users are instructed to provide a small amount of stool that covers the grooved portion of the stick (estimated to be < 50 mg³⁶). To mimic clinically used FIT kits, “mock FIT kits” were prepared using 50 mg aliquots of 1 of 4 different human fecal samples in 2 mL HEPES buffer in an anaerobic chamber. Aliquots were allowed to incubate for 0, 1, 3, 7, or 14 days at -80 °C, -20 °C, 4 °C, 25 °C, or 37 °C (with each unique condition tested in replicate) before being subjected to genomic DNA extraction and deep shotgun sequencing on the Illumina NextSeq 2000 platform (150 nt paired-end reads). Metagenomic analysis and calculation of CRC-association scores were performed as above.

Data analysis

Statistical comparisons were performed in R (version 4.0.0) and Python (version 3.8). Figures were generated using built-in functions in R as well as functions from the *ggplot2* (version 3.1.0) and *pheatmap* (version 1.0.12) packages. Boxplots represent median (center line), the 25th and 75th percentiles (box limits), and $1.5 \times$ interquartile range (whiskers); outliers beyond the whiskers are represented as points. Error bars in bar plots represent standard error. Figures were assembled in Adobe Illustrator.

Results

CAGs identify previously unknown CRC-microbiome associations

In our training dataset of ultra-deeply sequenced samples ($55,828,959 \pm 826,869$ reads/sample [mean $\pm$ SEM], median 56,177,744 reads/sample), we identified a total of 22,295 CAGs representing 1,858,307 microbial genes. Individual fecal microbiomes harbored between 10,280 and 426,560 microbial genes ($226,879 \pm 3,441$ [mean $\pm$ SEM]) and between 2655 and 13,094 CAGs (8284 ± 88 CAGs/sample [mean $\pm$ SEM]; Fig. 1A). Within microbiomes, numbers of assembled microbial genes and of CAGs were significantly correlated (Pearson correlation $\rho = 0.81$, $p < 10^{-15}$; Fig. 1B), and each was correlated with sequencing depth (genes: $\rho = 0.48$, $p < 10^{-15}$; CAGs: $\rho = 0.29$, $p < 10^{-9}$). CAGs were bimodal in size (Fig. S1) and much smaller on average (83.4 ± 4.0 genes/CAG [mean $\pm$ SEM]) than gut bacterial genomes³⁷. Neither bacterial genes per person nor CAGs per person were significantly different between CRC and healthy cohorts (Fig. 1A).

Using a beta-binomial regression model,²⁰ we identified 2319 distinct microbial CAGs harboring 427,261 total genes whose abundances in fecal microbiomes were consistently and significantly correlated ($q < 0.05$) with either CRC (1128 CAGs) or health (1191 CAGs) (Fig. S2A; Table S1A)¹⁷. These 2,319 CRC-or-health-associated CAGs were observed at varying abundances spanning several orders of magnitude, from extremely rare to highly prevalent, encoded in the genomes of phylogenetically diverse bacteria (Figs. 1C, S2B). Consistent with previously reported observations, CRC-associated CAGs are enriched in the *Bacteroidetes*, *Proteobacteria*, and *Fusobacteria* phyla, while health-associated CAGs were enriched in *Actinobacteria* and *Firmicutes* phyla ($p < 0.005$, Fisher’s exact test; Figs. 1D, S2C). Remarkably, many of the CAGs enriched in CRC were found in bacterial species that are not recognized as CRC-associated (e.g. *Bacteroides cellulosilyticus*, *Bacteroides thetaiotaomicron*, *Prevotella copri*, *Ruminococcus torques*, *Dialister pneumosintes*, *Dorea formicigenerans*, and *Faecalibacterium prausnitzii*; Table S1B). These associations were often complex and variable within species: for example, 596 CAGs mapped to *F. prausnitzii* genomes, and significant associations spanned from strongly health-linked to strongly CRC-linked, with Wald statistics ranging from -3.7 to 3.8 (Table S1A–D). In our benchmarking taxonomy-based analysis of the same data, *F. prausnitzii* had an estimated coefficient of -0.25 (health-associated) that was non-significant ($q = 0.4$). *F. prausnitzii* was not unique in this regard: members of numerous bacterial species lacked significant correlations with CRC or health in a taxonomy-based analysis but harbored CAGs with significant CRC-or-health associations that were often wide-ranging within species (Fig. 1E, Table S1A,D), illustrating the granularity of CAG-level resolution in metagenomic analysis.

Of the 22,295 CAGs identified in total, 10.5% of CAGs (representing 20.4% of total genes) were significantly enriched in CRC, while 12.0% of CAGs (16.9% of total genes) were significantly enriched in health (totaling $\sim 23\%$ of all CAGs). When CAGs were weighted by mean relative abundances across all samples in the aggregated dataset, 18.7% and 21.7% of CAGs were significantly enriched in CRC or in health, respectively, totaling 40.4% of all CAGs. In other words, upwards of $\sim 40\%$ of gut bacteria may be carcinogenic, chemoprotective, or secondarily affected by CRC. In contrast, in a traditional taxonomy-based analysis, just 25 bacterial species were significantly correlated with either CRC or health (i.e., 11.4% of all taxa identified at $\geq 0.1\%$ abundance; Table S1E). Of these, 8 were annotated taxonomically as being “unclassified” members of a particular genus (e.g. *Butyrivibrio*_unclassified), while our CAG-based approach identified species harboring CRC-or-health-associated CAGs (e.g. *Butyrivibrio crossotus*, *Butyrivibrio hungatei*, *Butyrivibrio* sp. M55, *Butyrivibrio* sp. XPD2006, *Butyrivibrio* sp. VCD2006), underscoring limitations of traditional taxonomy-based approach and further evidencing the potentially greater sensitivity of a CAG-based approach.

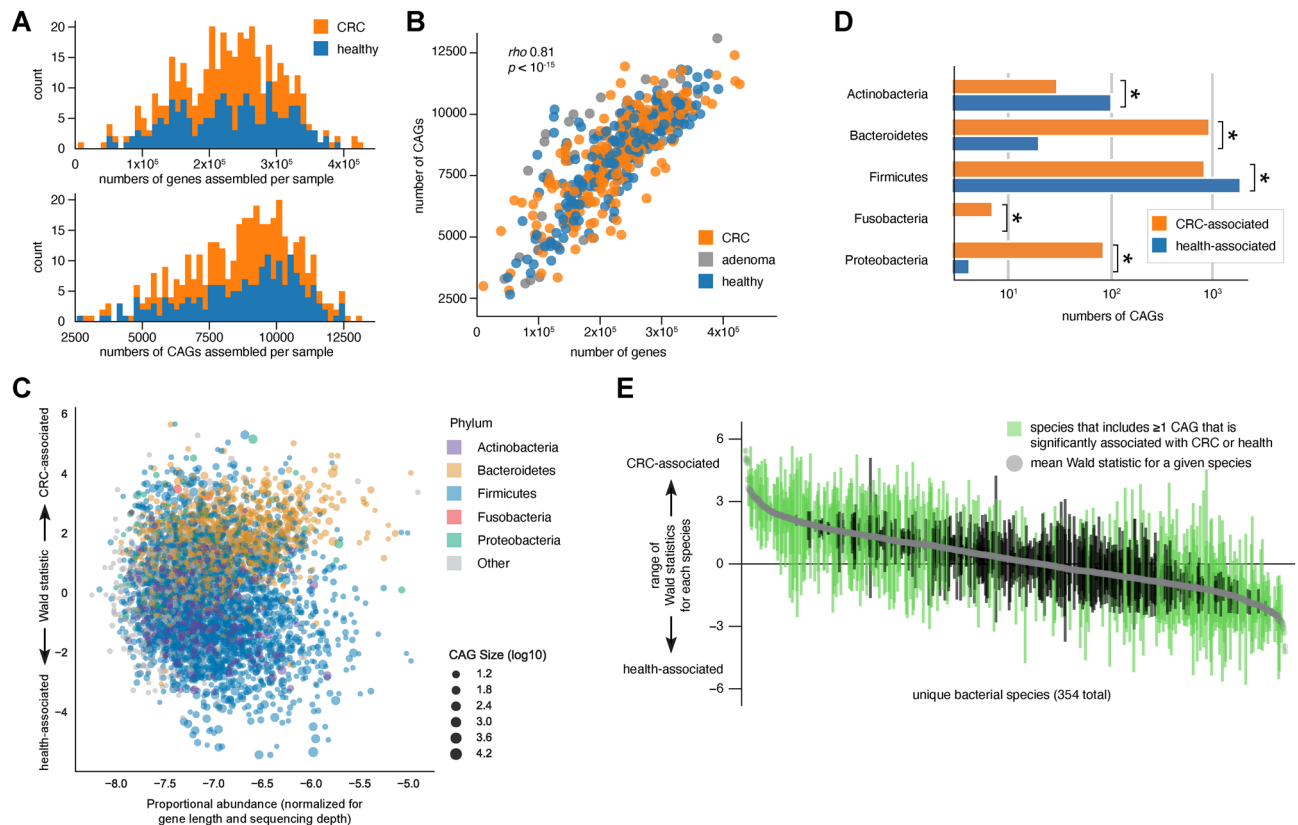


Fig. 1. Association of diverse microbial CAGs with CRC and health. **(A)** Stacked histogram depicting numbers of genes assembled per sample (top) and CAGs per sample (bottom). **(B)** Correlation between numbers of CAGs and genes across the aggregated training dataset, color-coded by cohort. The gray dots, which represent metagenomes from individuals with adenomas, were not used in generating our CAG-based model. **(C)** Bacterial CAGs that are enriched in CRC or in health are encoded in the genomes of phylogenetically diverse bacteria observed at varying abundances, from extremely rare to highly prevalent. **(D)** Phylum-level differences between CRC-associated and health-associated CAGs. **(E)** CRC-association Wald statistics of all identified CAGs, collapsed by species-level taxonomic classifications and rank ordered by mean Wald statistic.

Divergence of gut bacterial CAGs in CRC and in health

Consistent with expectations, a significantly greater proportion of sequencing reads generated from CRC microbiomes aligned to CRC-associated CAGs than to health-associated CAGs ($p < 10^{-16}$, Mann–Whitney U test; Fig. 2A). Conversely, healthy microbiomes were characterized by a greater proportion of reads mapping to health-associated CAGs than CRC-associated CAGs ($p < 10^{-18}$, Mann–Whitney U test; Fig. 2A). The log-ratio of sequencing reads mapping to CRC-associated CAGs versus health-associated CAGs was significantly greater in CRC microbiomes than healthy microbiomes ($p < 10^{-18}$, Mann–Whitney U test; Fig. 2B). We validated our findings in an independent dataset¹² and similarly observed a significantly greater representation of CRC-associated CAGs in CRC microbiomes than in healthy microbiomes ($p < 0.007$, Mann–Whitney U test comparing log-ratios; Fig. S3A).

Interestingly, the proportion of reads of any metagenomic sample that mapped to health-associated genes was inversely proportional to the proportion mapping to CRC-associated genes (Figs. 2C, S3B). Within microbiomes (i.e., in correlation analyses performed within individual metagenomes), CRC-associated CAGs were more likely to co-occur with other CRC-associated CAGs than they were with health-associated CAGs, and vice versa (Fig. 2D,E). This divergence was also observed across microbiomes despite the sample-to-sample variability, population-to-population variability, and diversity of CRC-associated or health-associated CAGs (Fig. S4). CRC-associated and health-associated CAGs may reflect antagonistic ecological forces and distinct ecosystems of synergistic CAGs.

We next asked whether this divergence exists within individual bacterial genomes as well. Using the NCBI Reference Sequence Database (RefSeq), we quantified the proportion of RefSeq bacterial genomes encoding CRC-or-health-associated CAGs by aligning the complete CAG catalog from our metagenomic analysis against the RefSeq database. Of the 12% of RefSeq genomes containing $\geq 1\%$ CRC-or-health-associated CAGs, genomes tended to contain only positively or negatively CRC-associated genes, but typically not both, with the exception of 9 genomes (Fig. 2E,G) that belong to *Enterobacter* (Figs. 2G, S5), a genus of gut bacteria which have previously been associated with human cancers^{38,39}. Nonetheless, the dominant trend seen in our metagenomic dataset of inversely proportional representation of CRC-associated or health-associated signatures was seen also within individual bacterial genomes.

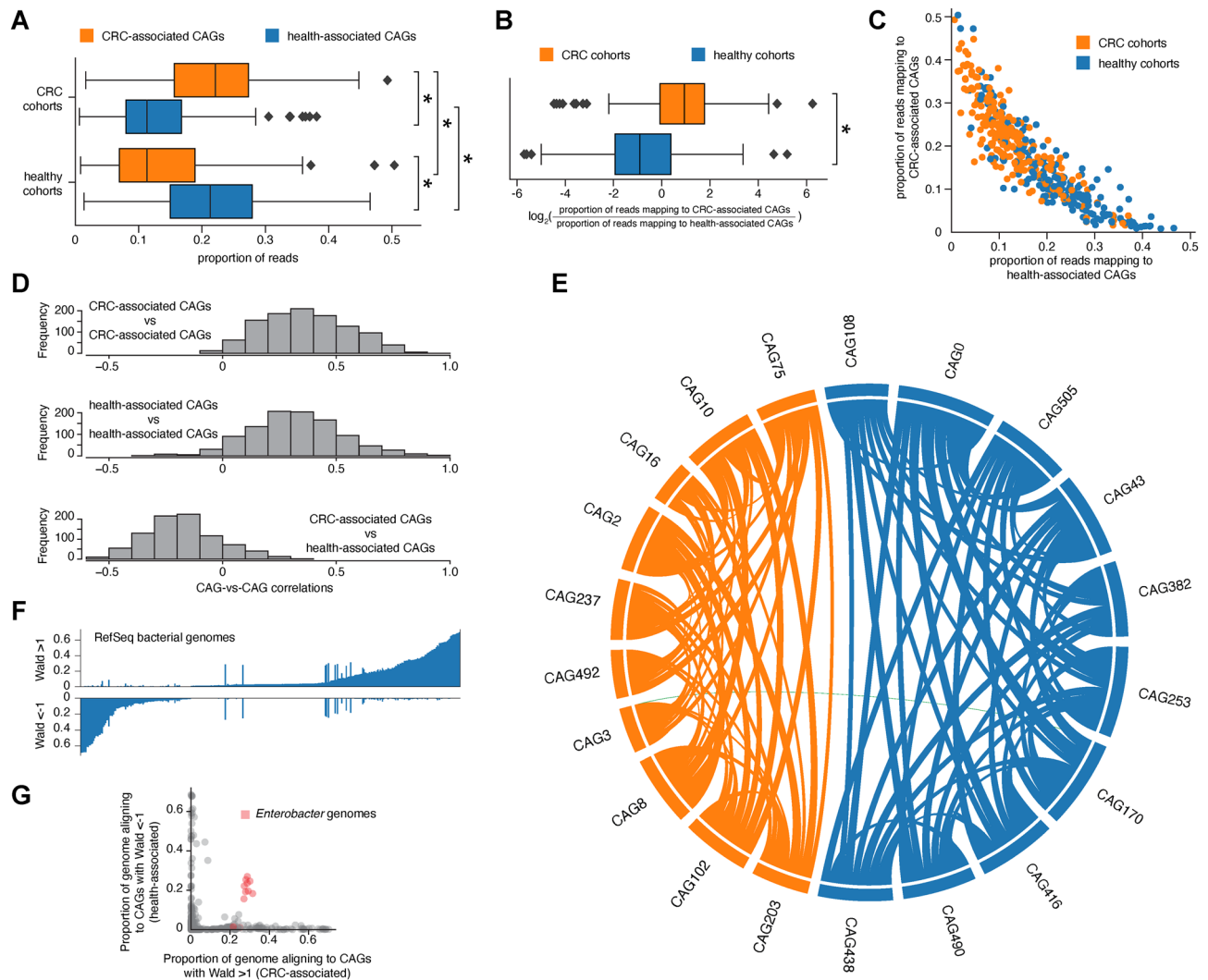


Fig. 2. Divergence of CRC-associated CAGs and health-associated CAGs. **(A)** Proportion of metagenomic shotgun sequencing reads from CRC cohorts vs healthy cohorts that align to CRC-associated or health-associated CAGs. **(B)** Log ratio of CRC-vs-health-associated CAG aligning proportions of reads in the training dataset. **(C)** Proportion of reads within samples aligning to CRC-associated or health-associated CAGs in the training dataset. **(D)** Significant correlations between CRC-associated CAGs, between health-associated CAGs, and between CRC-associated and health-associated CAGs. **(E)** Circos plot showing significant correlations between top CRC-associated and health-associated CAGs. Ribbon widths indicate correlation coefficients. Correlations between health-associated CAGs and between CRC-associated CAGs are represented by blue and orange ribbons, respectively, while a single significant correlation between a CRC-associated CAG (CAG3) and health-associated CAG (CAG170) is shown in green. **(F)** Proportions of individual RefSeq genomes aligning to CAGs with Wald statistic of > 1 or < -1 using a ≥ 1% genome alignment threshold. **(G)** Genomes of 9 *Enterobacteriaceae* aligned to both CRC-associated and health-associated CAGs.

CAG-based CRC-association scores are greater in precancer states than in health

To apply our findings more easily to microbiomes or individual bacterial strains, we then created a summary metric that can be calculated for any given metagenome or genome:

$$x = \sum_{i=1}^n (abundance_i * Wald_i)$$

where x is the ‘CRC-association score’; n is the total number of CAGs; i is each individual CAG; $abundance$ is the proportion of reads aligning to the set of genes in each CAG (adjusting for gene length) such that the sum of abundances across all n CAGs is 1; and $Wald$ metric is the degree of association of a CAG with either CRC (positive $Wald$) or health (negative $Wald$). This scoring metric faithfully recapitulated our earlier analysis: CRC cohorts had significantly higher CRC-association scores compared to healthy cohorts in each of the studies used

in our meta-analysis ($p < 0.05$, Mann–Whitney U test; Fig. 3A, Table S2). Notably, CRC-association score ranges varied by study, underscoring the importance of selecting controls from the same population.

Although our CAG-based model had been generated using only sequencing data from CRC or healthy cohorts, microbiomes associated with advanced adenomas (precancerous growths deemed to be high-risk based on size and/or histology; included in studies PRJEB7774 and PRJEB6070) had significantly higher CRC-association scores than healthy controls ($p = 0.04$, ANOVA), intermediate between healthy and CRC cohorts from the same populations (Fig. 3B).

We sought to further validate CRC-association scores in local cohorts of individuals diagnosed with CRC ($n = 13$), individuals with a history of advanced adenoma ($n = 11$), and healthy controls (stringently defined as having had 2 + colonoscopies and never having polyps; $n = 4$). We generated shotgun metagenomic sequencing data from their fecal samples ($6.8 \text{ M} \pm 2.9 \text{ k}$ reads per sample [mean $\pm$ SEM]). As predicted, CRC-association scores were significantly higher in CRC microbiomes than in the adenoma cohort ($p < 10^{-4}$, Student's two-tailed t -test) or in healthy individuals ($p < 10^{-5}$, Student's two-tailed t -test) (Fig. 3C). CRC-association scores in the adenoma cohort were intermediate and bimodally distributed, evidencing variation in microbiome-associated

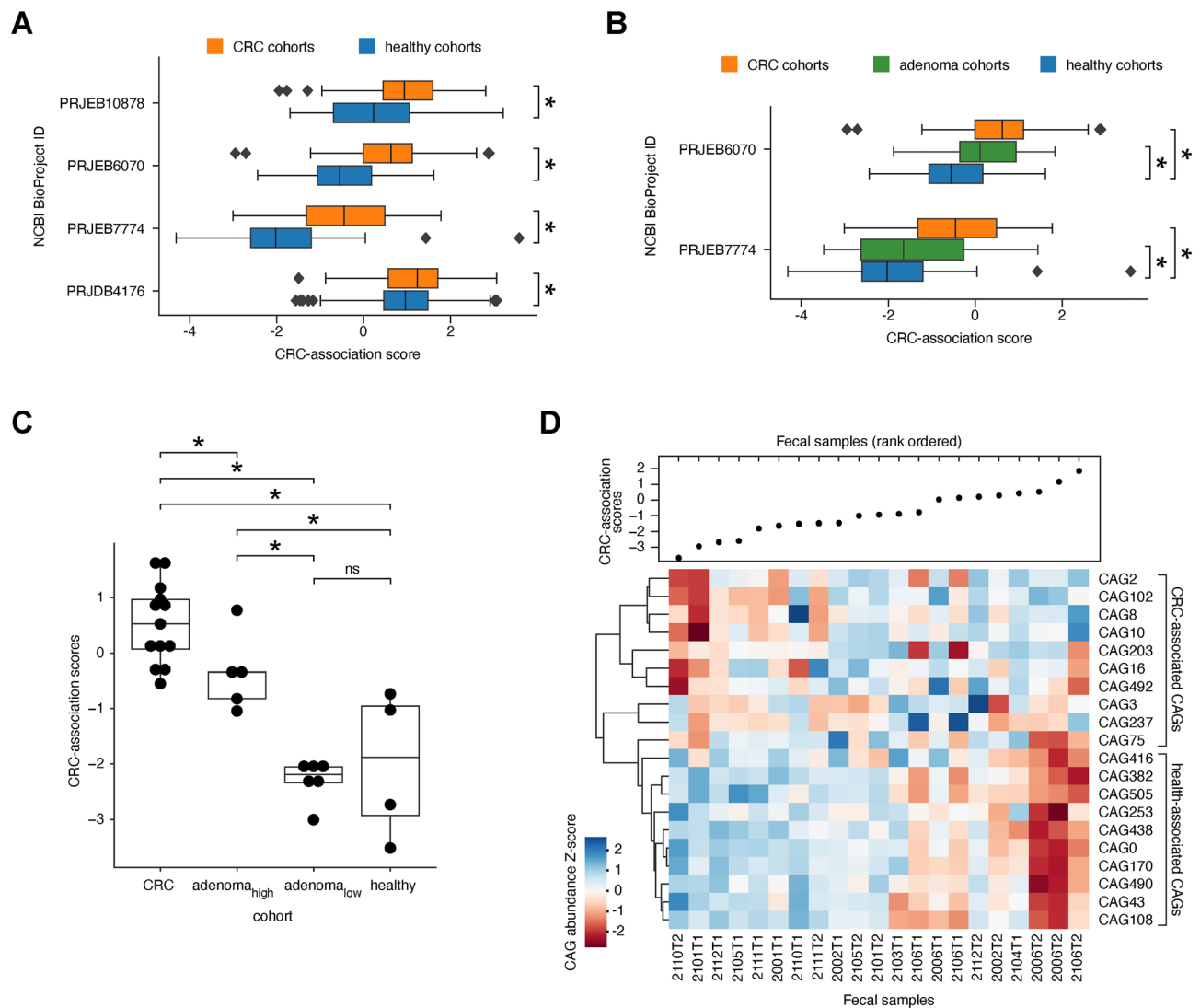


Fig. 3. Metagenome-based CRC-association scoring of individuals with CRC, adenoma, or history of advanced adenoma compared to healthy controls. **(A)** Significantly different CRC-association scores in CRC versus healthy cohorts in the datasets analyzed here. **(B)** CRC-association scores of fecal microbiomes from individuals with adenoma or CRC compared to healthy cohorts in two published studies. **(C)** Validation of CRC-association scores in local cohorts of individuals with CRC, history of advanced adenoma, or neither (i.e., healthy controls). A bimodal distribution of scores was seen in individuals with a history of advanced adenoma (adenoma_{high} and adenoma_{low}). In this plot, each individual is represented only once (first time point [T1] sample only) to avoid pseudoreplication in the statistical analysis. **(D)** Fecal microbiomes collected from individuals with history of advanced adenoma as a part of this study, rank-ordered by CRC-association scores (dot plot, top), and underlying abundances of top 10 CRC-associated CAGs and top 10 health-associated CAGs (heatmap, bottom).

CRC risk. The wide distribution of CRC-association scores corresponded to variability in the underlying CRC-or-health-associated CAGs (Fig. 3D).

We then assessed feasibility of calculating CRC-associated scores from residual stool in fecal immunochemical testing (FIT) kits, building upon preliminary reports of microbiome profiling in combination with FIT^{36,40,41}. To mimic real-world exposures, we prepared “mock FIT kits” from 4 fecal microbiomes from the adenoma cohort, subjected them to varying temperatures for various time intervals, generated shotgun metagenomic sequencing data, and calculated CRC-association scores. Fecal microbiota donor was the predominant and most significant factor determining CRC-association score, more so than time or temperature (microbiota donor: $p < 10^{-15}$, $F_{3,130} = 322$; time: $p < 0.01$, $F_{1,130} = 7$; temperature: $p = 0.2$, $F_{1,130} = 2$; three-way ANOVA; Fig. S6). These results suggest that fecal microbiome CRC-association scores may be robust to short-term real-world time and temperature exposures of stool samples.

CRC-association scores predict tumorigenicity

We next sought to prospectively demonstrate causality of microbiomes encoding CRC-associated CAGs in a preclinical model. We first performed an agnostic screen of a panel of 48 gut bacterial type strains that have undergone whole-genome sequencing. As in RefSeq genomes, we observed a distribution of CAG representation (Table S3) and CRC-association scores (Fig. 4A). Using these scores, we designed two consortia of bacteria not previously appreciated as cancer drivers (Figs. 4A, S7): our CRC-associated consortium was comprised of *Bacteroides fragilis* NCTC 9343 (i.e., non-toxicogenic *B. fragilis*), *Clostridium asparagiforme* DSM 15981, *Clostridium bolteae* ATCC BBA-613, *Clostridium symbiosum* DSM 934, and *Ruminococcus gnavus* ATCC 29149; and our health-associated consortium included *Bifidobacterium pseudocatenulatum* DSM 20438, *Coprococcus comes* ATCC 27758, *Dorea longicatena* DSM 13814, *Eubacterium rectale* ATCC 33656, and *Blautia obeum* ATCC 29174.

We colonized germ-free (GF) mice harboring a mutation in *Apc*, the tumor suppressor gene commonly mutated in CRCs and adenomas⁴², with either the CRC-associated or health-associated consortium ($n = 6-9$ /cohort). *Apc*^{Min/+} mice are genetically predisposed to developing intestinal tumors, but when rederived as GF, they have almost no tumor burden²¹ and therefore serve as a useful baseline for testing tumorigenic properties of microbes. We also used *Il10*-deficient *Apc*^{Min/+} mice as a second CRC model²¹ to test robustness. Metagenomic sequencing data generated from fecal pellets collected from these gnotobiotic mice confirmed that the mouse microbiomes retained the CRC-or-health-associated genomic signatures of the input communities ($p < 0.007$, two-tailed Student's *t*-test; Fig. 4B). While *Il10*-deficient mice had reduced tumor burden overall, the CRC-associated bacterial consortium induced significantly greater tumor burden in gnotobiotic mice compared to

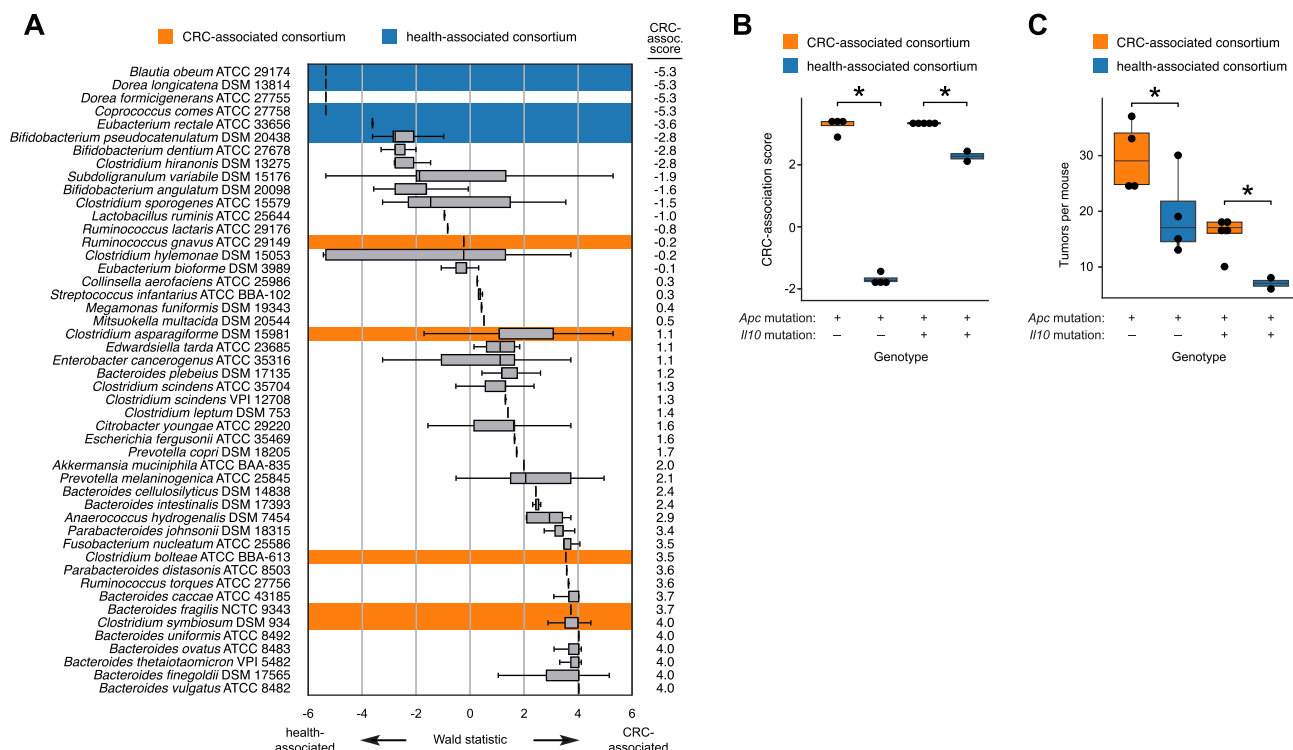


Fig. 4. CRC-association scores predict microbiome-induced tumorigenicity. **(A)** CRC-association scores and underlying Wald metrics of genome-aligned CAGs for a panel of bacterial strains. Consortium members highlighted. **(B)** CRC-or-health signatures seen in metagenomic sequencing of fecal pellets collected from gnotobiotic mice colonized with CRC-associated or health-associated consortia. **(C)** Small intestinal tumor burden in gnotobiotic mice colonized with CRC-associated consortium or health-associated consortium.

the health-associated consortium (microbiota: $p = 0.03$, $F_{1,168} = 5.9$; *Il10* genotype: $p < 10^{-3}$, $F_{1,630} = 22$; two-way ANOVA; Fig. 4C). Tumor size did not significantly differ between cohorts.

Histologic analysis revealed minimal surrounding inflammation observed (Fig. S8), with no significant difference in inflammation scores between treatment groups (both CRC consortium-colonized and health consortium-colonized mouse cohorts had a median score of 1, with range 0–2 out of maximum 5). Tumors were precancerous adenomas (typical in *Apc^{Min/+}* mice). These results suggest that members of our consortia transmit predicted tumor phenotypes, providing in vivo validation of in silico predictions. However, it should be noted that small intestinal (not colonic) tumor burden was the readout used here, a limitation necessitated by toxicity of colonic tumor-inducing dextran sodium sulfate (DSS) in the gnotobiotic setting, as discussed in *Supplemental Text*.

Microbiome-driven induction of senescence

We assessed direct growth-promoting effects of these bacterial consortia by culturing Caco-2 cells in media containing fecal filtrates from gnotobiotic mice colonized with either the CRC-associated or health-associated consortium²⁸. No significant difference was observed in Caco-2 cell viability between consortia, suggesting that there were no growth-promoting products secreted by the CRC-associated consortium that could explain our tumor phenotype (Fig. S9).

We therefore asked whether the CRC-associated consortium indirectly induced tumorigenesis. Field cancerization is the phenomenon in which visually and histologically normal tissue acquires molecular alterations that predispose it to cancer. In order to identify colonic field effects attributable to the CRC-associated consortium while minimizing detection of artifactual late-in-life responses to microbial colonization of GF mice⁴³, we utilized specific pathogen-free (SPF) *Apc^{Min/+}* mice, which we treated with antibiotics at 6–8 weeks of age, recolonized with CRC-associated or health-associated consortia, and euthanized 9–10 weeks later ($n = 8$ /cohort). We then assessed field effects in healthy-appearing colonic tissues. Mice colonized with these consortia harbored distinct colonic transcriptomic signatures ($p < 0.02$, PERMANOVA; Table S4). The CRC-associated consortium was associated with higher expression of *Cdkn2a* (which encodes p16^{INK4a}), *Gdf15*, *E2f2*, and *Ifng*, which are hallmark features of senescence (Fig. 5A), an aging phenomenon linked to CRC²⁸. Genes in epithelial-to-mesenchymal transition (*Cd44*, *Ptgs2*), DNA damage repair (*Pclaf*), the NF- κ B pathway (*Traf1*, *Ptgs2*, *Lat*), and other inflammatory pathways (*Cd44*, *Ifng*, *Pdcd1lg2* [which encodes PD-L2]) were also upregulated in mice colonized with the CRC-associated consortium (Figs. 5A, S10). Colonic tissues were stained for p21 to identify senescent cells, and while positively staining cells were detectable, overall levels were too low to reliably quantify and compare.

To investigate whether the microbiome can induce senescence, we first reanalyzed published single-cell RNA sequencing (scRNAseq) datasets—reasoning that microbial induction of senescence may be specific to discrete cell populations—from studies comparing colonic cells in GF and conventional SPF mice^{44,45}, reasoning that such a comparison would provide us insights into generalizable microbiome effects pertinent to CRC. To classify cells as senescent or not, we utilized a recently validated, gene expression-based method for defining senescence in which cells are scored on the basis of expression of 125 senescence-associated genes (“SenMayo panel”), with the top 10% of cells classified as senescence-associated secretory phenotype (SASP) cells³³. In our reanalysis of raw FASTQ single-cell sequencing data published in a study of colonic myeloid cells⁴⁴, we found that a greater proportion of cells from SPF mice were senescent compared to GF mice (12.1% vs 7.9%), and SenMayo enrichment scores were significantly higher in SPF mice (1054 ± 3 vs 1016 ± 3 , $p < 10^{-15}$, Student’s two-tailed t -test). Senescent cells were predominantly concentrated in stromal cells and macrophages, with 35% and 15% of each population exhibiting a senescent gene expression phenotype, respectively. This trend was reflected in a separate study⁴⁵ in which 78% of macrophages, 56% of monocytes, and 52% of neutrophils were senescent.

To assess reproducibility of these findings, we generated scRNAseq data from full-thickness healthy-appearing colonic tissues harvested from (i) age-matched SPF *Apc^{Min/+}* and wild-type mice, half of which had received antibiotics for microbiome depletion ($n = 4$; $3,800 \pm 733$ analyzed cells/sample, with 9927 ± 512 high-quality read pairs/cell); and (ii) mice colonized with either the CRC-associated or health-associated consortium ($n = 3$ – 4 /treatment group; 466 ± 136 analyzed cells/sample, with 6267 ± 804 high-quality read pairs/cell). After dissociating cells from the harvested colons of these mice, gene expression at single-cell resolution was assayed with the $10 \times$ Chromium single-cell sequencing platform. After clustering and classifying cells (Fig. 5B), we found that the cellular populations with the highest proportions of senescent cells were stromal cells (76%), neutrophils (49%), and macrophages (46%) (Fig. 5C). These cell types also had the highest SenMayo enrichment scores of 3781 ± 13 , 3455 ± 15 , and 3417 ± 26 , respectively, which was significantly greater than any other population of cells ($p < 10^{-15}$, Student’s one-tailed t -test; Fig. 5D). We also observed a moderately high distribution of SenMayo enrichment scores in endothelial cells (3064 ± 12) as well as mildly elevated scores in T cells (2407 ± 7) and EpCAM(+) cells (2337 ± 4) (Fig. 5C,D). Interestingly, the high SenMayo enrichment scores of the different senescent cell populations were attributable to expression of largely non-overlapping subsets of the SenMayo panel (Fig. 5E), evidencing heterogeneity in gene expression programs underlying senescence.

Mice harboring a complete microbiome had significantly higher SenMayo enrichment scores overall ($p < 0.01$, Student’s one-tailed t -test) and higher proportions of senescent cells (14.4% vs 9.2%) compared to mice that had undergone antibiotic-mediated microbiome depletion. This effect was largest in stromal cells, of which 79% vs 68% were classified as senescent in microbiome-harboring vs microbiome-depleted mice, respectively.

Mice colonized with the CRC-associated consortium had a greater proportion of senescent cells compared to mice harboring the health-associated consortium (10.4% vs 6.9%). SenMayo enrichment scores were significantly higher in stromal cells (3834 ± 21 vs 3753 ± 36 , $p < 0.03$, Student’s one-tailed t -test) and macrophages (3498 ± 32 vs 3186 ± 67 , $p < 10^{-4}$, Student’s one-tailed t -test), suggesting that there is a link between the microbiome, senescence, and tumorigenesis.

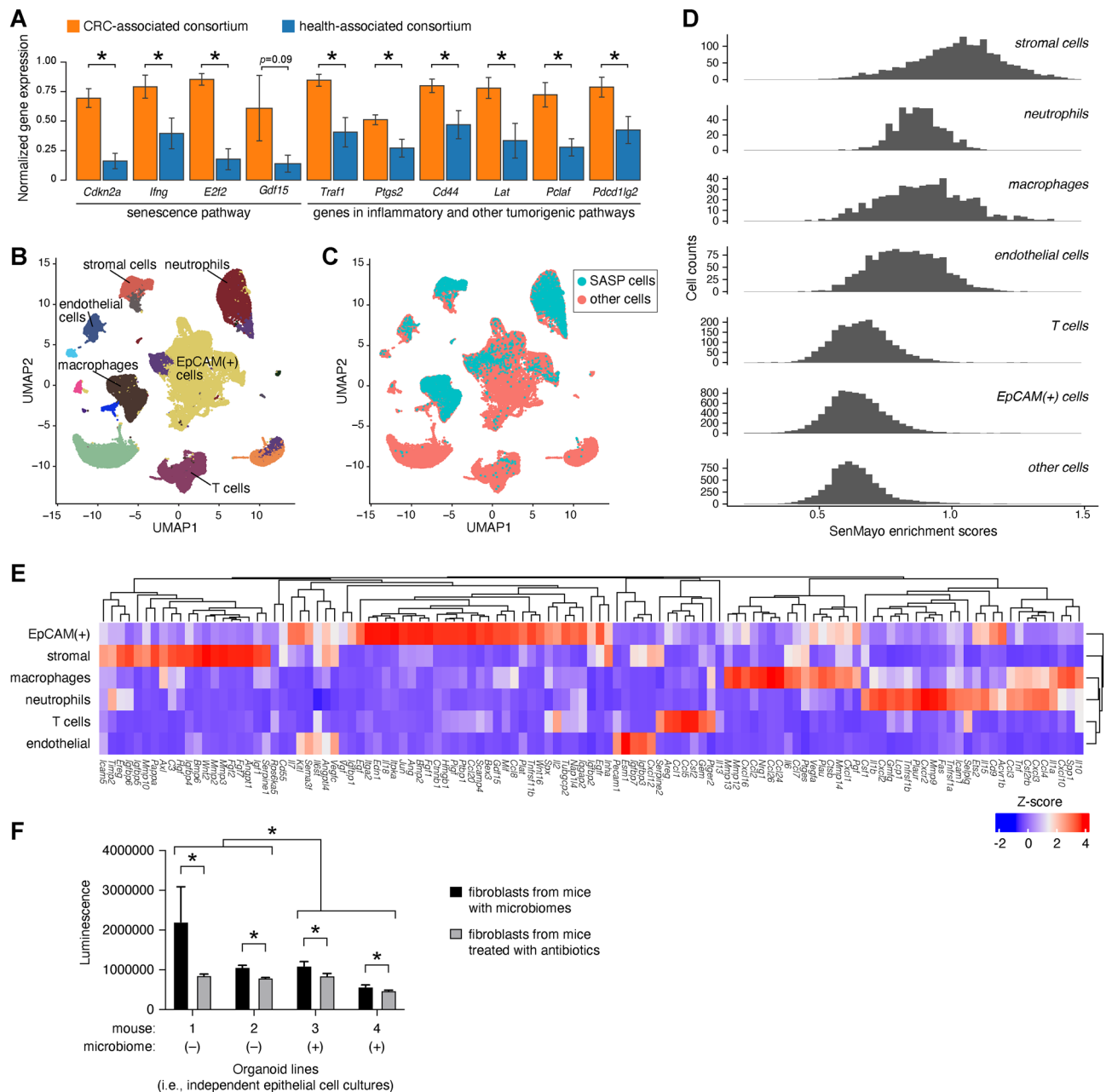


Fig. 5. The microbiome induces senescence in specific colonic cell populations, which then drives epithelial cell growth phenotypes. **(A)** Differentially expressed genes in healthy-appearing non-tumor colonic tissue harvested from mice colonized with a CRC-associated or health-associated consortium. **(B–C)** Single-cell RNAseq of healthy-appearing colons reveals discrete cell clusters of SASP cells, as shown in these UMAP plots. **(D)** Distribution of SenMayo enrichment scores and **(E)** expression of different subsets of SenMayo genes in different cell populations of the colon. **(F)** Microbiome conditioning of fibroblasts drives increased epithelial cell growth in an in vitro organoid model.

These findings were seen in both *Apc*^{Min/+} and wild-type mice, evidencing robustness to genotype. Overall, SenMayo enrichment scores were significantly higher in *Apc*^{Min/+} mice than in wild-type mice ($p < 10^{-15}$, Student's one-tailed *t*-test), and this was attributable to significant differences in multiple cell type-specific comparisons (stromal cells: 3769 ± 21 vs 3655 ± 33 , $p < 0.002$; endothelial cells: 3271 ± 32 vs 3110 ± 36 , $p < 0.0005$; T cells: 2503 ± 22 vs 2399 ± 17 , $p < 0.0002$; EpCAM(+) cells: 2427 ± 9 vs 2199 ± 11 , $p < 10^{-15}$; Student's one-tailed *t*-test).

Finally, we tested the hypothesis that microbiome-induced senescent tissue microenvironments can promote growth. In our scRNAseq analysis, stromal cells (which include fibroblasts) featured prominently, and we previously reported that senescent fibroblasts create a tumor-promoting tissue microenvironment in the colon²⁸. Thus, we set up an ex vivo co-culture system composed of primary colon fibroblasts and 3D colonic organoids derived either from unmanipulated SPF mice or littermates that had been treated with antibiotics, testing different combinations of fibroblasts and epithelial cells (4 organoid lines: 2 generated from microbiome-harboring mice and 2

from microbiome-depleted mice; 6 replicates/line). 3D organoid proliferation, as quantified by luminescence, was significantly higher in organoid co-cultures in which fibroblasts had been isolated from microbiome-harboring mice (Fig. 5F; 2-way ANOVA in which luminescence was the dependent variable: fibroblast source: $p < 0.0002$; epithelial cell source: $p < 0.0003$; interaction between fibroblast and epithelial cell sources: $p < 0.04$, $F_{1,45} = 4.48$). These results were consistent across all four independent organoid lines, evidencing reproducibility. Bulk gene expression signatures of fibroblasts sourced from mice harboring microbiomes confirmed findings from our single-cell analysis. Our findings suggest that the microbiome does indeed condition fibroblasts to promote epithelial cell growth.

Discussion

In summary, our findings suggest that upwards of an astounding ~40% of gut bacteria may be associated with CRC or with health (specifically, a state of not having CRC or precancerous changes). In our comparison with standard taxonomic analytic approaches, we found a greater number of associations at a higher level of taxonomic resolution, evidencing the potentially greater sensitivity of a CAG-based approach.

Many of the microbes we identified had not previously been linked to CRC. While only a subset of these associated microbes are expected to be causative, our findings nonetheless expand the scope of “known unknowns” of CRC-associated microbes. We validate our CAG-based model in preclinical models, in which we observed differential induction of precancerous field effects, namely senescence, that was attributable to the microbiome. Through reanalysis of published and newly generated single-cell RNA sequencing from preclinical models combined with an organoid-based co-culture model, we connect senescence to the gut microbiome at large. Our findings suggest generalizability of a prior report linking two *Porphyromonas* species to gut senescence⁴⁶. Senolytics can reduce colonic tumorigenesis in a preclinical model⁴⁷; by extension, targeting microbiome-induced senescence may be a viable CRC prevention strategy. Implications of our findings may be broad, as senescent cells, also known as ‘zombie cells’⁴⁸, are implicated in other disease contexts as well.

Our results serve as proof-of-principle that CAG-based analysis may offer clinically pertinent insights into microbiome-CRC relationships and discovery of cancer prevention strategies. One limitation of this study is that a comprehensive investigation of all CRC-associated CAGs and health-associated CAGs was not feasible. Instead, we rely upon an agnostic screen of a panel of bacteria to design experimental consortia that we compare in preclinical studies. This line of investigation led us to identify senescence as a microbiome-inducible precancer state, and fibroblasts as microbiome-responsive host cells. However, fibroblasts are just one type of stromal cell, and we identified several other cell types that likely experience senescence in microbiome-dependent ways (neutrophils, macrophages) that we were unable to test. Which specific microbes are strongest inducers of senescence, how those effects vary by cell type, and the significance of cell type-specific senescence gene expression responses to the microbiome remain unclear. Finally, senescence is unlikely to be the only mechanism through which microbiomes shape CRC risk. A comprehensive investigation and catalog of CRC-or-health-associated CAGs and their potentially carcinogenic or cancer-preventing effects on the colonic microenvironment is warranted. For example, given *F. prausnitzii*’s longstanding reputation as a butyrate-producing, anti-inflammatory, health-promoting bacterium, we found it striking that *F. prausnitzii*-encoded CAGs had wide-ranging associations spanning from strongly CRC-linked to strongly health-linked, an observation that evidences the biological importance of subspecies groups (e.g. the role of *F. prausnitzii* subspecies in atopic dermatitis⁴⁹). Strain-level variability in the roles of *F. prausnitzii* and other CRC-associated species in driving precancer states and CRC remains to be defined. Ultimately, targeted detection of CRC-associated CAGs may enhance CRC screening (perhaps leveraging population-specific CAG-based models) and the development of microbiome-based CRC prevention tools.

Nonetheless, our study and other related studies in this emerging area of research present an opportunity to address a public health gap: the need for improved non-invasive CRC screening. The microbiome may enable detection of precancerous lesions, which would be great value in terms of cancer prevention, as demonstrated through our experience with colonoscopy. The finding that microbiome-induced precancerous changes may precede adenoma development and may not be visually perceptible opens up prospects for even earlier screening. One explanation for the apparently significant effects of our microbiomes on our health is that our resident microbes are always with us, and their biological effects, even if modest in size, may accumulate over a lifetime of exposure. This ‘area under the curve’ of microbiome-induced effects may shape risk of developing CRC and other cancers, a prospect that is both daunting and brimming with opportunity.

Data availability

Bacterial shotgun sequencing datasets and single-cell sequencing datasets have been deposited in the NCBI Sequence Read Archives (PRJNA1157403 and PRJNA1158762, respectively).

Received: 1 February 2024; Accepted: 20 August 2024

Published online: 09 October 2024

References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Burkitt, D. P. Epidemiology of cancer of the colon and rectum. *Cancer* **28**, 3–13 (1971).
3. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679 (2019).
4. Wu, S. *et al.* A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat. Med.* **15**, 1016–1022 (2009).

5. Bonnet, M. *et al.* Colonization of the human gut by *E. coli* and colorectal cancer risk. *Clin. Cancer Res.* **20**, 859–867 (2014).
6. Yang, Y. *et al.* *Fusobacterium nucleatum* Increases proliferation of colorectal cancer cells and tumor development in mice by activating toll-like receptor 4 signaling to nuclear factor- κ B, and up-regulating expression of microRNA-21. *Gastroenterology* **152**, 851–866.e24 (2017).
7. Haghi, F. *et al.* The association between fecal enterotoxigenic *B. fragilis* with colorectal cancer. *BMC Cancer* **19**, 879 (2019).
8. Coker, O. O. *et al.* Altered gut metabolites and microbiota interactions are implicated in colorectal carcinogenesis and can be non-invasive diagnostic biomarkers. *Microbiome* **10**, 35 (2022).
9. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
10. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
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. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **5**(6), 968–976 (2019).
13. Ma, Z. S. Testing the Anna Karenina principle in human microbiome-associated diseases. *iScience* **23**, 101007 (2020).
14. Dey, N. *et al.* Association of gut microbiota with post-operative clinical course in Crohn's disease. *BMC Gastroenterol.* **13**, 131 (2013).
15. Tolstoy L. *Anna Karenina*. Penguin Classics; (1877).
16. Arthur, J. C. *et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120–123 (2012).
17. Minot, S. S. & Willis, A. D. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome* **7**, 110 (2019).
18. Minot, S. S. *et al.* geneshot: gene-level metagenomics identifies genome islands associated with immunotherapy response. *Genome Biol.* **22**, 135 (2021).
19. MetaHIT Consortium *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
20. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Stat.* **14**(1) 94–115 (2020).
21. Tomkovich, S. *et al.* Locoregional Effects of Microbiota in a Preclinical Model of Colon Carcinogenesis. *Cancer Res* **77**, 2620–2632 (2017).
22. Li, N. *et al.* Microbiome-encoded bile acid metabolism modulates colonic transit times. *iScience* **24**, 102508 (2021).
23. Dey, N. *et al.* Regulators of gut motility revealed by a gnotobiotic model of diet-microbiome interactions related to travel. *Cell* **163**, 95–107 (2015).
24. Erben, U. *et al.* A guide to histomorphological evaluation of intestinal inflammation in mouse models. *Int. J. Clin. Exp. Pathol.* **7**, 4557–4576 (2014).
25. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
26. Edgar, R. C. & Flyvbjerg, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**, 3476–3482 (2015).
27. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
28. Guo, Y. *et al.* Senescence-associated tissue microenvironment promotes colon cancer formation through the secretory factor GDF15. *Aging Cell* **18**, e13013 (2019).
29. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
30. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A* **102**, 15545–15550 (2005).
31. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* <https://doi.org/10.1101/060012> (2016).
32. Liberzon, A. *et al.* The Molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
33. Saul, D. *et al.* A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. *Nat. Commun.* **13**, 4827 (2022).
34. O'Rourke, K. P. *et al.* Isolation, culture, and maintenance of mouse intestinal stem cells. *Bio-Protocol* **6**, e1733 (2016).
35. Khalil, H. *et al.* Isolation of primary myofibroblasts from mouse and human colon tissue. *J Vis Exp* **80**, e50611 (2013).
36. Goedert, J. J. *et al.* Reusing a prepaid health plan's fecal immunochemical tests for microbiome associations with colorectal adenoma. *Sci Rep* **12**, 14801 (2022).
37. Tierney, B. T. *et al.* The landscape of genetic content in the gut and oral human microbiome. *Cell Host Microbe* **26**, 283–295.e8 (2019).
38. Tortora, S. C. *et al.* Microbiome and colorectal carcinogenesis: Linked mechanisms and racial differences. *World J. Gastrointest. Oncol.* **14**, 375–395 (2022).
39. Yuan, L. *et al.* The influence of gut microbiota dysbiosis to the efficacy of 5-Fluorouracil treatment on colorectal cancer. *Biomed. Pharmacother. Biomed. Pharmacother.* **108**, 184–193 (2018).
40. Baxter, N. T. *et al.* DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. *Microbiome* **4**, 59 (2016).
41. Krigul, K. L. *et al.* Using fecal immunochemical tubes for the analysis of the gut microbiome has the potential to improve colorectal cancer screening. *Sci. Rep.* **11**, 19603 (2021).
42. Powell, S. M. *et al.* APC mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235–237 (1992).
43. Camp, J. G. *et al.* Microbiota modulate transcription in the intestinal epithelium without remodeling the accessible chromatin landscape. *Genome Res.* **24**, 1504–1516 (2014).
44. Kang, B. *et al.* Commensal microbiota drive the functional diversification of colon macrophages. *Mucosal. Immunol.* **13**, 216–229 (2020).
45. Širvinskis, D. *et al.* Single-cell atlas of the aging mouse colon. *iScience* **25**, 104202 (2022).
46. Okumura, S. *et al.* Gut bacteria identified in colorectal cancer patients promote tumorigenesis via butyrate secretion. *Nat. Commun.* **12**, 5674 (2021).
47. Parihar, M. *et al.* Rapamycin extends life span in Apc colon cancer FAP model. *Clin. Colorectal Cancer* **20**, e61–e70 (2021).
48. Smith F. 'Zombie cells' could hold the secret to Alzheimer's cure. Natl Geogr 2023. Available at: <https://www.nationalgeographic.com/magazine/science/article/zombie-cells-could-hold-the-secret-to-alzheimers-cure>.
49. Song, H. *et al.* Faecalibacterium prausnitzii subspecies-level dysbiosis in the human gut microbiome underlying atopic dermatitis. *J. Allergy Clin. Immunol.* **137**, 852–860 (2016).

Acknowledgements

We would like to acknowledge Feinan Wu and Matt Fitzgibbon from the Genomics & Bioinformatics Core at Fred Hutch for technical guidance.

Disclosures

S.S.M. and N.D. are named inventors on a pending US non-provisional patent application (18/320,878: Bacterial Gene-Associated Methods and Compositions for Diagnosing and Treating Colorectal Cancer). All other authors declare no competing interests.

Author contributions

Conceptualization: S.S.M., N.D. Methodology: S.S.M., N.L., J.L.A., M.Y., S.T.K., J.A.D., W.M.G., N.D. Investigation: S.S.M., N.L., J.L.A., M.Y., S.T.K., M.M.S., R.B.H., N.D. Analysis: S.S.M., H.S., S.T.K., N.D. Visualization: S.S.M., N.L., H.S., S.T.K., M.M.S., N.D. Funding acquisition: W.M.G., N.D., R.B.H., M.Y. Project administration: N.D. Supervision: N.D. Writing—original draft: S.S.M., N.D. Writing—review & editing: S.S.M., N.D.

Funding

The study was funded by National Institutes of Health K08 DK111941 (N.D.), National Institutes of Health U54 CA274374 (N.D., W.M.G., R.B.H.), National Institutes of Health R50CA233042 (M.Y.), Microbiome Research Initiative at Fred Hutchinson Cancer Center (S.S.M., N.D.), and the Cancer Consortium (P30 CA015704) at Fred Hutchinson Cancer Center and University of Washington (N.D., Genomics & Bioinformatics Core at Fred Hutch).

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-70702-1>.

Correspondence and requests for materials should be addressed to N.D.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024