LETTERS
https://doi.org/10.1038/s41591-019-0458-7
nature medicine

Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer

Shinichi Yachida[1,2,17*], Sayaka Mizutani[3,17], Hirotsugu Shiroma[3], Satoshi Shiba[2], Takeshi Nakajima[4], Taku Sakamoto[4], Hikaru Watanabe[3], Keigo Masuda[3], Yuichiro Nishimoto[3], Masaru Kubo[3], Fumie Hosoda[2], Hirofumi Rokutan[2], Minori Matsumoto[4], Hiroyuki Takamaru[4], Masayoshi Yamada[4], Takahisa Matsuda[4], Motoki Iwasaki[5], Taiki Yamaji[5], Tatsuo Yachida[6], Tomoyoshi Soga[7], Ken Kurokawa[8], Atsushi Toyoda[9], Yoshitoshi Ogura[10], Tetsuya Hayashi[10], Masanori Hatakeyama[11], Hitoshi Nakagama[12], Yutaka Saito[4], Shinji Fukuda[7,13,14,15], Tatsuhiro Shibata[2,16] and Takuji Yamada[3,15*]

In most cases of sporadic colorectal cancers, tumorigenesis is a multistep process, involving genomic alterations in parallel with morphologic changes. In addition, accumulating evidence suggests that the human gut microbiome is linked to the development of colorectal cancer. Here we performed fecal metagenomic and metabolomic studies on samples from a large cohort of 616 participants who underwent colonoscopy to assess taxonomic and functional characteristics of gut microbiota and metabolites. Microbiome and metabolome shifts were apparent in cases of multiple polypoid adenomas and intramucosal carcinomas, in addition to more advanced lesions. We found two distinct patterns of microbiome elevations. First, the relative abundance of *Fusobacterium nucleatum* spp. was significantly ($P < 0.005$) elevated continuously from intramucosal carcinoma to more advanced stages. Second, *Atopobium parvulum* and *Actinomyces odontolyticus*, which co-occurred in intramucosal carcinomas, were significantly ($P < 0.005$) increased only in multiple polypoid adenomas and/or intramucosal carcinomas. Metabolome analyses showed that branched-chain amino acids and phenylalanine were significantly ($P < 0.005$) increased in intramucosal carcinomas and bile acids, including deoxycholate, were significantly ($P < 0.005$) elevated in multiple polypoid adenomas and/or intramucosal carcinomas. We identified metagenomic and metabolomic markers to discriminate cases of intramucosal carcinoma from the healthy controls. Our large-cohort multi-omics data indicate that shifts in the microbiome and metabolome occur from the very early stages of the development of colorectal cancer, which is of possible etiological and diagnostic importance.

Colorectal cancer (CRC) affects over a quarter of a million people each year worldwide[1]. Most sporadic CRCs develop through the formation of polypoid adenomas and are preceded by intramucosal carcinoma (high-grade dysplastic adenoma), which can progress into malignant forms[2]. This process is known as the adenoma–carcinoma sequence, which occurs through a multistep mechanism that involves specific mutations[2]. Because it takes decades before final malignancies develop[3], detection of early cancers and their endoscopic removal are priorities for cancer control.

Alterations in the gut ecosystem have been implicated in changes in human health and disease, including CRC[4]. Fecal metagenomics is a useful tool for the quantification of the gut microbiome and has possible diagnostic potential[5]. In particular, *F. nucleatum* has been associated with CRC[6,7]; its mechanism is being clarified in mice[8]. Intestinal metabolites, including amino acids and bacterial metabolites (for example, bile acids and short-chain fatty acids), have also been associated with cancerous conditions in the gut[9,10]. Comprehensive metagenomic and metabolomic analyses might therefore provide an alternative approach to understanding CRC development through associated changes in the gut environment.

We have performed whole-genome shotgun metagenomics and capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS)-based metabolomics on fecal samples obtained from patients with different stages of colorectal neoplasia to obtain evidence of distinct stage-specific phenotypes of fecal microorganisms and metabolites.

Metagenomic and metabolomic data were collected from 616 and 406 subjects, respectively (Supplementary Table 1 and Extended Data Fig. 1). Classification into nine groups was carried

[1]Department of Cancer Genome Informatics, Graduate School of Medicine, Osaka University, Osaka, Japan. [2]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan. [3]School of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan. [4]Endoscopy Division, National Cancer Center Hospital, Tokyo, Japan. [5]Epidemiology and Prevention Group, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan. [6]Department of Gastroenterology and Neurology, Faculty of Medicine, Kagawa University, Kagawa, Japan. [7]Institute for Advanced Biosciences, Keio University, Yamagata, Japan. [8]Genome Evolution Laboratory, National Institute of Genetics, Shizuoka, Japan. [9]Comparative Genomics Laboratory, National Institute of Genetics, Shizuoka, Japan. [10]Department of Bacteriology, Faculty of Medical Sciences, Kyushu University, Fukuoka, Japan. [11]Department of Microbiology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [12]National Cancer Center, Tokyo, Japan. [13]Intestinal Microbiota Project, Kanagawa Institute of Industrial Science and Technology, Kanagawa, Japan. [14]Transborder Medical Research Center, University of Tsukuba, Ibaraki, Japan. [15]PRESTO, Japan Science and Technology Agency, Saitama, Japan. [16]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [17]These authors contributed equally: Shinichi Yachida, Sayaka Mizutani. *e-mail: syachida@cgi.med.osaka-u.ac.jp; takuji@bio.titech.ac.jp

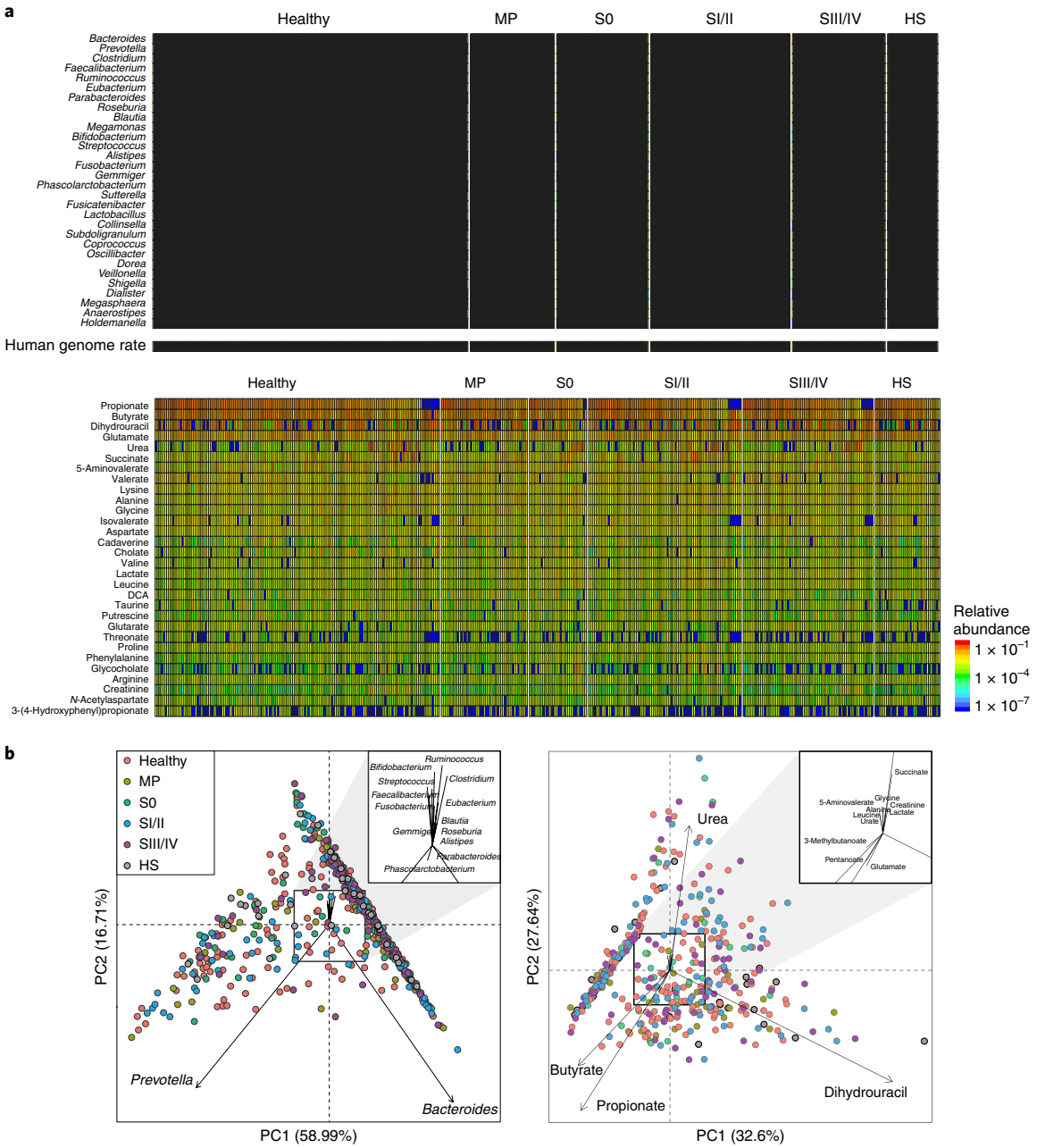


Fig. 1 | Global metagenomic and metabolomic characteristics of the fecal samples. a, Relative abundances of the top 30 genera (top) and fractions of the human genome (middle) for 616 subjects in the healthy control (normal and a few polyps) ($n=251$), MP ($n=67$), S0 ($n=73$), SI/II ($n=111$), SIII/IV ($n=74$) and HS ($n=40$) groups. Percentage concentrations of metabolites for 406 subjects in the healthy control ($n=149$), MP ($n=45$), S0 ($n=30$), SI/II ($n=80$), SIII/IV ($n=68$) and HS ($n=34$) groups (bottom). **b**, PCA of genera ($n=616$) (left) and metabolites ($n=406$) (right). PC, principal component.

out according to colonoscopic and histological findings: (1) normal (no remarkable colonoscopic findings); (2) a few polyps (up to two small (<5 mm) polyps); (3) multiple polypoid adenomas with low-grade dysplasia (MP, more than three adenomas, mostly more than five adenomas); (4) intramucosal carcinoma (polypoid adenoma(s) with high-grade dysplasia), stage 0/pTis CRC (S0); (5) stage I CRC; (6) stage II CRC; (7) stage III CRC; (8) stage IV CRC based on the eighth Union for International Cancer Control (UICC) TNM Classification of Malignant Tumors. The remaining group was (9) normal with a history of colorectal surgery (HS). The polypoid adenomas in MP were limited to pathologically proven conventional-type colorectal adenomas (that is, tubular adenoma, tubulovillous adenoma and villous adenoma) with low-grade dysplasia, not including serrated adenomas. Subjects in groups (1) and (2) were defined as healthy controls. Those with stage I and II CRCs were combined into stage I/II (SI/II), and stage III and IV CRCs were combined into stage III/IV (SIII/IV) for analysis. Clinical characteristics for the groups are shown in Supplementary Tables 1, 2 and Extended Data Fig. 2. Smoking history (Brinkman's index) differed among the groups; patients with more advanced stages tended to have a lower Brinkman's index (that is, less smoking) compared to patients with early stages, and patients with HS had even lower values. Metagenome data were obtained both before and after surgery from 28 patients with stage I–III CRCs (Supplementary Tables 1).

A broad overview of our taxonomic data from 616 subjects and metabolomic data from 406 subjects is given in Fig. 1. Subjects

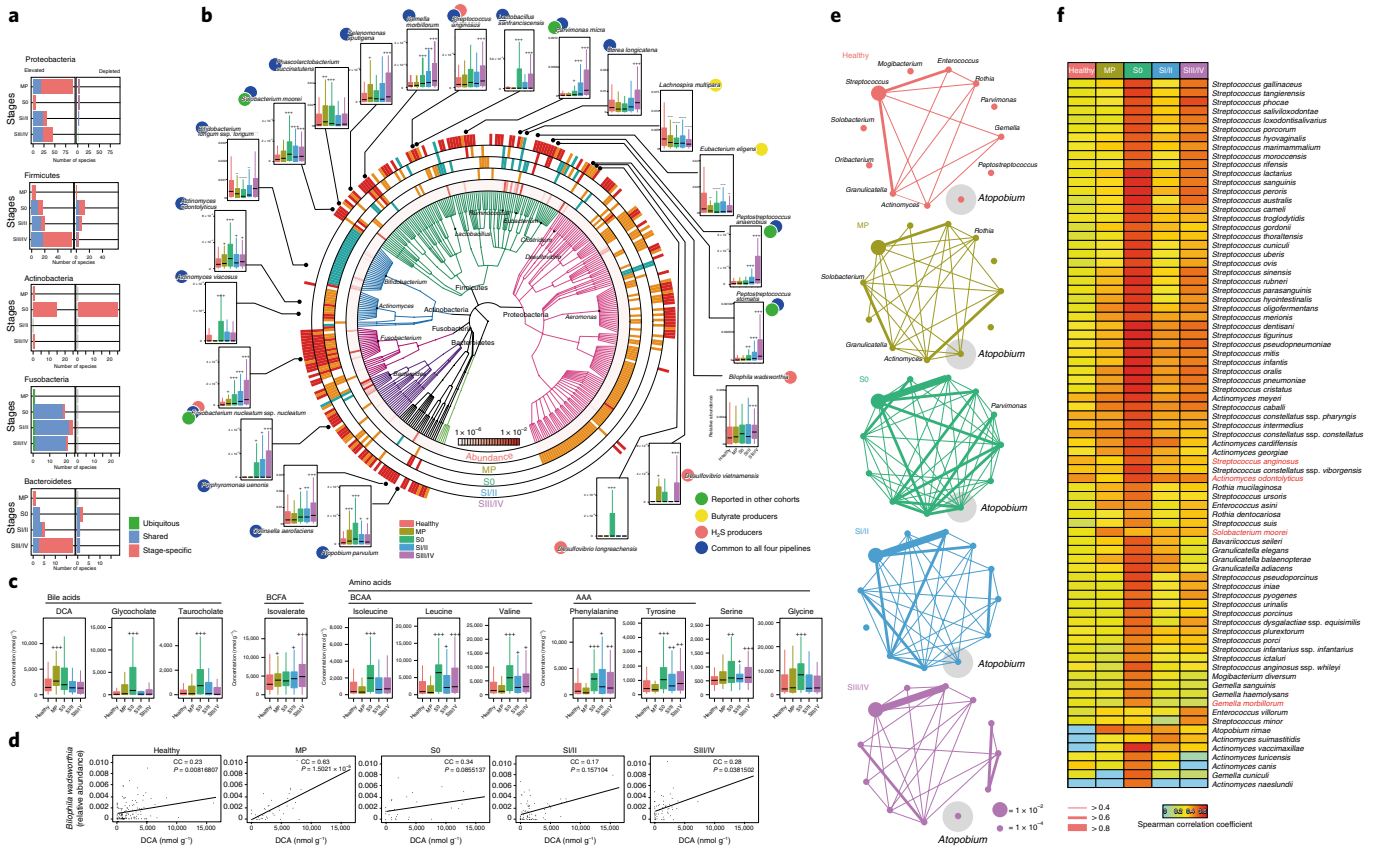


Fig. 2 | Distinct stage-specific taxonomic and metabolomic signatures with cancer progression. a,**b**, Species abundances were assessed for significant elevation or depletion ($P<0.005$; one-sided Mann–Whitney U test) in each of the four stages, MP ($n=67$), S0 ($n=73$), SI/II ($n=111$) and SIII/IV ($n=74$), compared to the healthy controls ($n=251$). **a**, Phylum distribution of the number of species that are either elevated or depleted in each of the four stages compared to the healthy controls. For each of the phyla (Proteobacteria, Bacteroidetes, Fusobacteria, Firmicutes and Actinobacteria), the number of species, which were either significantly elevated or depleted in each of the four stages compared to the healthy controls, was counted. The number observed with change over all stages (ubiquitous, green) was very small; most species demonstrated alterations that are specific to one stage (stage-specific, red) or shared with another stage(s) (shared, blue). **b**, In total, 361 differentially abundant species are shown in a phylogenetic tree, grouped in the phyla Proteobacteria, Bacteroidetes, Fusobacteria, Firmicutes and Actinobacteria. In the outer circles, species are marked for significant ($P<0.005$) elevation (orange) or depletion (green). Particularly highly elevated (false-discovery rate-corrected $P<0.1$; one-sided Mann–Whitney U test) species are marked in red. The innermost circle shows species relative abundances averaged over all samples. Species for which the mean relative abundances were lower than 1×10^{-6} were excluded from the phylogenetic tree. Box plots show the relative abundances of cancer-related species reported in other cohorts (green circles), butyrate producers (yellow circles), hydrogen sulfide producers (pink circles) or newly detected in all of four metagenomic pipelines (see Methods) with significant ($P<0.005$) changes compared to the healthy controls (blue circles). The y axis for each box plot is relative abundance. Each plot shows boxes in the healthy controls (left-most bar) and the four groups (MP, S0, SI/II and SIII/IV) in order of left to right. **c**, On the basis of the metabolome analysis, fecal concentrations of bile acids, branched-chain fatty acid (BCFA) and amino acids showed significant elevation or depletion ($P<0.005$; one-sided Mann–Whitney U test) in each of the four stages, MP ($n=45$), S0 ($n=30$), SI/II ($n=80$) and SIII/IV ($n=68$) compared j to the healthy controls ($n=149$). BCAA, branched-chain amino acids; AAA, aromatic amino acids. **d**, The relationship between *B. wadsworthia* and DCA. *B. wadsworthia* showed the highest Spearman's correlation coefficient (CC = 0.63, $P=1.50\times10^{-5}$ computed using asymptotic t approximation) with DCA in MP. **e**, Genus network analysis of *A. parvulum*. Genus correlation networks are constructed in the healthy controls ($n=251$), MP ($n=67$), S0 ($n=73$), SI/II ($n=111$) and SIII/IV ($n=74$) groups. In total, 22 genera with abundances $>1\times10^{-4}$ and Spearman's correlation coefficients of >0.4 with *Atopodium* in at least one of the stages are used. Node sizes are proportional to the abundance of genera. Edge widths are proportional to the strength of correlation. **f**, Species correlation coefficients with *A. parvulum* in each group. Species with abundance $>1\times10^{-5}$ and Spearman's correlation coefficients >0.5 are shown. Species indicated in red are shown in **b**. Significant changes (elevation and depletion) are denoted as follows: +++, elevation with $P<0.005$; ++, elevation with $P<0.01$; +, elevation with $P<0.05$; −−−, depletion with $P<0.005$; −−, depletion with $P<0.01$; −, depletion at $P<0.05$. The boxes represent 25th–75th percentiles, black lines indicate the median and whiskers extend to the maximum and minimum values within 1.5× the interquartile range.

enriched with the genus *Bacteroides* had low abundance of *Prevotella*. Notably, the genus *Megamonas* was detected as highly abundant (within the 10 most abundant genera) in 118 out of 616 patients (19.2%) across all groups, but was rarely detected in the remainder, showing an uncommon distribution in the population. *Megamonas* has not previously been reported as a dominant genus in gut microbiome studies with European and American subjects, but was found in studies with Chinese individuals, which suggests that this genus might be characteristic of Asian populations[11]. Human genome content was significantly higher in various stages of CRC (S0, $P=0.00175$; SI/II, $P=8.19\times10^{-5}$; SIII/IV, $P=1.05\times10^{-5}$; one-sided Mann-Whitney U test) than in the healthy controls (Extended Data Fig. 3). Principal component analysis (PCA) identified the two most variable clusters in all subjects (Fig. 1b) and in 251 healthy controls (Extended Data Fig. 3b), *Bacteroides* and *Prevotella*, both of which have been defined as major contributors to human gut enterotypes[12].

Fitting using the Dirichlet multinomial mixture model[13] resulted in four community types among all 616 individuals, one of which was dominated by genus *Prevotella* (Extended Data Fig. 3e). With regard to known risk factors (body mass index ($P=0.8773$), alcohol consumption ($P=0.2989$) and smoking history ($P=0.3151$)), no significant ($P<0.005$) differences were detected among the four community types. Neither the colonoscopic findings (healthy, MP, S0, SI/II or SIII/IV; $P=0.2864$) nor tumor locations (left colon, right colon or rectum; $P=0.5231$) were associated with any community types. Age was differentially distributed ($P=7.80\times10^{-10}$) and the *Prevotella* cluster was predominated found in males (79.1%) ($P=6.82\times10^{-8}$).

Propionate and butyrate, major energy sources in the large intestine[14], ranked as the two most abundant metabolites. The PCA showed large variation in dihydrouracil and urea in addition to propionate and butyrate in the population (Fig. 1b). On the basis of the PCA, none of the stages, tumor locations or genders was associated with variation in metabolite profiles (Extended Data Fig. 4).

Compared to the healthy controls, we found microbiome shifts in MP and S0, in addition to SI/II and SIII/IV, which were highly distinct across stages. A number of species in the phyla Firmicutes, Fusobacteria and Bacteroidetes was predominantly elevated in samples from S0, SI/II and SIII/IV, increasing with the degree of malignancy. Elevated species of Fusobacteria appeared in at least two stages, whereas many of the species from Firmicutes and Bacteroidetes were stage-specific. In the phylum Proteobacteria, a large number of species were elevated only in MP (Fig. 2a). The genus *Bifidobacterium* was depleted mainly in S0. We noted two patterns of significant ($P<0.005$) species elevation: the first increased across early to later stages, whereas the second was elevated only in the early stages. The former was characterized by *F. nucleatum* (for example, *F. nucleatum* ssp. *nucleatum* (S0, $P=7.64\times10^{-5}$, $q=0.0492$; SI/II, $P=1.47\times10^{-8}$, $q=5.63\times10^{-5}$; SIII/IV, $P=6.93\times10^{-11}$, $q=2.62\times10^{-7}$)), *Solobacterium moorei* (S0, $P=1.18\times10^{-5}$, $q=0.0381$; SI/II, $P=0.000601$, $q=0.1355$; SIII/IV; $P=5.91\times10^{-5}$, $q=0.0195$), *Peptostreptococcus stomatis* (SI/II, $P=4.62\times10^{-6}$, $q=3.22\times10^{-3}$; SIII/IV, $P=1.98\times10^{-10}$, $q=3.75\times10^{-7}$), *Peptostreptococcus anaerobius* (SI/II, $P=7.57\times10^{-5}$, $q=2.90\times10^{-3}$; SIII/IV, $P=1.83\times10^{-10}$, $q=3.75\times10^{-7}$), *Lactobacillus sanfranciscensis* (S0, $P=0.000616$, $q=0.118$; SIII/IV, $P=0.00328$, $q=0.101$), *Parvimonas micra* (SI/II, $P=1.89\times10^{-5}$, $q=1.04\times10^{-3}$; SIII/IV, $P=6.15\times10^{-13}$, $q=4.66\times10^{-9}$) and *Gemella morbillorum* (S0, $P=0.000257$, $q=0.0800$; SI/II, $P=5.62\times10^{-7}$, $q=6.15\times10^{-4}$; SIII/IV, $P=1.79\times10^{-9}$, $q=2.11\times10^{-6}$). The latter pattern was characterized by *Atopobium parvulum* (MP, $P=0.00338$, $q=0.153$; S0, $P=7.03\times10^{-5}$, $q=0.0492$), *Actinomyces odontolyticus* (S0, $P=0.000164$, $q=0.0636$), *Desulfovibrio longreachensis* (S0, $P=0.00164$, $q=0.188$) and *Phascolarctobacterium succinatutens* (S0, $P=0.00236$, $q=0.242$). Abundance of *A. parvulum* was validated using quantitative PCR (Extended Data Fig. 5).

In addition, we identified species newly associated with CRC, of which *Colinsella aerofaciens* ($P=0.000840$, $q=0.0544$), *Dorea longicatena* ($P=0.000925$, $q=0.0557$), *Porphyromonas uenonis* ($P=0.000439$, $q=0.0475$), *Selenomonas sputigena* ($P=0.00369$, $q=0.101$) and *Streptococcus anginosus* ($P=0.00177$, $q=0.0788$) were significantly elevated in SIII/IV with all four analytic pipelines used (see Methods). In line with a previous study[5], two butyrate producers, *Lachnospira multipara* (S0, $P=0.000596$, $q=0.585$; SI/II, $P=0.000801$, $q=0.725$; SIII/IV, $P=0.000116$, $q=0.877$) and *Eubacterium eligens* (S0, $P=0.00147$, $q=0.698$), were significantly depleted in CRC stages. Sulfide-producing bacteria, including *Desulfovibrio vietnamensis* (SIII/IV, $P=0.00109$, $q=0.0565$), *D. longreachensis* (S0, $P=0.00164$, $q=0.188$) and *Bilophila wadsworthia* (SIII/IV, $P=0.00408$, $q=0.101$), were elevated.

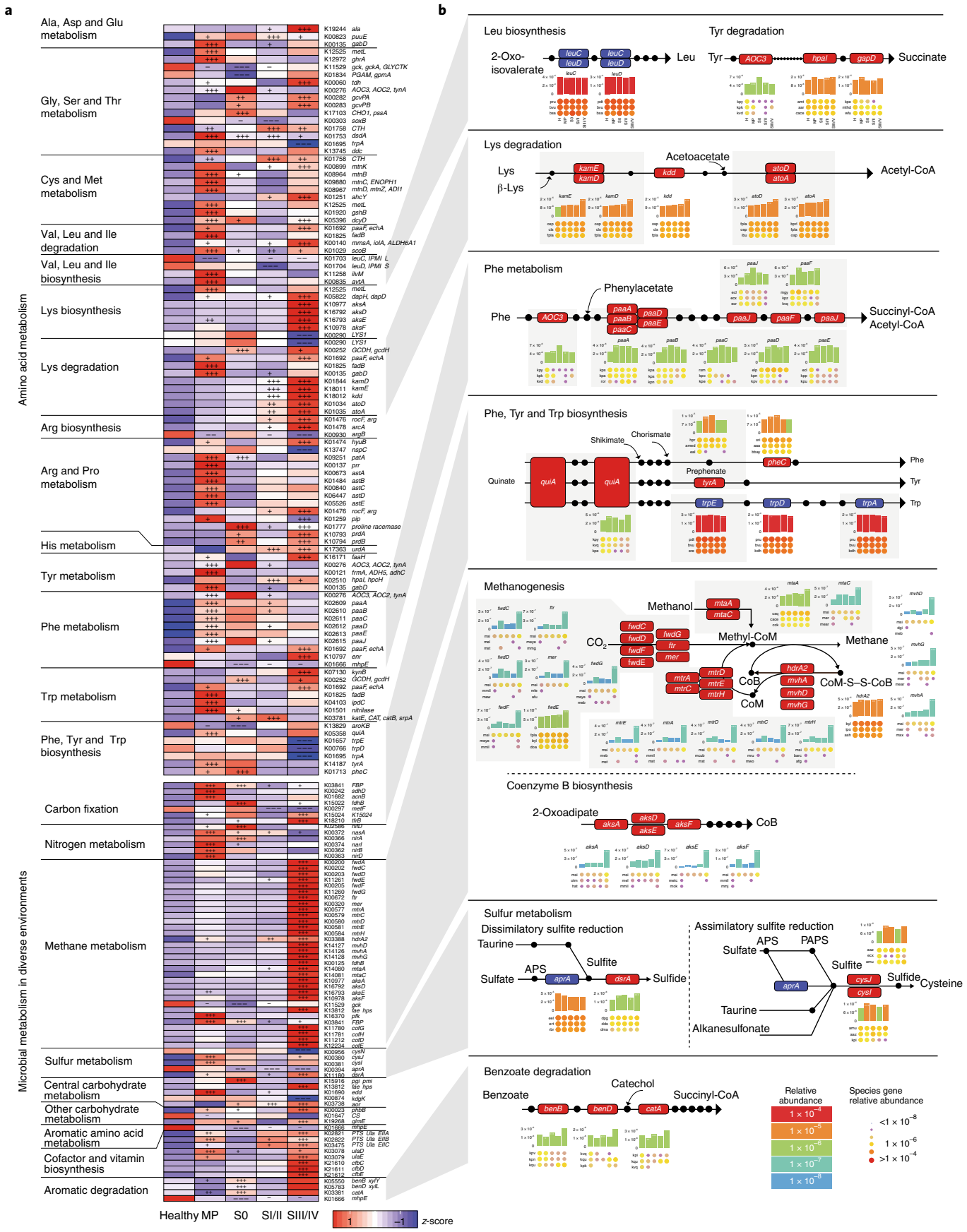
Bacterial replication rates[15] were significantly ($P<0.005$) higher for *G. morbillorum* (SI/II, $P=0.000436$; SIII/IV, $P=5.09\times10^{-7}$), *P. micra* (SI/II, $P=0.000114$; SIII/IV, $P=2.00\times10^{-6}$) and *P. stomatis* (SI/II, $P=0.00424$; SIII/IV, $P=8.07\times10^{-6}$) for the indicated stages, and those of *F. nucleatum* ssp. *nucleatum* ($P=0.000386$) and *D. longicatena* ($P=0.000246$) were significantly higher in SIII/IV, compared to healthy controls (Extended Data Fig. 6). Higher replication rates could explain higher abundances of species, suggesting the possibility that these bacteria may be metabolically active. Notably, *S. moorei* (MP, $P=0.000299$) and *C. aerofaciens* (S0, $P=0.000215$) showed significantly higher replication rates in MP and S0, respectively, before elevation in relative abundance in later stage(s). Overall, our findings indicate that microbial alterations of a number of species might be related to early CRC progression.

In total, 65 metabolites showed significant ($P<0.005$) differences in at least one of the stages compared to healthy controls (Extended Data Fig. 7 and Supplementary Table 3). Bile acids, short-chain fatty acids, amino acids and elements of central carbon metabolism were focuses of attention, as they have a close relationship with intestinal microbiota[16] (Fig. 2c and Extended Data Fig. 8).

Compared to the healthy controls, we found significant increases in the concentration of deoxycholate (DCA) in MP ($P=0.000118$, $q=0.0503$), and glycocholate ($P=0.00100$, $q=0.0508$) and taurocholate ($P=0.00195$, $q=0.0655$) in S0. In addition, the concentrations of branched-chain amino acids (isoleucine (S0, $P=0.00124$, $q=0.0508$), leucine (S0, $P=0.000371$, $q=0.0507$; SIII/IV, $P=0.00314$, $q=0.0931$), valine (S0, $P=0.000483$,

Fig. 3 | CRC-associated changes in microbial genes summarized in KO genes and KEGG pathway modules. a,**b**, Gene abundances were assessed for significant elevation or depletion ($P<0.005$; one-sided Mann–Whitney *U* test) in each of the four stages, MP ($n=67$), S0 ($n=73$), SI/II ($n=111$) and SIII/IV ($n=74$), compared to the healthy controls ($n=251$). **a**, Relative abundance of KO genes involved in amino acid metabolism and representative microbial metabolism (for example, methane metabolism, sulfur metabolism and aromatic degradation) that showed significant difference(s) in at least one of the four stages are shown in the heat map. KO genes with prevalence 5% or higher (KO genes detected in more than 5% out of 576 subjects) are shown. Significant changes (elevation and depletion) are denoted as follows: +++, elevation with $P<0.005$; ++, elevation with $P<0.01$; +, elevation with $P<0.05$; −−−, depletion with $P<0.005$; −−, depletion with $P<0.01$; −, depletion at $P<0.05$. **b**, Representative KO genes appearing in **a** are shown in pathway modules modified from KEGG pathway maps 'Valine, leucine and isoleucine biosynthesis', 'Lysine degradation', 'Tyrosine metabolism', 'Phenylalanine metabolism', 'Phenylalanine, tyrosine and tryptophan biosynthesis', 'Methane metabolism', 'Sulfur metabolism' and 'Benzoate degradation'. Each box in a pathway represents a KO gene, and is marked in red for elevation or in blue for depletion at any of the stages. Bar plots show relative gene abundances averaged over samples within each of the five groups (healthy (H), MP, S0, SI/II and SIII/IV from left to right) and are colored according to the order of the values. Each KO gene is composed of organism genes represented by circles. The sizes and colors of the circles are proportional to the relative abundances of the organism genes. Organism genes are grouped into one row and indicated by the organism name. The three most abundant organisms in the healthy controls are shown using three letter codes (for example, pru for *Prevotella ruminicola*, bvu for *Bacteroides vulgatus*). Other organism names are abbreviated and denoted in Supplementary Table 4. Gene numbers linked to each of the genes are listed in Supplementary Table 5. For other amino acids, see Extended Data Fig. 8. Dots in each pathway represent intermediate metabolites. APS, adenosine 5′-phosphosulfate; PASP, 3′-phosphoadenosine 5′-phosphosulfate.

$q = 0.0508$)), as well as the concentrations of phenylalanine (S0, $P = 0.000697$, $q = 0.0508$), tyrosine (S0, $P = 0.00136$, $q = 0.0508$), glycine (S0, $P = 0.00497$, $q = 0.120$) and serine (SIII/IV, $P = 0.00178$, $q = 0.0900$) were increased. Isovalerate, a branched-chain fatty acid produced by bacterial fermentation from leucine[17] increased gradually from S0 to SIII/IV (SIII/IV, $P = 0.00188$, $q = 0.0900$).



As DCA was elevated in MP, we searched for species that might correlate with this metabolite. *B. wadsworthia* was the only species significantly associated with DCA in MP (Fig. 2d). Correlation coefficients were positive in other stages but not significant. *B. wadsworthia* is known to grow in medium that contains taurocholate[18], a conjugate form of the DCA precursor, cholate.

Given that *A. parvulum* was significantly elevated in MP and S0, and has been reported to constitute a network hub of H_2S-producing bacteria via high co-occurrence relationships with *Streptococcus* in patients with inflammatory bowel disease[19], we examined correlations of abundance of this species with other species. Numbers of bacteria correlated with *Atopobium* markedly increased in S0 at both genus and species levels (Fig. 2e,f). There was a strong correlation between *A. parvulum* and *A. odontolyticus*, *S. anginosus*, *S. moorei* and *G. morbillorum*, for which the relative abundances increased in S0 and/or at subsequent stages. Increases in *Atopobium* even in the early stage of CRC suggests that it could have a strong influence on the H_2S-producing bacterial community.

A total of 1,243 Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology genes (KO genes) showed significant ($P<0.005$) elevation and 96 significant depletion in at least one of the stages, compared to healthy controls. Because fecal amino acid concentrations were markedly changed from S0 (Fig. 2c), we examined microbial gene abundances to check the roles of the microorganisms in amino acid metabolism (Fig. 3a).

Changes in the gene abundances are shown in a pathway representation (Fig. 3b and Extended Data Fig. 8b). Pathway modules, which set a limit to bacterial biosynthesis and degradation, were manually constructed by modifying KEGG pathway maps that referred to 'Microbial metabolism in diverse environments' (map01120) or the literature[20–23].

Among the most differentially abundant pathways, aromatic amino acid metabolism and sulfide-producing pathways were found to be associated with CRCs. Genes involved in phenylalanine and tyrosine biosynthesis were significantly elevated, among which *pheC* ($P=1.94\times10^{-5}$, $q=0.0297$ in S0) was identified as a top-scoring marker to distinguish S0 cases from healthy controls (Fig. 4b and Extended Data Fig. 9). In the catabolic pathways, genes involved in phenylalanine degradation through the production of toxic phenylacetate[24–26] were elevated mainly in MP. Genes involved in the biosynthesis of tryptophan were significantly ($P<0.005$) depleted in SIII/IV. Dissimilatory sulfate reductase subunit A (*dsrA*), which is responsible for production of genotoxic hydrogen sulfide[27], was significantly elevated in SIII/IV ($P=0.00499$, $q=0.0729$) (Fig. 3b). *dsrA* is found to be active in a number of sulfate-reducing bacteria including *Desulfovibrio* spp.[28] and — for instance, *Desulfovibrio piger* ($P=0.0178$, $q=0.226$) — was substantially high in SIII/IV.

Most S0 lesions can be cured by endoscopic approaches and there is a broad window of opportunity for detection[3]. In order to investigate the potential of gut metagenomic and metabolomic parameters to act as diagnostic markers, we constructed random-forest and LASSO logistic regression classifiers to discriminate S0 and SIII/IV cases from healthy controls. Four types of models were constructed based on species only, KO gene only, metabolite only data or a combination of the three. Comparison of the classification potential among the four models resulted in outperformance of the combination model over the individual models in the classification of both S0 and SIII/IV (Fig. 4b,c). Random-forest classifiers achieved an area under the receiver-operating characteristic (ROC) curve (AUC) of 0.78 and 0.85 to detect a patient with S0 and SIII/IV CRC, respectively. The results obtained using the LASSO logistic regression classifier are shown in Extended Data Fig. 9.

In the S0 classification, high-ranking features were mainly KO genes, including *pheC* (which encodes cyclohexadienyl dehydratase). Other features included *D. longreachensis*, *S. moorei*, leucine, valine, phenylalanine and succinate (Fig. 4b), which were detected as differentially distributed (Fig. 2b,c). The top-ranking features of the SIII/IV classification included oral anaerobes, such as *P. micra*, *P. stomatis*, *F. nucleatum* and *P. anaerobius*, which have previously been identified as marker species for CRCs[5,29,30] (Fig. 4b).

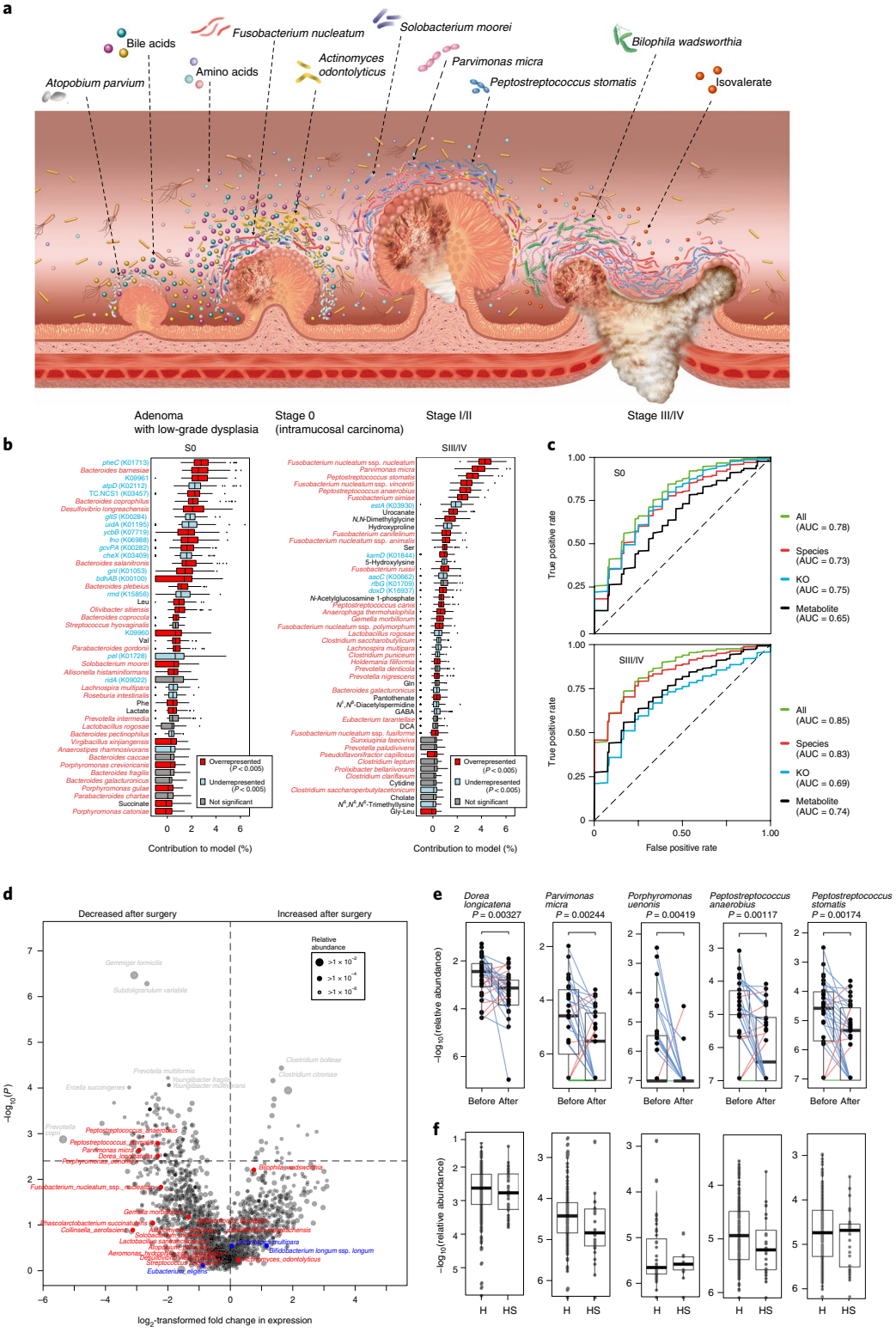
Metagenome data were obtained from 28 patients with CRC (SI/II and SIII/IV) before and approximately 1 year after surgical treatment. The relative abundances of *P. stomatis*, *P. anaerobius*, *P. micra*, *P. uenonis* and *D. longicatena* among the 22 species discussed in Fig. 2b were reduced after tumor removal (Fig. 4d,e). Comparison of these five species with fecal samples from HS subjects did not show significant differences (Fig. 4f). The main results obtained in the present study are shown in Fig. 4a.

The present study on the relationships between the gut ecosystem and multistep tumorigenesis yielded information on microorganism and microorganism-derived metabolite profiles in CRC. Our results demonstrate that microbiome and metabolome shifts in MP and S0 occurred, in addition to the more advanced stages. The shifts were highly distinct across stages. Two patterns of species elevation were found: the first consisted of a continuous increase from early stages onwards, whereas the other showed elevation only in early stages. The latter pattern was the main focus of the present study,

Fig. 4 | Microbial dynamics and their diagnostic potential during multistep CRC progression. a, Graphical representation of major microbial and metabolomic alterations during multistep CRC progression. **b**,**c**, Metagenomic and metabolomic markers for detecting patients with S0 ($n=27$) (left) and SIII/IV ($n=54$) (right) CRCs from the healthy controls ($n=127$) identified from random-forest classifiers based on species (red), KO genes (blue) or metabolites (black) alone, or the combination (green) of the three features. For the S0 classification, individual models used 29 species, 16 KO genes and 24 metabolites as features. For SIII/IV classification, individual models used 55 species, 5 KO genes and 62 metabolites as features. In the combination models, species, KO gene and metabolite features were selected from the individual models. GABA, γ-aminobutyric-acid; KO, KEGG orthology gene. The *x* axis presents the percentage contribution of the features to the model in each test (see Methods). The boxes represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers. The boxes are marked in red for overrepresentation, in light-blue for underrepresentation or in gray for no significant changes ($P<0.005$; one-sided Mann–Whitney *U* test) in S0 or SIII/IV compared to the healthy controls. Performance of the classifiers using AUCs was evaluated using 10 randomized 10-fold cross-validation. **d**, Comparison of species relative abundances before and after surgery ($n=28$). The *x* axis indicates log-transformed fold change in expression and the *y* axis *P* values analyzed using a one-sided Wilcoxon signed-rank test. The horizontal dashed line indicates a *P* value of 0.005. The sizes of the circles indicate the abundance of each species averaged over pre- and post-operative statuses. Species discussed in Fig. 2 as previously reported as increased or decreased in CRCs, known as butyrate producers, H_2S producers, cancer-related species reported in other cohorts, or newly associated with CRCs in the present study as elevated or depleted species, are highlighted in red (increased) and blue (decreased). **e**, Relative abundances of five species, *D. longicatena*, *P. micra*, *P. uenonis*, *P. anaerobius* and *P. stomatis*, which are significantly decreased after surgery ($P<0.005$; one-sided Wilcoxon signed-rank test) in comparison to before surgery in 28 patients with SI/II/III CRCs. The boxes represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers. Increases and decreases are colored in red and blue, respectively. **f**, Relative abundances of the same five species in HS samples ($n=40$) for whom fecal samples before the operation were not available are shown in comparison to the healthy controls ($n=251$).

given the possibility that changes in the state of the gut microbiome predispose individuals to develop CRC. In particular, our results showed that the abundances of *F. nucleatum* and *S. moorei* were elevated in S0, which may indicate contributions of these species to initial events in tumorigenesis, in addition to the known association with advanced CRCs[5,31].

Notably, *A. parvulum* and *A. odontolyticus* showed significant increases only in MP and/or S0. Our network analyses demonstrated the co-occurrence of the two species in S0 and a strong correlation of *A. parvulum* with *Streptococcus* spp., which are known H_2S producers[32]. *A. parvulum* has been shown to constitute a network hub of H_2S-producing bacteria through high co-occurrence relationships with *Streptococcus* in patients with inflammatory bowel disease[19]. *A. odontolyticus* is often present in the oral cavity and gastrointestinal tract of healthy humans and, in particular, has been shown to be one of the predominant *Actinomyces* species in developing



biofilms on tooth surfaces[33]. It has previously been reported that the presence of *A. odontolyticus* in fecal samples of patients with carcinoma in adenoma, although the sample size was small[34]. Further studies will be necessary to clarify the precise mechanisms by which these bacteria might contribute to tumorigenesis.

The concentration of DCA was significantly increased in patients with MP. DCA is known to lead to increased DNA damage and mutations[35]. In animal studies, administration of bile acids resulted in a higher incidence of tumors in the gut[36]. *B. wadsworthia*, the growth of which is stimulated by bile[18], was the only species that was significantly correlated with DCA in the present study (Fig. 2d). Concentrations of conjugated bile acids (taurocholate and glycocholate) were also increased in S0. Our questionnaires demonstrated that *B. wadsworthia* was positively correlated with intake of dietary protein ($P=0.00278$) and meat ($P=0.00248$) in S0 (Extended Data Fig. 10). *B. wadsworthia*, a close relative of *Desulfovibrio* species, is known to cause inflammation, but has not been intensively studied in association with carcinogenesis[29], although it is often discussed in the context of dysbiosis[9,18]. Although these microorganisms are a normal part of the gut ecosystem, an overabundance of the bacteria and their metabolites in the colorectum may lead to inflammation and damage to DNA.

The following limitations should be considered with our study. First, although this study includes a large CRC cohort, a validation cohort is needed in the future. Second, we did not determine the total cell count in fecal samples and recent reports have shown that microbial load is a key driver of observed alterations in the microbiota in some diseases[37]. Absolute rather than relative abundances might be a better indicator of multistep CRC carcinogenesis. Third, although our methods (that is, fecal samples collected immediately at first defecation after starting oral administration of a bowel-cleansing agent) appear to be appropriate for collecting and freezing material from subjects undergoing colonoscopy at the hospital, there is a lack of similar study reports for comparison. In addition, the effect of bowel cleansing on the resulting metabolome data in cases with CRC remains to be investigated. Therefore, further studies are warranted to validate our sampling protocol.

In the present study, we could not clarify any potential microbial involvement in stages prior to the formation of adenomas and more detailed causalities between microbiome and/or metabolome and tumors. Therefore, we are prospectively collecting fecal and tissue biopsy samples from individuals who undergo colonoscopy at regular intervals for future microbiome analyses with a possible longitudinal component. In addition, clarification of relationships between the gut microbiome and tumor molecular characteristics in individual patients with CRC will be necessary to understand the roles of the microbiome in CRC carcinogenesis. Cases who had a hereditary or suspected hereditary disease were excluded in the present study. The adenomas were limited to conventional-type colorectal adenomas, not including serrated lesions. Recently, it has been reported that these latter may be associated with tissue bacteria[38–40]. Metagenomic and metabolomic data derived from fecal and/or tissue samples from patients with hereditary or suspected hereditary diseases and patients with serrated lesions may clarify other aspects of CRC tumorigenesis.

In conclusion, we observed dynamic shifts in microbial composition, gene abundance in gut microbiota and metabolites during multistep CRC progression. Although it is not clear whether these species and metabolites directly cause tumorigenesis, structural shifts in gut microbiota may lead to changes in the oncogenic microenvironment. Furthermore, the present study highlights that CRC progression may be influenced by the metabolic output of the entire microbiota, as well as the presence of cancer-associated organisms. We believe that CRC is fundamentally not only a genetic but also a microbial disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0458-7.

Received: 10 September 2018; Accepted: 11 April 2019;
Published online: 6 June 2019

References

1. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *Lancet* **383**, 1490–1502 (2014).
2. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
3. Jones, S. et al. Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105**, 4283–4288 (2008).
4. Ashktorab, H., Kupfer, S. S., Brim, H. & Carethers, J. M. Racial disparity in gastrointestinal cancer risk. *Gastroenterology* **153**, 910–923 (2017).
5. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
6. Castellarin, M. et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).
7. Kostic, A. D. et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
8. 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-kB, and up-regulating expression of microRNA-21. *Gastroenterology* **152**, 851–866 (2017).
9. Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **12**, 661–672 (2014).
10. Hirayama, A. et al. Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. *Cancer Res.* **69**, 4918–4925 (2009).
11. Liao, M. et al. Comparative analyses of fecal microbiota in Chinese isolated Yao population, minority Zhuang and rural Han by 16sRNA sequencing. *Sci. Rep.* **8**, 1142 (2018).
12. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
13. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509**, 357–360 (2014).
14. Wong, J. M., de Souza, R., Kendall, C. W., Emam, A. & Jenkins, D. J. Colonic health: fermentation and short chain fatty acids. *J. Clin. Gastroenterol.* **40**, 235–243 (2006).
15. Emiola, A. & Oh, J. High throughput in situ metagenomic measurement of bacterial replication at ultra-low sequencing coverage. *Nat. Commun.* **9**, 4956 (2018).
16. Brestoff, J. R. & Artis, D. Commensal bacteria at the interface of host metabolism and the immune system. *Nat. Immunol.* **14**, 676–684 (2013).
17. Zarling, E. J. & Ruchim, M. A. Protein origin of the volatile fatty acids isobutyrate and isovalerate in human stool. *J. Lab. Clin. Med.* **109**, 566–570 (1987).
18. Devkota, S. et al. Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in *Il10*$^{-/-}$ mice. *Nature* **487**, 104–108 (2012).
19. Mottawea, W. et al. Altered intestinal microbiota–host mitochondria crosstalk in new onset Crohn's disease. *Nat. Commun.* **7**, 13419 (2016).
20. Xu, H. et al. Isoleucine biosynthesis in *Leptospira interrogans* serotype lai strain 56601 proceeds via a threonine-independent pathway. *J. Bacteriol.* **186**, 5400–5409 (2004).
21. Bui, T. P. et al. Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. *Nat. Commun.* **6**, 10062 (2015).
22. Prieto, M. A., Díaz, E. & García, J. L. Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of *Escherichia coli* W: engineering a mobile aromatic degradative cluster. *J. Bacteriol.* **178**, 111–120 (1996).
23. Teufel, R. et al. Bacterial phenylalanine and phenylacetate catabolic pathway revealed. *Proc. Natl Acad. Sci. USA* **107**, 14390–14395 (2010).
24. Russell, W. R. et al. High-protein, reduced-carbohydrate weight-loss diets promote metabolite profiles likely to be detrimental to colonic health. *Am. J. Clin. Nutr.* **93**, 1062–1072 (2011).
25. Russell, W. R. et al. Major phenylpropanoid-derived metabolites in the human gut can arise from microbial fermentation of protein. *Mol. Nutr. Food Res.* **57**, 523–535 (2013).
26. Windey, K., De Preter, V. & Verbeke, K. Relevance of protein fermentation to gut health. *Mol. Nutr. Food Res.* **56**, 184–196 (2012).
27. Attene-Ramos, M. S., Wagner, E. D., Plewa, M. J. & Gaskins, H. R. Evidence that hydrogen sulfide is a genotoxic agent. *Mol. Cancer Res.* **4**, 9–14 (2006).

28. Loubinoux, J., Bisson-Boutelliez, C., Miller, N. & Le Faou, A. E. Isolation of the provisionally named *Desulfovibrio fairfieldensis* from human periodontal pockets. *Oral Microbiol. Immunol.* **17**, 321–323 (2002).
29. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
30. 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).
31. Bullman, S. et al. Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443–1448 (2017).
32. Carbonero, F., Benefiel, A. C., Alizadeh-Ghamsari, A. H. & Gaskins, H. R. Microbial pathways in colonic sulfur metabolism and links with health and disease. *Front. Physiol.* **3**, 448 (2012).
33. Könönen, E. & Wade, W. G. *Actinomyces* and related organisms in human infections. *Clin. Microbiol. Rev.* **28**, 419–442 (2015).
34. Kasai, C. et al. Comparison of human gut microbiota in control subjects and patients with colorectal carcinoma in adenoma: terminal restriction fragment length polymorphism and next-generation sequencing analyses. *Oncol. Rep.* **35**, 325–333 (2016).
35. Bernstein, H., Bernstein, C., Payne, C. M. & Dvorak, K. Bile acids as endogenous etiologic agents in gastrointestinal cancer. *World J. Gastroenterol.* **15**, 3329–3340 (2009).
36. Suzuki, K. & Bruce, W. R. Increase by deoxycholic acid of the colonic nuclear damage induced by known carcinogens in C57BL/6J mice. *J. Natl Cancer Inst.* **76**, 1129–1132 (1986).
37. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
38. Tahara, T. et al. *Fusobacterium* in colonic flora and molecular features of colorectal carcinoma. *Cancer Res.* **74**, 1311–1318 (2014).
39. Ito, M. et al. Association of *Fusobacterium nucleatum* with clinical and molecular features in colorectal serrated pathway. *Int J. Cancer* **137**, 1258–1268 (2015).
40. Mima, K. et al. *Fusobacterium nucleatum* in colorectal carcinoma tissue and patient prognosis. *Gut* **65**, 1973–1980 (2016).

Acknowledgements

We thank all patients and their families who participated in this study, S. Goto for technical advice and A. Kaya, C. Shima, K. Igarashi, R. Usui, K. Murakami, I. Take, M. Sezawa, M. Iwahara, M. Komori, Z. Nakagawa, Y. Ohara and K. Kamezaki for expert technical assistance. Computations were partially performed on the NIG supercomputer at the ROIS National Institute of Genetics. This work was supported by grants from the National Cancer Center Research and Development Fund (25-A-4 and 28-A-4 to S.Y., S.F and T. Yamada and 29-A-6 to T. Yamada and T. Shibata); Practical Research Project for Rare/Intractable Diseases from the Japan Agency for Medical Research and Development (AMED) (JP18ek0109187 to S.Y., S.M., Y.S., S.F. and T. Yamada; JP18jk0210009 to S.Y. and T. Shibata); AMED-CREST (JP18gm0710003 to S.Y. and T. Soga.); JST (Japan Science and Technology Agency)-PRESTO (JPMJPR1537 to S.F. and JPMJPR1507 to T. Yamada); JSPS (Japan Society for the Promotion of Science) KAKENHI (16H04901, 17H05654 and 18H04805 to S.F.; 16J10135, 142558 and 221S0002 to T. Yamada); Joint Research Project of the Institute of Medical Science, the University of Tokyo (2017–2107 to S.Y. and T. Shibata); Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University (to S.Y.); the Takeda Science Foundation (to S.Y.) and Suzuken Memorial Foundation (H25–2–11 to S.Y.).

Author contributions

S.Y., T.N., Y.S., S.F., T. Shibata and T. Yamada contributed to the study concept and design. S.Y., T.N., T. Sakamoto, S.S., M.M., H.T., M.Y., T.M., M.I., T. Yamaji, T. Yachida and Y.S. collected the clinical samples and information. S.Y., F.H., H.R., T. Soga, A.T., Y.O., T.H. and S.F. performed metagenome and metabolome experiments. S.M., H.S., H.W., K.M., Y.N., M.K. and T. Yamada performed bioinformatics analyses. S.Y., S.M., H.S., K.M. and T. Yamada wrote the manuscript. K.K., M.H., H.N. and T. Shibata supervised the study.

Competing interests

S.F. and T. Yamada are founders of Metabologenomics. The company is focused on the design and control of the gut environment for human health. The company had no control over the interpretation, writing or publication of this work. The terms of these arrangements are being managed by Keio University and Tokyo Institute of Technology in accordance with its conflict of interest policies. S.Y., S.F. and T. Yamada are currently applying for a patent (2018–18134/PCT/JP2019/3825, 'Method for diagnosing the early stage of colorectal cancers based on the gut microbiome and metabolome profiles').

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0458-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0458-7.

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

Correspondence and requests for materials should be addressed to S.Y. or T. Yamada.

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Methods

Study subjects and sample collection. This study was conducted with subjects undergoing total colonoscopy in the National Cancer Center Hospital, Tokyo, Japan. The samples and clinical information used in this study were obtained under conditions of informed consent and with approval of the institutional review boards of each participating institute (National Cancer Center, 2013–244; Tokyo Institute of Technology, 2014018). Stool samples were collected immediately at first defecation after starting oral administration of a bowel-cleansing agent at hospital and stored frozen on dry ice. We previously demonstrated high pairwise Pearson's correlation coefficients for taxonomic profiles and no significant differences in taxonomic abundance of 20 dominant genera between this type of sample and frozen fecal examples taken one day before colonoscopy (standard samples)[41]. The subjects were provided with the same commercial low residue diet the day before the colonoscopy procedure. Data on life style, including dietary habits, were obtained with detailed questionnaires (475 question items, 25 pages), based on the example used in the Japan Public Health Center-based prospective study[42]. Cases who had a hereditary or suspected hereditary disease (for example, familial adenomatous polyposis, hereditary non-polyposis colorectal cancer, microsatellite instability-high), an inflammatory bowel disease, an abdominal surgical history or for which stool samples were insufficient for data collection were excluded from the study. Inter-group (healthy, MP, S0, SI/II, SIII/IV and HS) distributions of body-mass index (BMI), alcohol consumption (grams per day) and smoking habits (Brinkman index) were analyzed using a Kruskal–Wallis rank-sum test. Inter-group distribution of gender was analyzed using Pearson's χ^2 test with d.f. = 5.

DNA extraction. DNA was extracted from frozen fecal samples with the bead-beating method as previously described[43] using a GNOME DNA Isolation Kit (MP Biomedicals). DNA quality was assessed with an Agilent 4200 TapeStation (Agilent Technologies). After final precipitation, the DNA samples were resuspended in TE buffer and stored at −80 °C before further analysis.

Whole-genome shotgun sequencing. Sequencing libraries were generated with a Nextera XT DNA Sample Prep Kit (Illumina). Library quality was confirmed with an Agilent 4200 TapeStation. Whole-genome shotgun sequencing of fecal samples was carried out on the HiSeq2500 platform (Illumina). All samples were paired-end sequenced with a 150-bp read length to a targeted data size of 5.0 Gb.

Quality control. A total of 31,797,649,036 (49,375,231 on average) paired-end reads, covering 4,772,084,552,120 (7,410,069,180 on average) base pairs in total, were underwent quality control as follows. Raw reads containing the letter 'N' (base pair not identified) were discarded. Reads containing the bacteriophage *phiX* DNA sequences were identified by mapping them against the reads using Bowtie 2 (version 2.2.9)[44] with preset options in '–fast-local' and discarded. Reads were trimmed for adapter sequences and primer sequences using cutadapt (version 1.9.1)[45] for which the following options were used ('-a CTGTCTCTTATACACATCTCCGAGCCCACGAGAC -O 33 -q 17' for the forward primer sequence; '-a CTGTCTCTTATACACATCTGACGCTGCCGACGA -O 32 -q 17' for the reverse primer sequence). Reads containing quality values of 17 or less consecutively were tailed-cut at the 3′ termini within the cutadapt program. Next, reads of lengths less than 50 base pairs were discarded. Reads of average quality values of 25 or less were discarded. Next, reads were mapped against the human genome (24 gi numbers: from 568336000 to 568336023, http://www.ncbi.nlm.nih.gov/nuccore/568336023/, GRCh38) using Bowtie2 (version 2.2.9). Those that were mapped were considered to be derived from the human genome and were discarded. Finally, unpaired reads were discarded. As a result, a total of 28,482,269,496 (44,227,127 on average) paired-end reads with 4,114,878,107,497 (6,389,562,279 on average) base pairs in total (referred to as the 'high-quality reads' hereafter) (Supplementary Table 6) were used for the following analyses.

Taxonomic profiling. The high-quality reads were aligned to a pre-calculated operational taxonomic unit (OTU) dataset stored in VITCOMIC2[46] using BLAST+ (version 2.2.30)[47] (cut-off: $E < 1 \times 10^{-8}$) so that reads were filtered for bacterial and archaeal 16S rRNA sequences, whereas tRNA, 23S rRNA and the internal transcribed spacer (ITS) sequences were excluded. There are a number of reference strategies applicable to taxonomic assignment for metagenome data. We used the All-Species Living Tree Project (LTP) of the SILVA database for two reasons. First, it has the advantage that the single-locus database contains more taxonomic entries than those covering multiple loci or complete genomes. Secondly, the reference genes in this database consist only of type strains that have been isolated and could therefore be cultured under particular conditions and relatively easily manipulated in an animal experiment. The filtered reads were aligned to the LTP of the SILVA database (version 123)[48] using BLASTn (cut-offs: $E < 1 \times 10^{-8}$, sequence identity > 97%, alignment coverage > 80%, bit score > 70)[48,49]. Only top hits were selected. As a result, a total of 8,367 species and 1,941 genera were identified. The relative abundance of a species was computed per sample, defined as the number of reads assigned to the species divided by the total number of aligned reads in the sample. If a read was aligned to more than one taxonomic sequence in the database with equal alignment scores, these taxonomies were given a value 1 divided by the number of the taxonomies so that they could 'share' the read. The relative abundance of the genus was computed as the sum of all species belonging to the genus. The generated profiles are referred to as 'species profile' and 'genus profile' hereafter (Supplementary Table 7).

In order to validate our species profiling pipeline, we used three other pipelines to obtain species-level profiles. We used the above high-quality reads to obtain species-level metagenomic OTUs (mOTUs) from taxonomic profiling by the mOTU profiler[50]. The database of this profiler is based on ten universal single-copy marker genes. These marker genes consist of reference genome-derived and metagenome-derived examples. We used two databases; with and without metagenome data-derived OTUs. The generated profiles encode 651 species and 838 species (Supplementary Table 8). High-quality reads were also used to obtain species-level taxonomic profiling by MetaPhlAn2[51] (version 2.6.0) with default parameters, which resulted in 623 species (Supplementary Table 9). The MetaPhlAn2 is based on species specific marker genes.

Microbial community structure analysis. The overall community structures of 616 metagenome and 406 metabolome data reads were examined using PCA. The community types of metagenomic samples were also analyzed by the Dirichlet multinomial mixture model using counts of sequencing reads[13]. The R package 'DirichletMultinomial'[52] was used. The maximum number of the clusters was four, among which the third cluster was predominated by *Prevotella* (Extended Data Fig. 3e). The other three clusters were of more diverse structure with the genus *Bacteroides* predominating. The cancer stages, tumor locations and gender were examined across different community types. Distributions of age, BMI, smoking habits (Brinkman index) and alcohol consumption (grams per day) in the four community types were tested using analysis of variance. Distributions of gender, colonoscopic findings (healthy, MP, S0, SI/II, SIII/IV and HS), and tumor locations (left colon, right colon, rectum and double or triple cancers, in addition to three other groups without cancer (healthy, MP, and HS)) were tested using Pearson's χ^2 test (gender, d.f. = 3; colonoscopic findings, d.f. = 15; tumor location, d.f. = 18).

Taxonomy tree construction. A taxonomy tree was constructed using GraPhlAn (version 0.9.7)[53]. Taxonomic hierarchy information was obtained from the SILVA database. Seven levels, including domain, phylum, family, genus and species, were used for Fig. 2b. Species were filtered so that those with an average abundance of 10^{-6} or higher and $P < 0.005$ (one-sided Mann–Whitney U test) for any of the stages are shown.

Analysis of microbial replication rates. We calculated replication rates using growth rate index (GRiD) (version 1.2). The algorithm is based on the estimation of the ratio between coverage at the peak (*ori*) and trough (*ter*) for the reference bacterial genome using M estimator with Tukey's biweight function. GRiD values demonstrate a positive relationship with the replication rate. Replication rates were computed for 20 out of 22 species presented in Fig. 2b (*A. odontolyticus, Actinomyces viscosus, A. parvulum, Bifidobacterium longum* ssp. *longum, B. wadsworthia, C. aerofaciens, D. longicatena, E. eligens, F. nucleatum* ssp. *nucleatum, G. morbillorum, L. multipara, L. sanfranciscensis, P. micra, P. anaerobius, P. stomatis, P. succinatutens, P. uenonis, S. sputigena, S. moorei* and *S. anginosus*) using GRiD[15] with parameter 'single' to refer a single reference genome at each calculation. Reference genomes were downloaded from the NCBI website (Supplementary Table 10) for which the reference genome IDs were cross-referenced with SILVA LTP identifiers using NCBI accession numbers. If SILVA identifiers were not cross-referenced with the NCBI database, a 'representative genome' was selected under the 'RefSeq category'. *D. longreachensis* and *D. vietnamensis* reference genomes were not found in the NCBI database. Replication rates were compared across each of the stages (MP, S0, SI/II and SIII/IV) with the healthy controls using one-sided Mann–Whitney U tests with an α level of 0.005 for significance. Owing to the coverage requirement of GRiD, for each reference genome, samples with coverage of 0.2 or less were omitted from the statistical test. For *A. odontolyticus, A. viscosus, A. parvulum, L. multipara, S. anginosus* and *S. sputigena*, only few samples had high replication rates based on this analysis.

Genus and species network analysis. Spearman's correlation coefficients were computed using relative abundance profiles of the genus and species for each of the stages (MP, S0, SI/II and SIII/IV). Genus correlation networks were constructed with species that had correlation coefficients of 0.4 or higher or −0.4 or lower with the genus *Atopobium*. Construction of networks was performed using yEd Graph Editor (version 3.18.11) (https://www.yworks.com/products/yed).

Sequence assembly. The high-quality reads were assembled per sample using IDBA_UD (version 1.1.1)[54] with parameters –mink 20 –maxk 120 –step 10. A total of 81,002,850 (125780.8 per sample on average) scaffolds were generated.

Functional profiling. Open reading frames (ORFs) were predicted on the obtained scaffolds using MetaGeneMark (version 3.26)[55] with parameter –g 11. As a result, 156,163,520 ORFs with amino acid lengths of 50 or longer were annotated with the KEGG GENES database (as of 2017)[56] using DIAMOND (version 0.9.10) (cut-offs: sequence identity > 40, bit score > 70, coverage > 80), giving a total of 126,761,506 ORFs, or annotated 'genes'. Gene abundances were computed as

follows. High-quality reads were mapped back onto the scaffolds using Bowtie 2 version 2.2.9[44]. Each ORF on each scaffold was scored for read coverage, which was defined as the number of base pairs mapped onto the corresponding scaffold regions divided by the lengths of the ORFs. As more than two ORFs match up to one gene, each gene abundance was computed as the average of the score values. Gene abundances were then summed up to a total of 7,242 KO genes, a KEGG-defined functional unit. The generated profile is referred to as the 'KO gene profile' (Supplementary Table 11).

Pathway functional characterization. Amino acid-related KO genes with pathway information were obtained from KEGG BRITE 'ko00001.keg' (list of KO genes with pathway maps) under the 'Amino acid metabolism' category. In order to investigate pathway modules known in microorganisms, we collected KEGG modules listed under 'Microbial metabolism in diverse environments' (map01120). In Fig. 3a, KO genes for which the prevalence was higher than 5% of all 576 samples are shown, as analyzed by Mann-Whitney *U*test ($P < 0.005$) for any of the stages (MP, S0, SI/II and SIII/IV) compared to the healthy controls. Representative reaction pathways shown in Fig. 3b and Extended Data Fig. 8b were manually constructed either by referring to the literature or modified from the reference maps in the KEGG pathway. In Fig. 3b and Extended Data Fig. 8b, KO genes are shown with $P < 0.005$ for any of the stages (MP, S0, SI/II and SIII/IV), whereas the remaining KO genes in the pathways were omitted. Genes and KEGG orthology genes are linked in KEGG. For each of the KO genes, the abundance of microbial genes is summed so that each component in each KO gene represents an organism. Organism names are stored using three letter codes in KEGG.

Validation of *A. parvulum* using quantitative PCR. Abundance of *A. parvulum* was validated using quantitative PCR (qPCR) from fecal samples of 73 patients with S0 CRC and 73 healthy controls. The copy number of the targeted region of *A. parvulum* 16S rRNA gene was estimated (Supplementary Table 12) in 1 µg of the extracted DNA. PCR products were sequenced using an *A. parvulum* F-primer (5′-TGGATAATACCGAATACTTCGAGACT-3′) and *A. parvulum* R-primer (5′-TGCAGGTACCGTCACTTTCG-3′) and qPCR was performed on a Rotor-Gene Q (QIAGEN) with TB Green Premix Ex Taq II (Takara Bio).

Metabolome analysis. Quantitative analysis of charged metabolites by CE-TOFMS was performed as previously described[57]. Fecal metabolites were extracted by vigorous shaking with methanol containing 20 µM each of methionine sulfone and D-camphol-10-sulfonic acid as the internal standards[58]. All CE-TOFMS experiments were performed using an Agilent CE system. CE-TOFMS metabolome data were obtained for 517 compounds (Supplementary Table 3). For the analysis, concentrations below the detection limit were substituted with zero, and metabolites for which levels were below the detection limit in all of the samples were excluded. The metabolite profile is provided in Supplementary Table 13.

Identification of metagenomic and metabolomic markers for detecting colorectal cancer. In order to identify metagenomic and metabolomic markers that can distinguish samples from patients with S0 ($n = 27$) and SIII/IV ($n = 54$) CRCs from samples of healthy controls ($n = 127$), we constructed classification models based on the species, KO gene and metabolite profiles using two different methods, random-forest and LASSO logistic regression[59]. Models were validated by 10-fold stratified cross-validation testing (we resampled dataset partitions 10 times). In each test, the accuracy of the model was examined using a ROC, and abundance filtering was carried out to remove low-abundance features by calculating the average relative abundance in each stage. The abundance threshold was determined to optimize the AUC (see Supplementary Table 14). Features were then standardized (by centering to mean 0 and dividing by the s.d. of each feature). Models were designed using one of three types of features (that is, species, KO genes or metabolites) alone, or the combination of the three.

In the random-forest model, two steps were carried out. In the first step, the model was constructed using each of the three profiles (species, KO gene and metabolite profiles) independently, for which all of the prefiltered features were used to perform a random-forest function with the indicated parameters (500 trees, balanced class weight, max features = square root of all features) to compute 'feature importance' (see the description below), for which the optimal number of features was determined using the recursive feature elimination method[60] with parameter step = 0.1 using five different random seeds. In the second step, the combination model was constructed using features determined in the separate species, KO gene and metabolite models. Features were then further selected using the recursive feature elimination method[60] as above (see Supplementary Table 14 for the number of features used in the model constructions). All analyses were carried out using the Python package 'scikit-learn'. Feature contributions to the models were output as feature importances. Figure 4 displays features with a nonzero value for the feature importances in at least 50% of the tests. In the LASSO logistic regression model, features were selected using L_1 regularization. The LASSO hyperparameter was set as shown in Supplementary Table 14. Feature contributions to the models were computed using the percentage absolute values of the regression coefficients. Extended Data Fig. 9 displays features with a nonzero coefficient in at least 50% of the tests. Assessment of the diagnostic potential of the known CRC risk factors (age, gender, BMI, smoking and alcohol consumption) resulted in lower predicting accuracy (Extended Data Fig. 9c).

Species–metabolome correlation. We computed pairwise correlation coefficients using Spearman's correlation coefficients between species and metabolites for each of the stages (MP ($n = 40$), S0 ($n = 27$), SI/II ($n = 69$) and SIII/IV ($n = 54$)). We focused on species–metabolite pairs for which the abundances were greater in MP or S0 samples compared to the healthy control samples ($P < 0.005$; one-sided Mann–Whitney *U* test) with a correlation coefficient greater or equal to 0.6 ($P < 0.005$) in MP or S0. This process left us with 169 pairs. Among these, selecting pairs with species abundances higher than 10^{-4} left us with only one pair in MP—that is, *B. wadsworthia* and DCA.

Statistical analysis. Abundances of each species, KO gene or metabolite, as well as bacterial replication rates were determined to be significantly elevated or depleted in each of the stages (MP, S0, SI/II and SIII/SIV) by a pairwise comparison with the healthy controls using one-sided Mann-Whitely *U* tests (Supplementary Table 15). $P < 0.005$ was considered statistically significant. In addition, a Benjamini–Hochberg false-discovery rate-corrected P value (*q* value) was estimated.

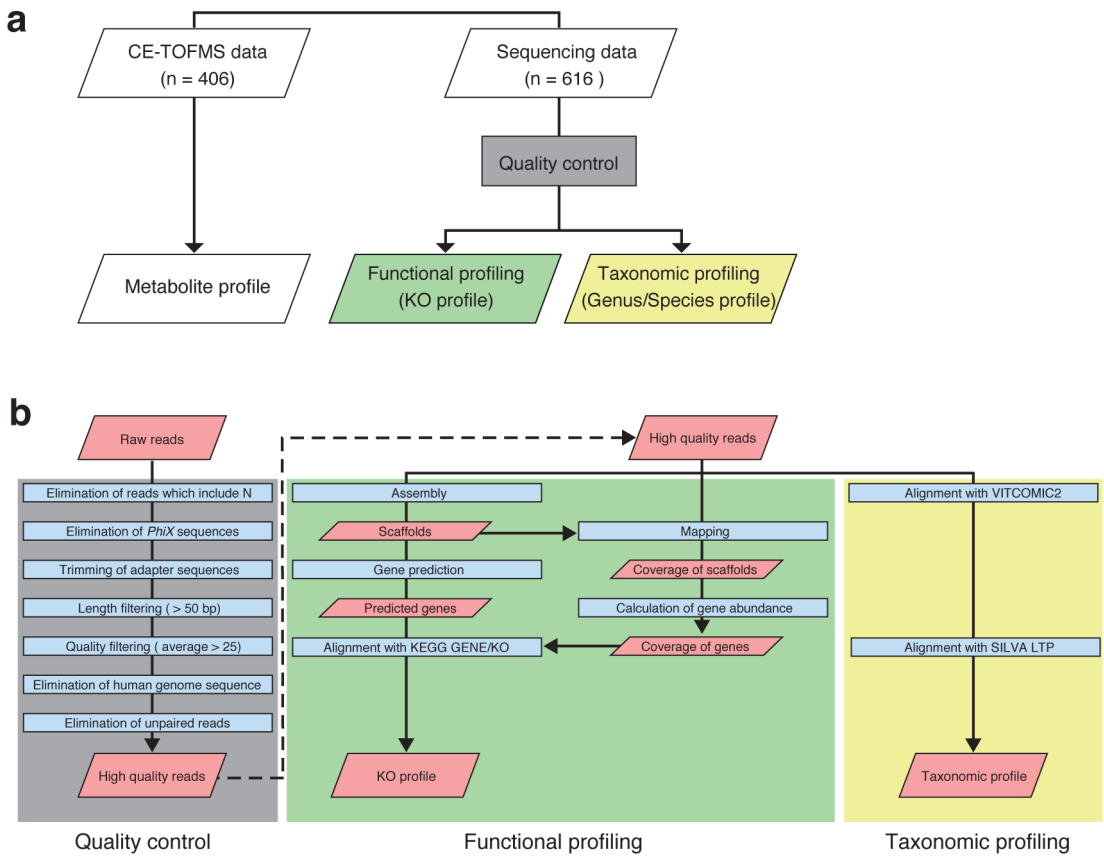
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

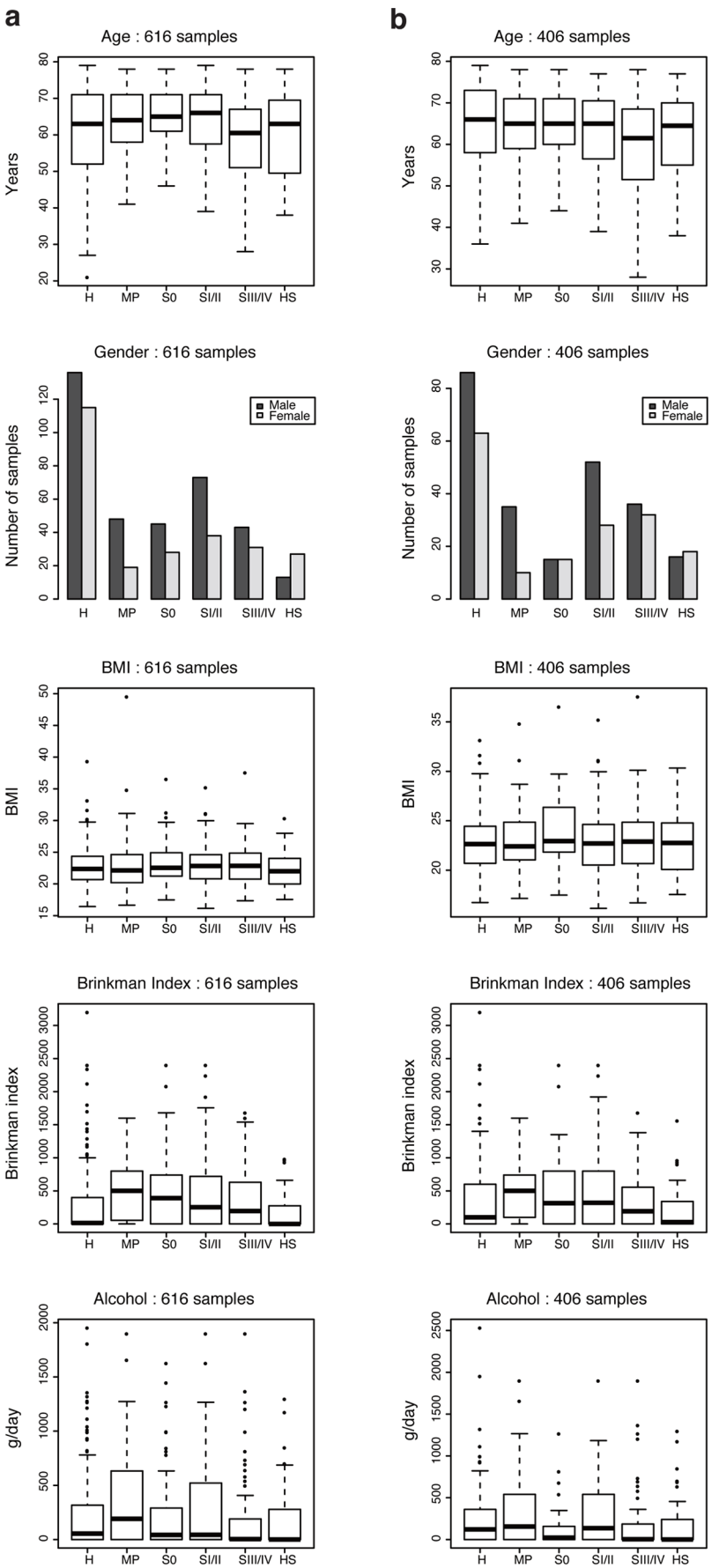
The raw sequencing data reported in this paper have been deposited in DDBJ Sequence Read Archive (DRA) as DRA006684 and DRA008156.

References

41. Nishimoto, Y. et al. High stability of faecal microbiome composition in guanidine thiocyanate solution at room temperature and robustness during colonoscopy. *Gut* **65**, 1574–1575 (2016).
42. Tsugane, S. & Sawada, N. The JPHC study: design and some findings on the typical Japanese diet. *Jpn J. Clin. Oncol.* **44**, 777–782 (2014).
43. Furet, J. P. et al. Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. *FEMS Microbiol. Ecol.* **68**, 351–362 (2009).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
46. Mori, H., Maruyama, T., Yano, M., Yamada, T. & Kurokawa, K. VITCOMIC2: visualization tool for the phylogenetic composition of microbial communities based on 16S rRNA gene amplicons and metagenomic shotgun sequencing. *BMC Syst. Biol.* **12**, 30 (2018).
47. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
48. Yarza, P. et al. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Syst. Appl. Microbiol.* **33**, 291–299 (2010).
49. Yarza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
50. Sunagawa, S. et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
51. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
52. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).
53. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).
54. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
55. Besemer, J. & Borodovsky, M. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* **27**, 3911–3920 (1999).
56. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
57. Soga, T. et al. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* **2**, 488–494 (2003).
58. Mishima, E. et al. Evaluation of the impact of gut microbiota on uremic solute accumulation by a CE-TOFMS-based metabolomics approach. *Kidney Int.* **92**, 634–645 (2017).
59. Tibshirani, R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
60. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

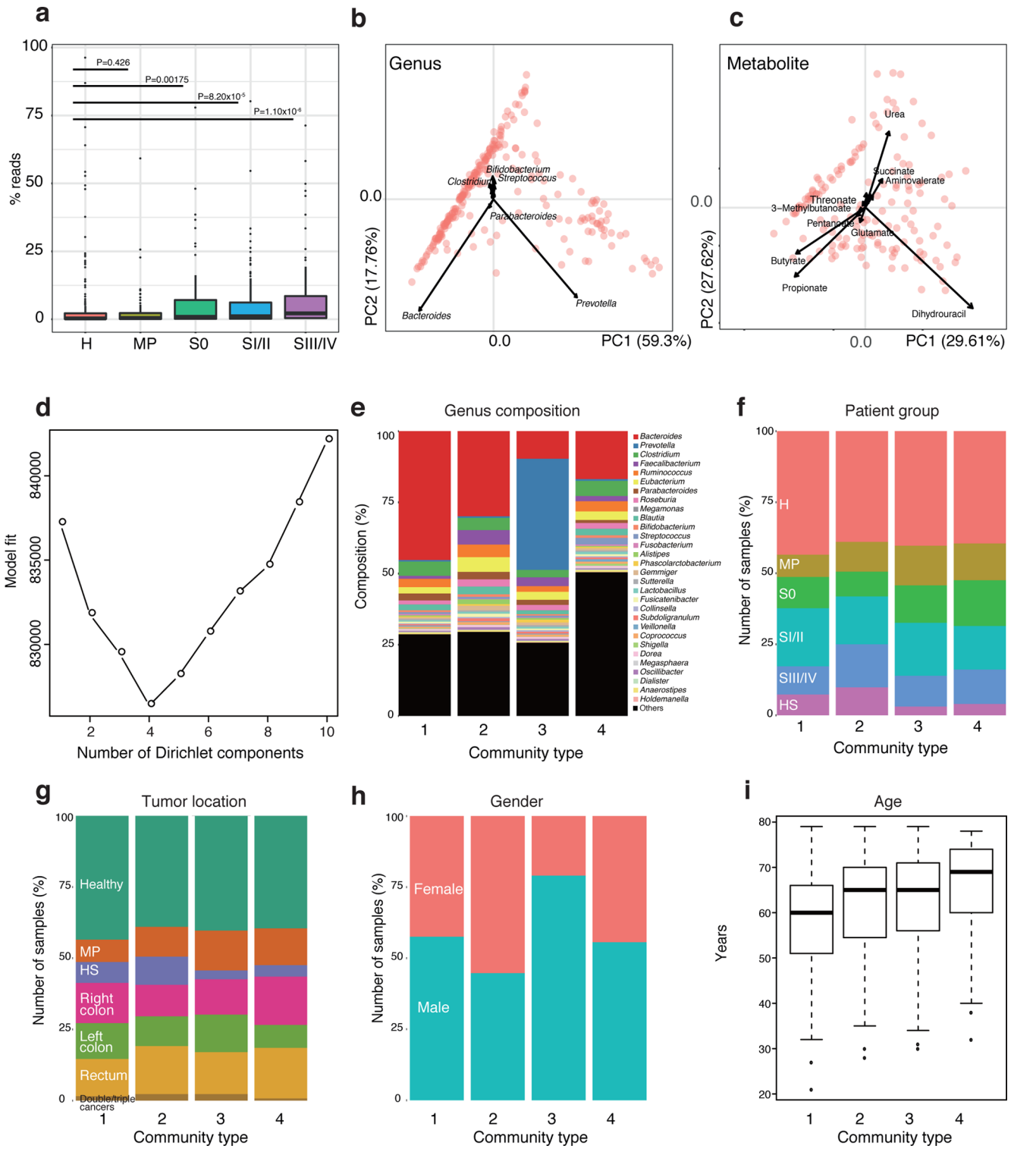


Extended Data Fig. 1 | Overview of the study and metagenomic analysis pipeline. a, Overview of the study. Fecal samples from 616 subjects were used to collect whole-genome shotgun sequencing data, from which functional and taxonomic profiles were generated. Fecal samples from 406 subjects were used to perform CE-TOFMS analysis to generate metabolite profiles. Samples from 347 subjects were available for both the sequencing and CE-TOFMS data analyses. KO, KEGG orthology. **b**, Flow chart of the pipeline used for the metagenomics analysis. Our metagenomics pipeline consists of three parts, quality control, functional profiling and taxonomic profiling, in which raw reads first undergo a quality control check and are then used to run several analytical steps to finally generate the KEGG orthology gene-based functional and taxonomic profiles.

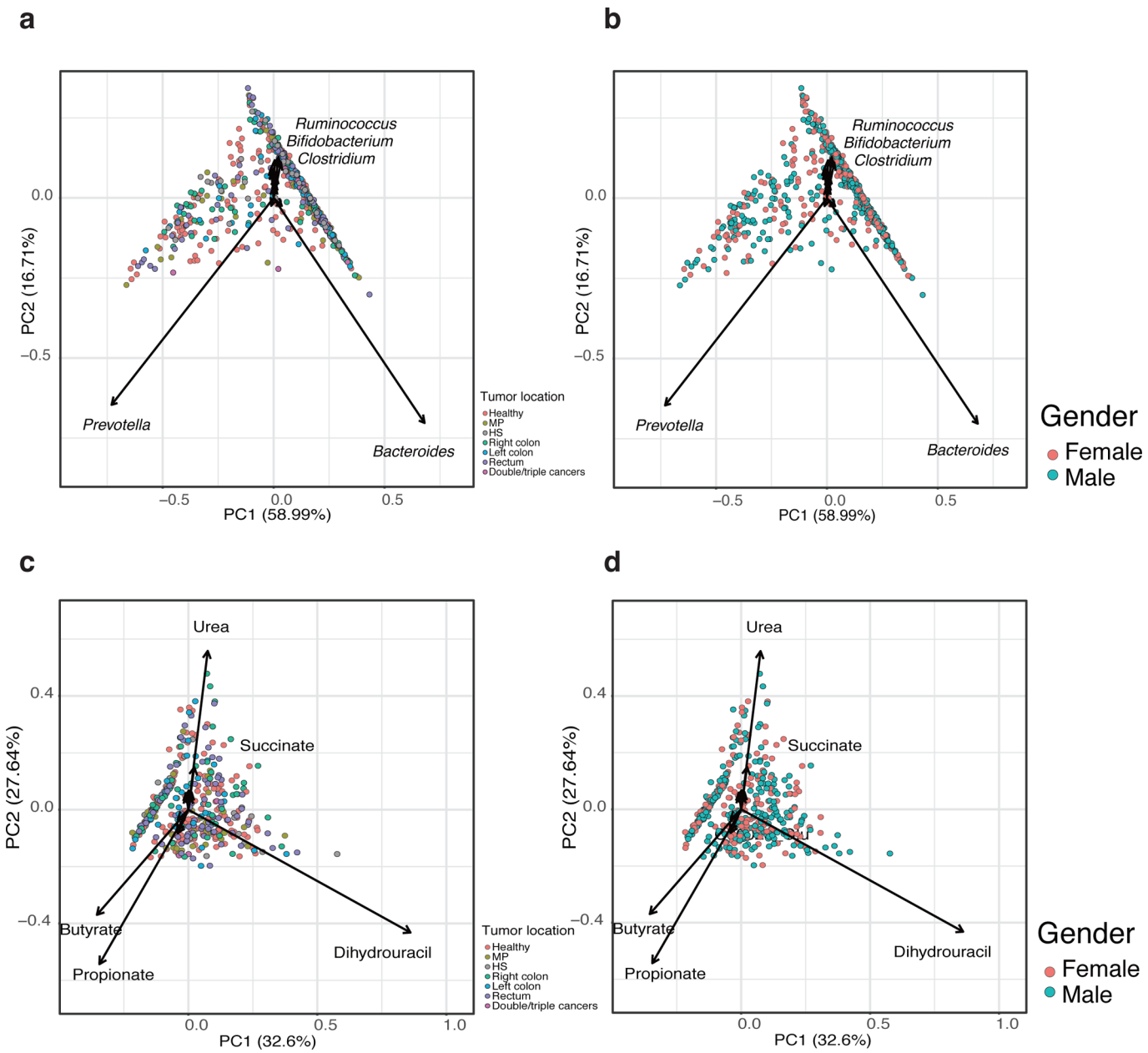


Extended Data Fig. 2 | see next page for caption.

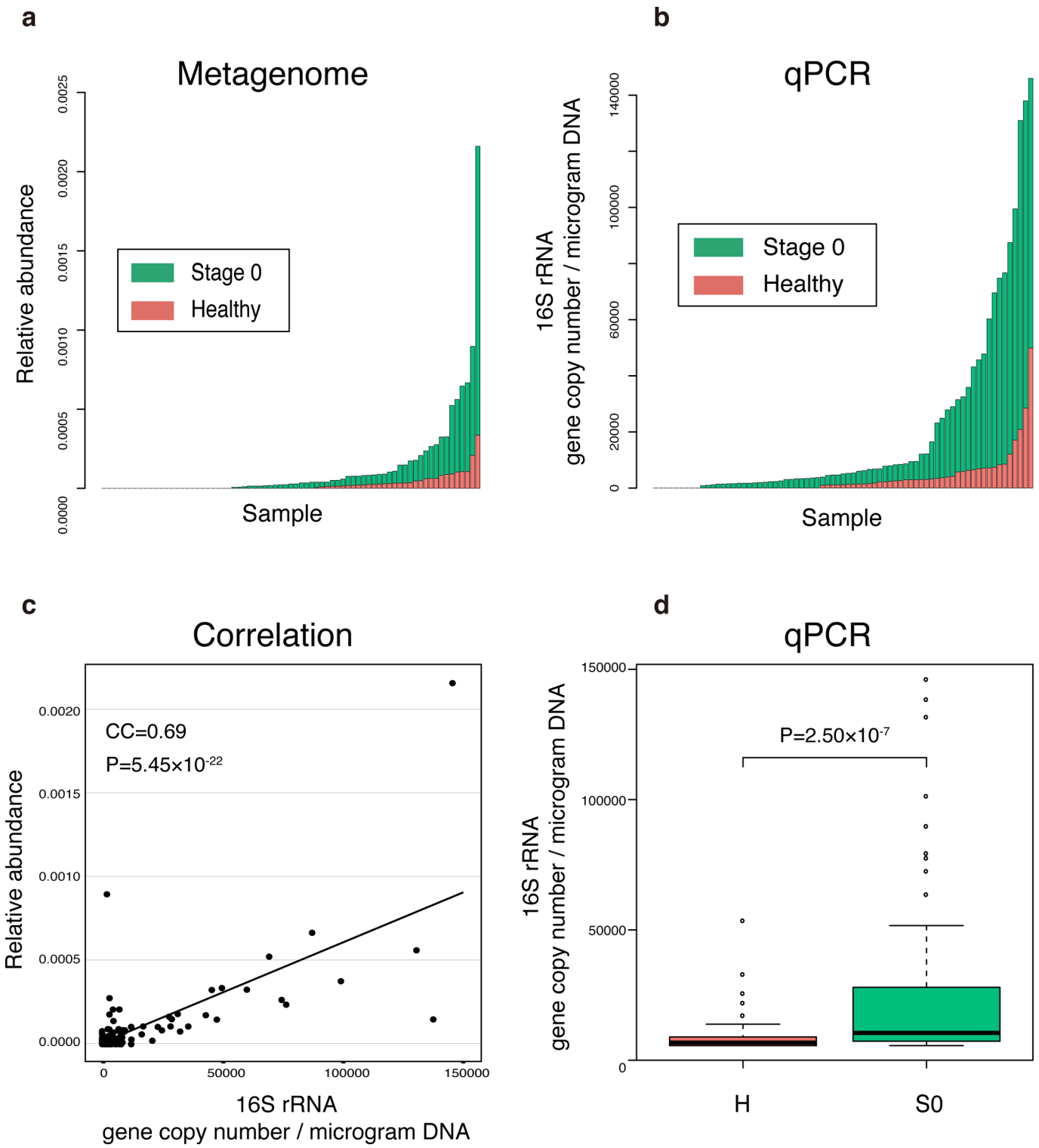
Extended Data Fig. 2 | Clinical information for the study subjects. a,**b**, Distribution of age, gender, BMI, Brinkman index and alcohol consumption in 616 subjects with metagenome data (**a**) and 406 subjects with metabolome data (**b**). The boxes represent 25th–75th percentiles, black lines indicate the median and whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers.



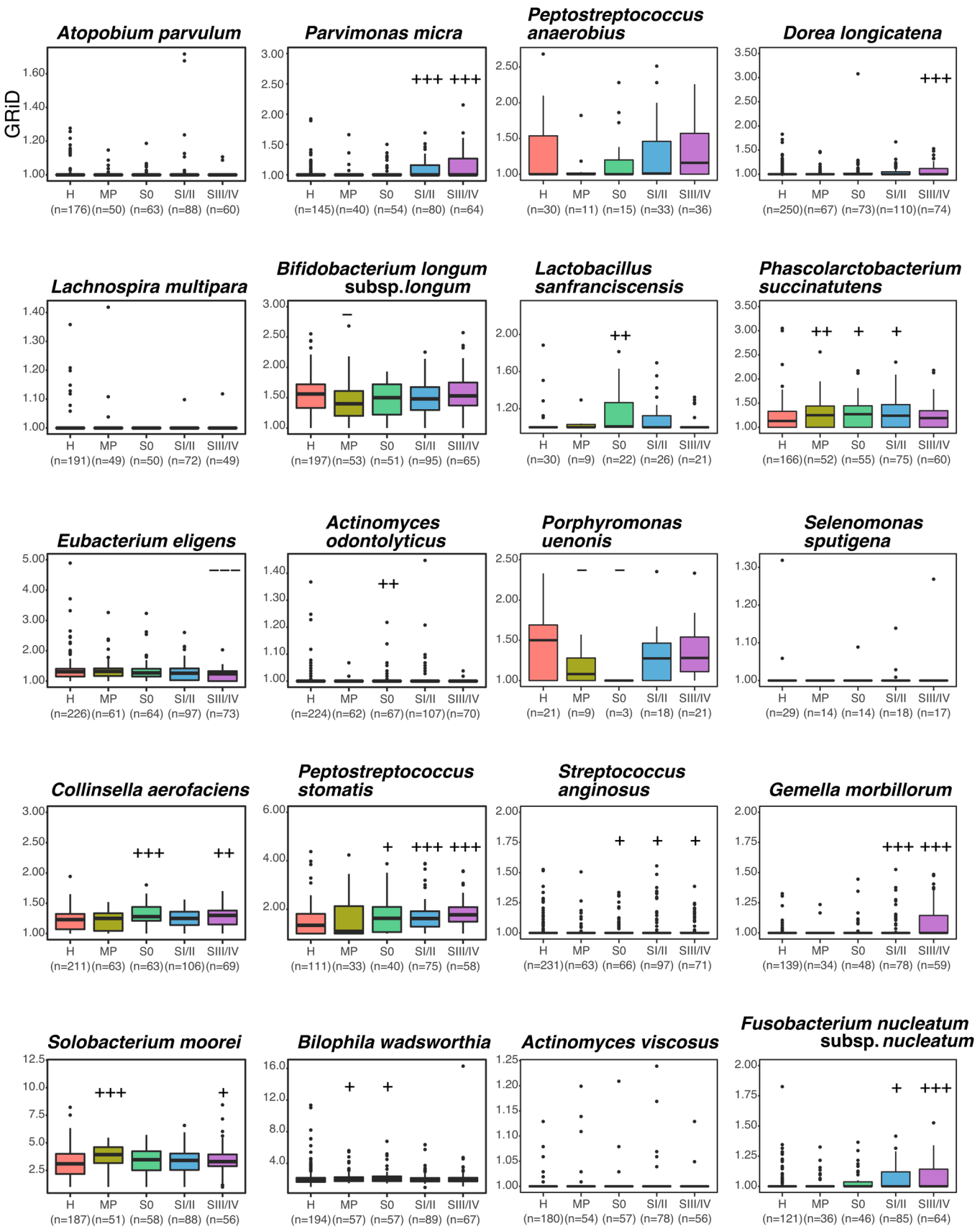
Extended Data Fig. 3 | Microbial community structure and human genome content in fecal metagenomes. a, Change in the fraction of human genome reads out of the total number of raw reads along with colorectal cancer progression. Significant increases ($P < 0.005$; one-sided Mann–Whitney U test) in the human genome ratio (percentage reads mapped to the human genome) in feces in the MP ($n = 67$), S0 ($n = 73$), SI/II ($n = 111$) and SIII/IV ($n = 74$) groups are evident compared to the healthy group (H, $n = 251$). The boxes represent 25–75th percentiles, black lines indicate the median, vertical lines show maximum values within 1.5× the interquartile range and dots indicate outliers beyond the 1.5× interquartile range. **b**,**c**, PCAs of genera ($n = 251$) (**b**) and metabolites ($n = 149$) (**c**) in healthy controls. **d**, Fitting to the Dirichlet multinomial mixture model indicated optimal classification of fecal metagenomes ($n = 616$) into four community types. **e**, Composition of top 30 genera in each of the four community types. **f**–**i**, Distribution of stages (healthy, MP, S0, SI/II, SIII/IV and HS) (**f**), tumor locations (right colon, left colon, rectum and double or triple cancers) (**g**), gender (**h**) and age (**i**) in each of the four community types. Patient distribution of the community types are as follows: community type 1, $n = 191$; community type 2, $n = 172$; community type 3, $n = 129$; community type 4, $n = 124$. The boxes in **i** represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers.



Extended Data Fig. 4 | Distributions of tumor locations and gender in the overall structures of metagenomes and metabolomes. a,**b**, PCA plots of genus profiles ($n = 616$) grouped by tumor location (**a**) and gender (**b**). **c**,**d**, PCA plots of metabolite profiles ($n = 406$) grouped by tumor location (**c**) and gender (**d**).

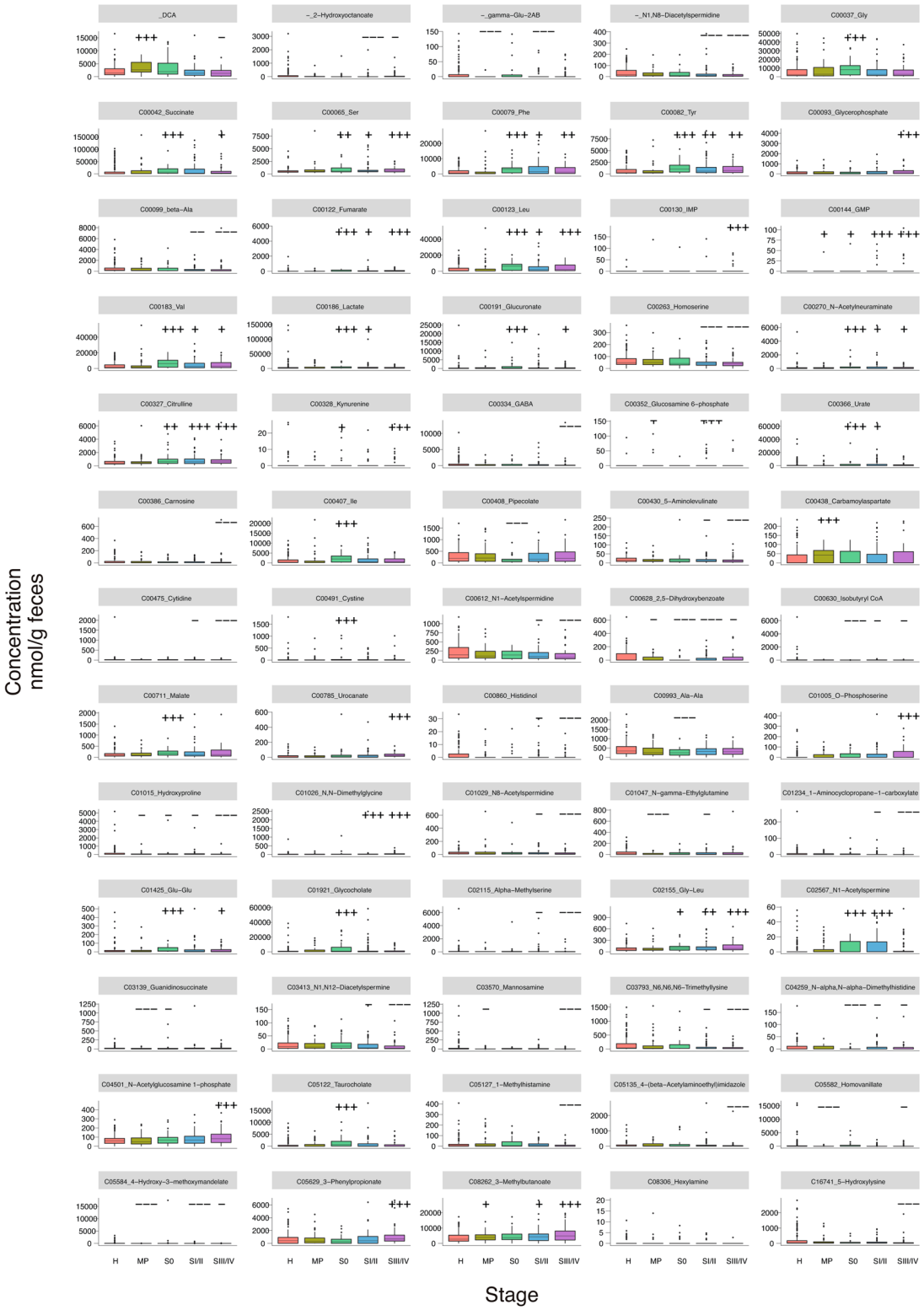


Extended Data Fig. 5 | Comparison of *A. parvulum* abundance in metagenomic and qPCR analyses. a,**b**, Abundance of *A. parvulum* estimated with whole-genome shotgun metagenomic sequence data (**a**) and by qPCR using the 16S rRNA gene copy number (**b**) in samples from 73 patients with S0 CRCs (green) and 73 healthy controls (red). **c**, Spearman's correlation coefficient of *A. parvulum* abundances between the two methods was calculated with *P*-value computation using asymptotic approximation. **d**, qPCR demonstrated statistically significant differences in the gene number of *A. parvulum* between the healthy controls and patients with S0 CRCs (one-sided Mann–Whitney *U* test). The boxes represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers.

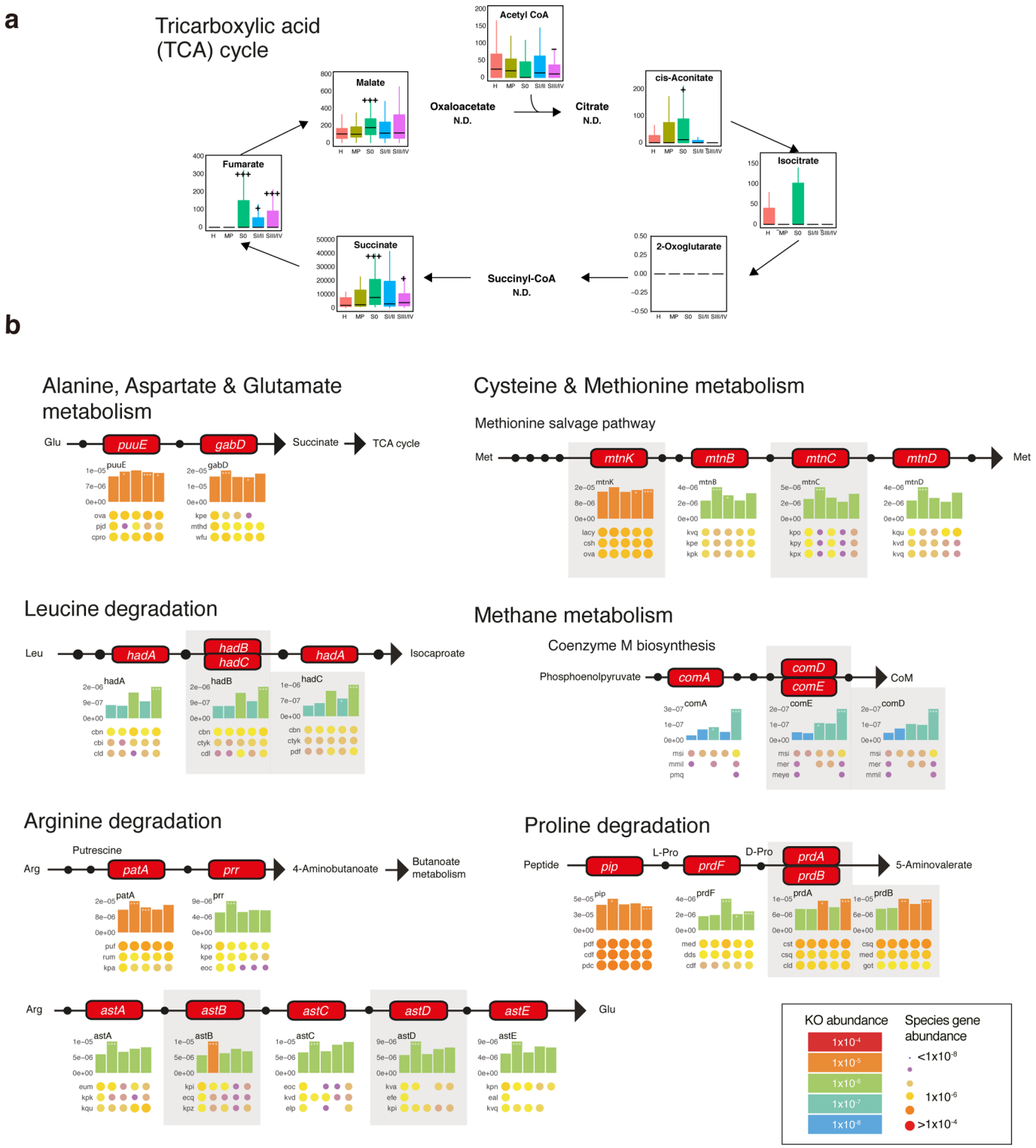


Extended Data Fig. 6 | see next page for caption.

Extended Data Fig. 6 | Replication rates were estimated using GRiD. Replication rates were plotted for the 20 species shown in Fig. 2. The *y* axis (GRiD) is defined as the ratio between coverage at the peak (*ori*) and trough (*ter*) for the reference bacterial genome. Samples that had sufficient coverage are plotted (coverage > 0.2) for mapping against the reference genomes. Sample numbers varied with dependent species and are indicated in parentheses. *P* values were calculated using one-sided Mann–Whitney *U* tests for each of the stages (MP, S0, SI/II, SIII/IV) and were compared to the healthy controls. Significant elevation or depletion are denoted as follows: +++, elevation with $P < 0.005$; ++, elevation with $P < 0.01$; +, elevation with $P < 0.05$; −−−, depletion with $P < 0.005$; −−, depletion with $P < 0.01$; −, depletion at $P < 0.05$. The boxes represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers.

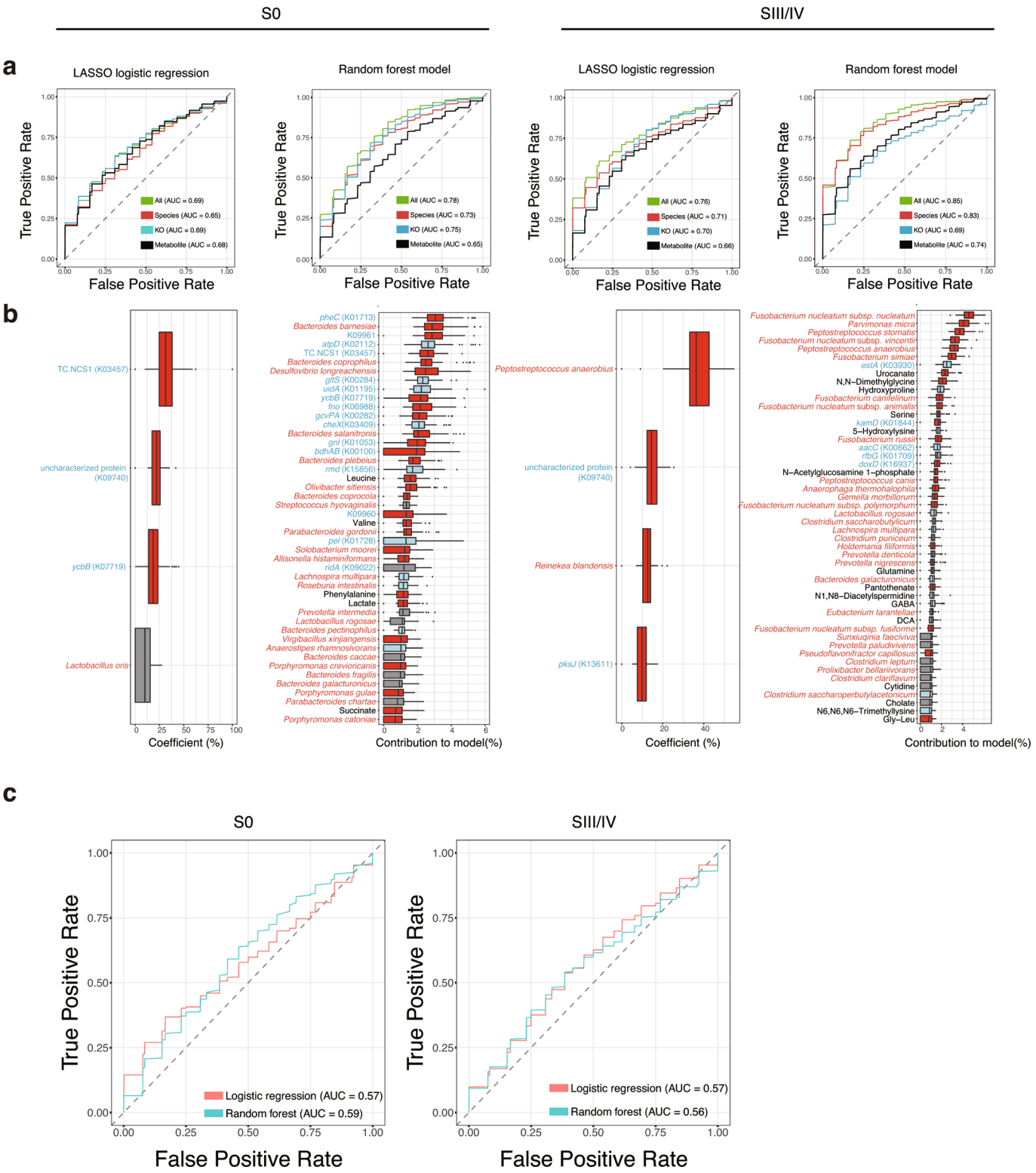


Extended Data Fig. 7 | Metabolite changes in colorectal cancer stages. Sixty-five metabolites with statistically significant ($P < 0.005$; one-sided Mann–Whitney U test) differences for any of the stages (MP, $n = 45$; S0, $n = 30$; SI/II, $n = 80$; SIII/IV, $n = 68$ compared to the healthy controls ($n = 149$) on capillary electrophoresis time of flight mass spectrometer (CE-TOFMS) analysis. Significant changes (elevation and depletion) are denoted as follows: +++, elevation with $P < 0.005$; ++, elevation with $P < 0.01$; +, elevation with $P < 0.05$; −−−, depletion with $P < 0.005$; −−, depletion with $P < 0.01$; −, depletion at $P < 0.05$. The boxes represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers.



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

Extended Data Fig. 8 | Metabolomic changes in the tricarboxylic acid (TCA) pathways and metagenomic changes in amino acid metabolism and other representative pathways. a, Quantified levels of metabolites involved in the tricarboxylic acid (TCA) pathway. The levels of the three TCA metabolites, succinate, fumarate and malate, were significantly higher $P < 0.005$; one-sized Mann–Whitney U test) in S0 samples (and SIII/IV samples for fumarate) (+++, $P < 0.005$; ++, $P < 0.01$; +, $P < 0.05$) compared to healthy control samples. It is uncertain what is the cause of accumulation of succinate, fumarate and malate in the feces of patients with early colorectal cancer, despite extremely low concentrations of other TCA intermediates such as 2-oxoglutarate. It is known that some bacteria synthesize ATP using a reverse reaction of succinate dehydrogenase and produce succinate as a byproduct, as part of fumarate respiration, in which fumarate rather than molecular oxygen is used as electron acceptor. The boxes represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers. The concentration is shown on the y axis (nmol g^{-1}). Healthy ($n = 127$), MP ($n = 45$), S0 ($n = 30$), SI/II ($n = 80$), SIII/IV ($n = 68$). N.D., not detected and/or not determined. **b**, Pathway modules for metabolism types omitted from Fig. 3b. The pathway modules are modified from KEGG pathway maps 'Alanine, aspartate and glutamate metabolism', 'Cysteine and methionine metabolism', 'Methane metabolism' and 'Arginine and proline metabolism'. 'Leucine degradation' is constructed based on leucine metabolism of *Clostridium difficile*, as the bacterial map is not available in KEGG. For each KO gene, bar plots show KO gene abundances averaged over samples within each of the five groups, healthy ($n = 251$), MP ($n = 67$), S0 ($n = 73$), SI/II ($n = 111$) and SIII/IV ($n = 74$) in order of left to right, and colored according to the order of the values. Each KO gene is composed of organism genes represented by circles. The sizes and colors of the circles are proportional to the relative abundances of the organism genes. Organism genes are grouped into one row and indicated by the organism name. The three most abundant organisms in the healthy controls are shown using three letter codes (for example, ova for *Oscillibacter valericigenes*, kpe for *Klebsiella pneumoniae* 342). Abbreviations for other organism names can be found in Supplementary Table 4. Gene numbers linked to each of the genes are listed in Supplementary Table 5. Dots in each pathway represent intermediate metabolites. The colors of the boxes of pathway components are marked in red for significant elevation ($P < 0.005$; one-sided Mann–Whitney U test) for any of the stages (MP, S0, SI/II and SIII/IV) compared to the healthy controls.



Extended Data Fig. 9 | see next page for caption.

Extended Data Fig. 9 | Metagenomic and metabolomic potentials as diagnostic markers for early (S0) and advanced (SIII/IV) CRCs. a, ROC curves as performance evaluation for LASSO logistic regression and random-forest models and to distinguish samples from patients with S0 (left two panels) and SIII/IV (right two panels) CRCs from samples of healthy controls. Models were designed based on species (red), KO genes (blue) or metabolites (black) individually, or a combination (green) of the three features. For the S0 classification, the species-based model used 29 species, the KO gene-based model used 16 KO genes and the metabolite-based model used 24 metabolites. For SIII/IV classification, the species-based model used 55 species, the KO gene-based model used 5 KO genes and the metabolite-based model used 62 metabolites. In the combination models, species, KO gene and metabolite features were selected from the individual models. Classification accuracy was evaluated on the AUC using 10 randomized 10-fold cross-validation testing. For the LASSO logistic regression models, all features that satisfied the abundance thresholds were used to construct both the individual models and the combination model. The discriminant features among all are shown. **b**, Discriminant features identified from LASSO logistic regression classifiers and random-forest classifiers for distinguishing S0 ($n = 27$) and SIII/IV ($n = 54$) cases from the healthy controls ($n = 127$). The colors of box plots represent significant increases (red) and decreases (light blue) in each group, compared to the healthy controls ($P < 0.005$; one-sided Mann–Whitney U test). Dark gray boxes indicate features without statistical significance. The x axis shows the percentage contribution of the features to the model in each test (see Methods). The boxes represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers. **c**, Analysis of potential confounding factors that might affect metagenomic and metabolomic classifiers. We analyzed AUCs for factors such as age, gender, BMI, smoking and alcohol exposure. Smoking and alcohol values are indicated as the Brinkman index and the amount of alcohol consumption, respectively. Whereas the gender of the patient and the Brinkman index significantly differed between groups (Supplementary Table 1), neither the random-forest nor the logistic regression model achieved high accuracy.

Extended Data Fig. 10 | Correlation between dietary intake and the gut microbiome. *Fusobacterium* spp., *Akkermansia muciniphila* and sulfidogenic bacteria (*B. wadsworthia* and *Pyramidobacter piscolens*), which were previously reported to exhibit relationships with dietary intake, were examined for Spearman's correlation coefficients with dietary fiber (water-soluble, insoluble and total dietary fiber), dietary protein intake (protein and meat), dietary fat (lipid), dietary calcium (calcium), serving of dairy products (milk) and energy intake (energy). +++, correlation with $P < 0.005$; +, correlation with $P < 0.05$. Samples that lack dietary data are omitted from the computation of correlation coefficients. Healthy, $n = 242$; MP, $n = 67$; S0, $n = 72$; SI/II, $n = 109$; SIII/IV, $n = 71$.

nature research

Corresponding author(s): Shinichi Yachida, Takuji Yamada

Last updated by author(s): Mar 21, 2019

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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 has been used for data collection, all the software used is reported in the paper.
Data analysis	Bowtie 2 (version 2.2.9), cutadapt (version 1.9.1), BLAST+ (version 2.2.30), mOTU profiler (version 1), MetaPhlAn2 (version 2.6.0), R (version 3.4.3), DirichletMultinomial (version 1.16.0), GraPhlAn (version 0.9.7), GRiD (version 1.2), yEd Graph Editor (version 3.18.11), IDBA_UD (version 1.1.1), MetaGeneMark (version 3.26), DIAMOND (version 0.9.10), scikit-learn (version 0.19.1).

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Nucleotide sequences are available in DDBJ Sequenced Read Archive (DRA) as DRA006684 and DRA008156.

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	No sample size calculation was performed because the present study is observational, not interventional.
Data exclusions	Details of the exclusion criteria are reported in the paper. Samples not passing quality-control have been excluded. For metabolomic analysis, concentrations below the detection limit were substituted with zero, and metabolites whose levels were below the detection limit in all the samples were excluded.
Replication	Abundances of species obtained from metagenomic analysis were validated using quantitative PCR (qPCR). The external validation was not performed.
Randomization	Not applicable for this observational study.
Blinding	Blinding was not possible because statistical analyses depended on information about cancer status.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	A total of 616 (358 males and 258 females, aged between 21 and 79 years old, 61.8 on average) individuals with metagenomic data, and a total of 406 (240 males and 166 females, aged between 28 and 79 years old, 62.9 on average) individuals with metabolomic data were enrolled, of which 347 individuals were covered both with metagenomic and metabolomic data (Supplementary Table S1).
Recruitment	This study was conducted with subjects undergoing total colonoscopy in the National Cancer Center Hospital, Tokyo, Japan, who were explained about this study and consent to it. Selection bias (e.g., referral bias) cannot be ruled out because this is not a completely randomized sampling selection.
Ethics oversight	The samples and clinical information used in this study were obtained under conditions of informed consent and with approval of the institutional review boards of each participating institute (National Cancer Center, 2013-244; Tokyo Institute of Technology, 2014018).

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