ORIGINAL ARTICLE

Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer

Jun Yu,[1†] Qiang Feng,[2,3†] Sunny Hei Wong,[1†] Dongya Zhang,[2†] Qiao yi Liang,[1†] Youwen Qin,[2] Longqing Tang,[2] Hui Zhao,[2] Jan Stenvang,[4] Yanli Li,[2] Xiaokai Wang,[2] Xiaoqiang Xu,[2] Ning Chen,[2] William Ka Kei Wu,[1] Jumana Al-Aama,[2,5] Hans Jørgen Nielsen,[6] Pia Kiilerich,[3] Benjamin Anderschou Holbech Jensen,[3] Tung On Yau,[1] Zhou Lan,[2] Huijue Jia,[2] Junhua Li,[2] Liang Xiao,[2] Thomas Yuen Tung Lam,[1] Siew Chien Ng,[1] Alfred Sze-Lok Cheng,[1] Vincent Wai-Sun Wong,[1] Francis Ka Leung Chan,[1] Xun Xu,[2] Huanming Yang,[2] Lise Madsen,[2,3,7] Christian Datz,[8] Herbert Tilg,[9] Jian Wang,[2] Nils Brünner,[2,4] Karsten Kristiansen,[2,3] Manimozhiyan Arumugam,[2,10] Joseph Jao-Yiu Sung,[1] Jun Wang[2,3,5,11]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2015-309800).

For numbered affiliations see end of article.

Correspondence to
Professor Jun Wang, Beijing Genomics Institute at Shenzhen, Shenzhen 518000, China; wangj@genomics.org.cn and Professor Joseph JY Sung, The Chinese University of Hong Kong, Hong Kong; jjysung@cuhk.edu.hk and Dr Manimozhiyan Arumugam, University of Copenhagen, 2200 Copenhagen, Denmark; arumugam@sund.ku.dk

JuY, QF, SHW, DZ, QL contributed equally.

Received 21 April 2015
Revised 26 August 2015
Accepted 1 September 2015
Published Online First 25 September 2015

CrossMark

To cite: Yu J, Feng Q, Wong SH, *et al*. *Gut* 2017;**66**:70–78.

ABSTRACT

Objective To evaluate the potential for diagnosing colorectal cancer (CRC) from faecal metagenomes.

Design We performed metagenome-wide association studies on faecal samples from 74 patients with CRC and 54 controls from China, and validated the results in 16 patients and 24 controls from Denmark. We further validated the biomarkers in two published cohorts from France and Austria. Finally, we employed targeted quantitative PCR (qPCR) assays to evaluate diagnostic potential of selected biomarkers in an independent Chinese cohort of 47 patients and 109 controls.

Results Besides confirming known associations of *Fusobacterium nucleatum* and *Peptostreptococcus stomatis* with CRC, we found significant associations with several species, including *Parvimonas micra* and *Solobacterium moorei*. We identified 20 microbial gene markers that differentiated CRC and control microbiomes, and validated 4 markers in the Danish cohort. In the French and Austrian cohorts, these four genes distinguished CRC metagenomes from controls with areas under the receiver-operating curve (AUC) of 0.72 and 0.77, respectively. qPCR measurements of two of these genes accurately classified patients with CRC in the independent Chinese cohort with AUC=0.84 and OR of 23. These genes were enriched in early-stage (I–II) patient microbiomes, highlighting the potential for using faecal metagenomic biomarkers for early diagnosis of CRC.

Conclusions We present the first metagenomic profiling study of CRC faecal microbiomes to discover and validate microbial biomarkers in ethnically different cohorts, and to independently validate selected biomarkers using an affordable clinically relevant technology. Our study thus takes a step further towards affordable non-invasive early diagnostic biomarkers for CRC from faecal samples.

Significance of this study

What is already known on this subject?

- Changes in the gut microbial composition are associated with colorectal cancer (CRC), but causality is yet to be established.
- *Fusobacterium nucleatum* potentiates intestinal tumorigenesis through recruitment of infiltrating immune cells and via activation of β-catenin signalling.
- Faecal microbiota holds promise for early non-invasive diagnosis of CRC.
- However, a simple and affordable targeted approach to diagnosing CRC from faecal samples is still lacking.

What are the new findings?

- Discovery of significant enrichment of novel species, including *Parvimonas micra* and *Solobacterium moorei*, and a strong co-occurrence network between them in the faecal microbiomes of patients with CRC.
- Identification of 20 gene markers that significantly differentiate CRC-associated and control microbiomes in a Chinese cohort, and trans-continental validation of four of them in a Danish cohort.
- Further validation of the four gene markers in published cohorts from the French and Austrian cohorts with areas under the receiver-operating curve (AUC) of 0.72 and 0.77.
- Quantitative PCR abundance of two gene markers (butyryl-CoA dehydrogenase from *F. nucleatum*, and RNA polymerase subunit β, *rpoB*, from *P. micra*) clearly separates CRC microbiomes from controls in an independent Chinese cohort consisting of 47 cases and 109 healthy controls, with AUC=0.84 and odds ratio of 23.

INTRODUCTION

Colorectal cancer (CRC), the third most common cancer in the world affecting >1.36 million people every year,[1] arises due to complex interactions

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Significance of this study

How might it impact on clinical practice in the foreseeable future?

- The four microbial gene markers shared between the Chinese, Danish, Austrian and French cohorts suggest that even though different populations may have different gut microbial community structures, signatures of CRC-associated microbial dysbiosis could have universal features.
- Our study takes a step further towards affordable early diagnosis of CRC by targeted analysis of metagenomic biomarkers in faecal samples.

between genetic, lifestyle and environmental factors. Despite massive efforts in whole-genome sequencing and genome-wide association studies, genetic factors only explain a small proportion of disease variance[2]—heritability may account for up to 35% all CRCs,[3] but only about 5% of cancers occur in the setting of a known genetic predisposition syndrome.[4] These findings support lifestyle and environment as additional major disease determinants.

Emerging evidence indicates that microbial dysbiosis in the human gut may be an important environmental factor in CRC. Early evidence for gut microbial contribution to CRC pathogenesis came from *Apc*$^{min/+}$ mice, a genetic mouse model of CRC, where mice housed in germ-free conditions showed a reduction of tumour formation in the intestine compared with mice housed in specific pathogen-free conditions.[5] Further studies have suggested that several bacteria, including *Bacteroides fragilis* and a strain of *Escherichia coli*, may promote colorectal carcinogenesis.[6–11] In humans, bacterial culture-based studies have reported associations between CRC and clinical infections by specific bacteria such as *Streptococcus bovis*[12] and *Clostridium septicum*.[13] Additionally, culture-free 16S ribosomal RNA sequencing studies have associated faecal microbial composition with CRC.[14–16] Independent studies have identified *Fusobacterium nucleatum* to be more abundant in human CRC tissues,[17 18] and follow-up studies showed that *F. nucleatum* potentiates intestinal tumorigenesis through recruitment of infiltrating immune cells[19] and by modulating β-catenin signalling.[20] Two recent studies investigated gut microbial dysbiosis in patients with CRC[21 22] and reported diagnostic potential using metagenomic sequencing. These promising results are still far from directly translating to diagnostic tests for CRC, as a simple and affordable targeted approach to diagnosing CRC from faecal samples is still lacking.

Here we present the first study that (i) uses deep metagenomic profiling of CRC faecal microbiomes to discover and validate microbial gene biomarkers in ethnically different cohorts, and (ii) independently validates them using an affordable technology that can translate to clinical practice.

MATERIALS AND METHODS

Sample collection and DNA preparation

Cohorts C1 and C2 were from Hong Kong, China. C1 (see online supplementary table S1) comprised 128 individuals: 74 patients with CRC (15 stage I, 21 stage II, 34 stage III and 4 stage IV; median age 67 years; 26 were females) and 54 controls (median age 62 years; 21 were females). C2 (see online supplementary table S16) comprised 156 individuals: 47 patients with CRC (4 stage I, 24 stage II, 15 stage III and 4 stage IV; median age 69 years; 22 were females) and 109 controls (median age 58 years; 69 were females). Cohort D from Copenhagen, Denmark (see online supplementary table S18), comprised 40 individuals: patients with CRC (n=16; 1 stage I, 9 stage II, 5 stage III and 1 stage IV; median age 67.5 years; 6 were females) and controls (n=24; median age 65.5 years; 17 were females). Cancer staging in all three cohorts was performed using the tumour, node, metastasis staging system[23] maintained by the American Joint Committee on Cancer and the International Union for Cancer Control. Stool samples were collected by individuals at home, followed by immediate freezing at −20°C. DNA from Chinese samples was extracted using Qiagen QIAamp DNA Stool Mini Kit (Qiagen) according to manufacturer's instructions. DNA from Danish samples was extracted using previously published method.[24] For comprehensive description of sample collection and DNA extraction as well as ethical committee approval numbers, see online supplementary methods.

Metagenomic sequencing and annotation

Metagenomic sequencing using Illumina HiSeq 2000 platform, generating gene profiles using gene catalogue, constructing metagenomic linkage groups (MLGs), generating Kyoto Encyclopedia of Genes and Genomes (KEGG) ortholog, module and pathway profiles, were all done using previously published methods.[25] Species-level molecular operational taxonomic units (mOTUs) were obtained using mOTU profiling software.[26] Reads were mapped to the Integrated Microbial Genome (IMG) reference database[27] (v400) to generate IMG species and IMG genus profiles. Genes of MLGs were mapped to the IMG database, and MLGs were annotated to an IMG genome when >50% of genes were mapped. MLG species were constructed by grouping MLGs using this annotation. For comprehensive description of these procedures, see online supplementary methods.

Data analysis

Permutational multivariate analysis of variance (PERMANOVA) was used to assess effects of different phenotypes on gene profiles. Enrichments of genes, KEGG features, mOTUs, IMG species and MLG species were calculated using Wilcoxon rank-sum tests. When appropriate, we adjusted for confounding effects of sample collection before/after colonoscopy: Wilcoxon rank-sum tests were performed using 'colonoscopy before/after sampling' as a stratifying factor using COIN package in R, and ORs were estimated using Mantel–Haenszel test after stratifying by 'colonoscopy before/after sampling'. We controlled for multiple testing with Benjamini–Hochberg false discovery rate (FDR). Minimum-redundancy maximum-relevancy (mRMR) feature selection method[28] was used to select optimal gene markers, which were then used in constructing a CRC index. Co-occurrence networks were constructed using Spearman's correlation coefficient (>0.5 or <−0.5) and visualised in Cytoscape V.3.0.2. Metagenomic sequences from French (F) and Austrian (A) cohorts were downloaded from NCBI Short Read Archive using study identifiers ERP005534 and ERP008729, respectively. For comprehensive description of biodiversity analysis, rarefaction analysis, identification of CRC-associated genes/species, estimation of FDR, mRMR feature selection framework, definition and validation of CRC index, and receiver operator characteristic (ROC) analysis, see online supplementary methods.

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

Validation of gene markers by qPCR

Abundances of selected gene markers were estimated in stool samples using TaqMan probe-based quantitative PCR (qPCR). Primer and probe sequences were designed manually and then tested using Primer Express V.3.0 (Applied Biosystems, Foster City, California, USA) for determination of Tm, guanine-cytosine (GC) content and possible secondary structures. Each probe carried a 5′ reporter dye 6-carboxy fluorescein or 4,7,2′-trichloro-7′-phenyl-6-carboxyfluorescein and a 3′ quencher dye 6-carboxytetramethyl-rhodamine. Primers and hydrolysis probes were synthesised by Invitrogen (Carlsbad, California, USA). Nucleotide sequences of primers and probes are listed in online supplementary table S27. qPCR was performed on an ABI7500 Real-Time PCR System using TaqMan Universal PCR Master Mixreagent (Applied Biosystems). Universal 16S rDNA was used as internal control and abundance of gene markers were expressed as relative levels to 16S rDNA.

RESULTS

Dysbiosis in CRC gut microbiome

We recruited 128 individuals (74 patients with CRC and 54 control subjects) from China (cohort C1; see online supplementary table S1), performed metagenomic sequencing on their stool samples and generated 751 million metagenomic reads (5.86 million reads per individual on average; see online supplementary table S2) using Illumina HiSeq 2000 platform. Among the recorded metabolic parameters, elevated fasting blood glucose and reduced high-density lipoproteins showed significant associations with CRC status (Wilcoxon rank-sum test, $q=0.0014$ for both) agreeing with previous findings reporting them as risk factors.[29 30] We also observed that a significantly higher number of CRC patient samples were collected after colonoscopy than before (Fisher's exact test, $q=0.0165$; see online supplementary table S1). We adjusted for this as a confounding factor in subsequent analyses when appropriate (see section 'Materials and methods'). Rarefaction analysis using a previously published gut microbial gene catalogue consisting of 4 267 985 genes[25] showed a curve reaching plateau, suggesting that this catalogue covers most prevalent microbial genes present in cohort C1 (see online supplementary figure s1A). Therefore, we based subsequent analyses on mapping the metagenomic reads to this catalogue. CRC patient microbiomes exhibited reduced gene richness (see online supplementary figure 1A, B; Wilcoxon rank-sum test, $p<0.01$) and gene alpha diversity (Wilcoxon rank-sum tests on Shannon and Simpson indices: $p=0.075$ and 0.028, respectively; see online supplementary figure S1C,D and table S3). However, these differences exhibited $p>0.5$ after correcting for colonoscopy.

To ensure robust comparison of gene content among 128 metagenomes from cohort C1, we created a set of 2 110 489 genes that were present in at least 6 subjects and generated 128 gene abundance profiles using these 2.1 million genes. When we performed multivariate analysis using PERMANOVA on 17 different covariates, only CRC status and CRC stage were significantly associated with these gene profiles ($q<0.06$, all other factors: $q>0.27$; see online supplementary table S4). Thus, the data suggest an altered gene composition in CRC patient microbiomes that cannot be explained by other recorded factors. When we performed a principal component analysis (PCA) based on gene profiles, the first and fifth principal components, which explained 6.6% and 3.2% of total variance, respectively, were associated with CRC status (Wilcoxon rank-sum test, PC1: $p=0.029$; PC5: $p=1\times10^{-6}$; see online supplementary figure S2 and table S5). Together, these results suggest a state of dysbiosis of the gut microbiome in patients with CRC.

Gut microbial genes associated with CRC

We performed a metagenome-wide association study (MGWAS) to identify genes contributing to the altered gene composition in CRC. From 2.1 million genes, we identified 140 455 genes that were associated with disease status (Wilcoxon rank-sum test $p<0.01$ and FDR 11.03%; see online supplementary figure S3). Interestingly, CRC-enriched genes occurred less frequently and at lower abundance compared with control-enriched genes (see online supplementary figure S4), suggesting that microbial dysbiosis associated with CRC may not involve dominant species. Such patterns of frequency and occurrence have been observed in two earlier metagenomic case–control studies on type 2 diabetes[25] in Chinese individuals and CRC in Austrian individuals,[31] suggesting that this may be a common trend in disease-associated gut microbial dysbiosis.

We annotated the 140 455 genes using KEGG[32] functional database (V.59) to investigate whether certain microbial functions were associated with CRC. None of the KEGG pathways passed our stringent criteria (Wilcoxon rank-sum test, $q<0.05$; see online supplementary table S6), suggesting that bacterial metabolic pathways present in KEGG database may not be involved in CRC pathogenesis. However, two KEGG modules were enriched in CRC microbiomes: leucine degradation ($q=0.0148$) and guanine nucleotide biosynthesis ($q=0.0241$; see online supplementary table S6). Leucine stimulates both protein synthesis and degradation,[33 34] suggesting possible links between leucine metabolism and cancer. At the gene level, several KEGG orthologous groups showed significant associations with disease status (Wilcoxon rank-sum test, $q<0.05$; see online supplementary table S7).

Taxonomic alterations in CRC gut microbiomes

We examined taxonomic differences between CRC-associated and control microbiomes to identify microbial taxa contributing to the dysbiosis. For this, we used species profiles derived from three different methods—IMG species, species-level mOTUs and MLG species (see section 'Materials and methods')—as supporting evidence from multiple methods would strengthen an association. Our analysis identified 28 IMG species, 21 mOTUs and 85 MLG species that were significantly associated with CRC status after adjusting for colonoscopy as a confounding factor (Wilcoxon rank-sum test, $q<0.05$; see online supplementary table S8). *Eubacterium ventriosum* was consistently enriched in control microbiomes across all three methods (IMG: $q=0.002$; mOTU: $q=0.0049$; MLG: $q=3.33\times10^{-4}$). On the other hand, *Parvimonas micra* ($q<7.73\times10^{-6}$), *Solobacterium moorei* ($q<0.011$) and *F. nucleatum* ($q<0.00279$) were consistently enriched in CRC patient microbiomes across all three methods (figure 1A and online supplementary figure S5), while *Peptostreptococcus stomatis* ($q<7.73\times10^{-6}$) was enriched according to two methods. PERMANOVA analysis showed that only CRC status ($p\leq0.013$ from all three methods) and colonoscopy ($p=0.079$ from two methods) explained the quantitative variation in the three CRC-enriched species. All other non-CRC-specific factors could not explain the variation with statistical significance ($p>0.18$; see online supplementary table S9). *P. stomatis* has recently been shown to significantly associate with CRC,[22] and *S. moorei* has previously been associated with bacteraemia.[35] However, a highly significant enrichment of *P. micra*—an obligate anaerobic bacterium that can cause oral infections like *F. nucleatum*[36]—in CRC-associated microbiomes is a novel finding.

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

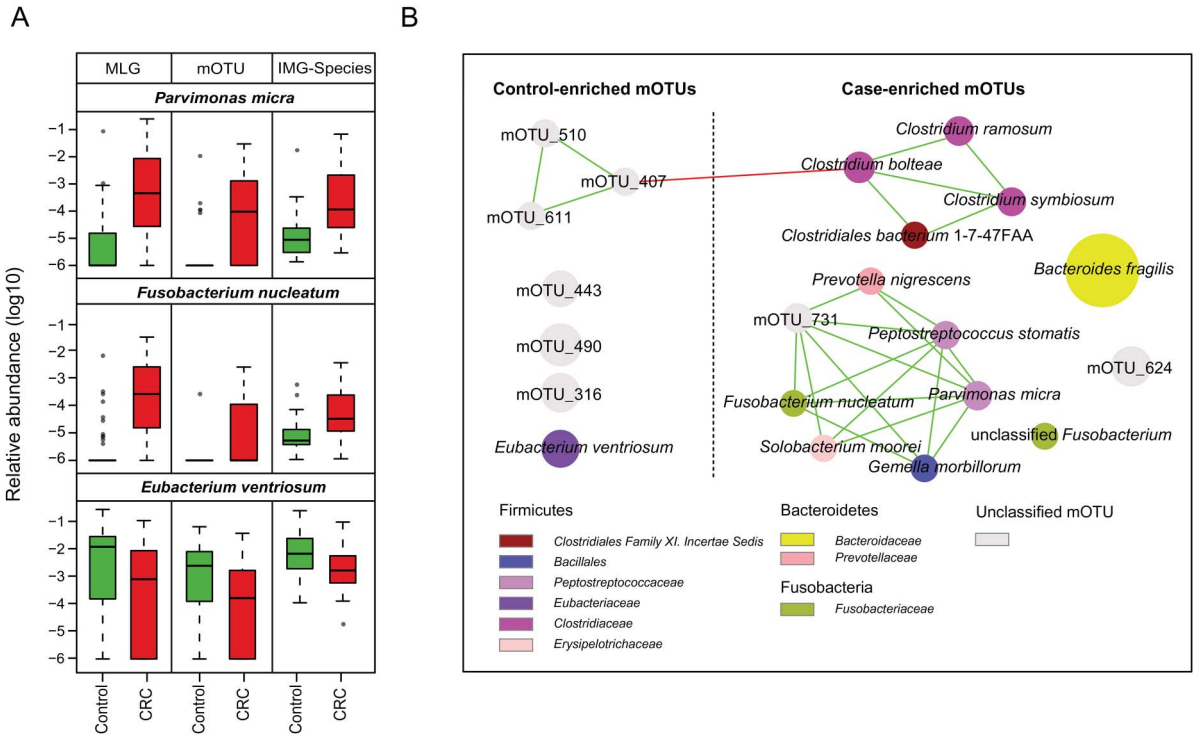


Figure 1 Species involved in gut microbial dysbiosis associated with colorectal cancer (CRC). (A) Differential relative abundance of two CRC-enriched and one control-enriched microbial species consistently identified using three different methods: metagenomic linkage group (MLG), molecular operational taxonomic unit (mOTU) and Integrated Microbial Genome (IMG) database. (B) A co-occurrence network deduced from relative abundance of 21 mOTUs significantly associated with CRC. Species are rearranged in two sides based on their enrichment in CRC or control microbiomes. Spearman correlation coefficient values below −0.5 (negative correlation) are indicated as red edges, and coefficient values above 0.5 (positive correlation) are indicated as green edges. Node size shows the average relative abundance for each species, and node colour shows their taxonomic annotation.

Species co-occurrence networks derived from pairwise correlations of species abundances showed a strong positive association between three oral pathogens: *P. micra*, *F. nucleatum* and *S. moorei* (figure 1B and online supplementary figure S6). Previous reports suggest that *P. micra* commonly occurs together with *F. nucleatum* in infected root canals, where they could account for up to 90% of the endodontic microbiome.[36] Given this, our results could suggest cooperation between these two species in CRC-associated gut environment.

Although several bacterial genera corresponding to the CRC-associated species identified earlier (including *Parvimonas*, *Fusobacterium*, *Solobacterium* and *Peptostreptococcus*) showed significant associations with CRC status (see online supplementary table S10), we observed some exceptions as well. While we

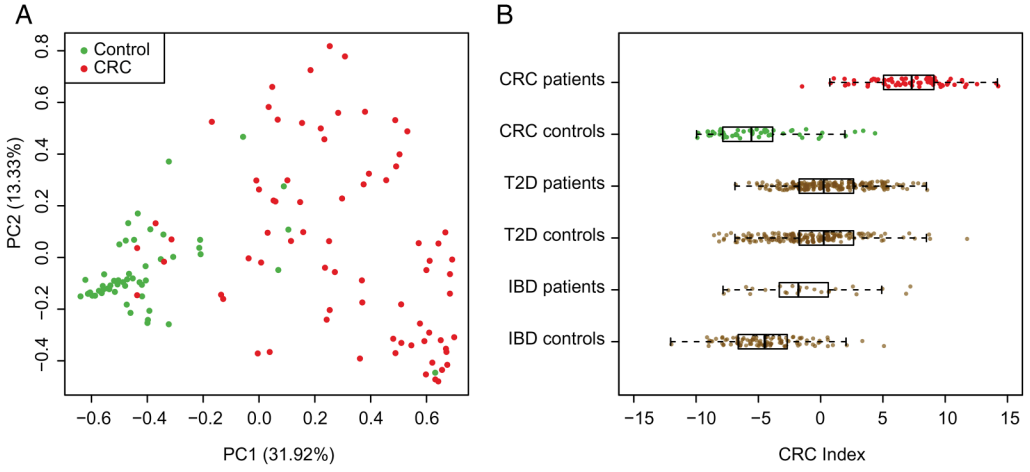


Figure 2 Discovering gut microbial gene markers associated with colorectal cancer (CRC). (A) Principal component analysis based on abundances of 20 gene markers separates CRC cases and control individuals in cohort C1. First and second principal components associate with CRC status (PC1 and PC2 explain 31.9% and 13.3% of variance, respectively). Compare this with online supplementary figure S2 based on 2.1 million genes, where no separation can be observed. (B) CRC index computed using a simple unweighed linear combination of log-abundance of 20 gene markers for patients with CRC (red) and control individuals (green) from this study, shown together with patients and control individuals (brown) from earlier studies on type 2 diabetes[25] and IBD.[38] CRC indices for CRC patient microbiomes are significantly different from the rest ($p<0.001$), suggesting that the 20 gene markers are CRC-specific. The box depicts the IQRs between the first and third quartiles, and the line inside denotes the median.

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

identified a significant over-representation of *B. fragilis* in patients with CRC (mOTU: q=0.0158; MLG: $q=3.02\times10^{-4}$; see online supplementary table S8), there was no association with *Bacteroides* genus. At the phylum level, only Fusobacteria and Basidiomycota were significantly enriched in CRC-associated microbiomes (q<0.0002; see online supplementary table S11).

In order to evaluate the predictive power of these taxonomic associations, we used random forest ensemble learning method[37] to identify 17 IMG species, 7 species-level mOTUs and 27 MLG species that were highly predictive of CRC status (see online supplementary table S12), with predictive power of 0.86, 0.89 and 0.96 in ROC analysis, respectively (see online supplementary figure S7). *P. micra* was identified as a key species from all three methods, while *F. nucleatum*, *P. stomatis* and *S. moorei* were identified from two out of three methods, providing further statistical support for their association with CRC status.

CRC biomarker discovery

We used the mRMR feature selection method[28] to identify potential CRC biomarkers from the 140 455 genes identified by MGWAS. First, to eliminate confounding effects of colonoscopy, we performed blocked independent Wilcoxon rank-sum tests on these genes with colonoscopy as a stratifying factor. This resulted in 102 514 genes at a significance level of p<0.01 (FDR ≤13%) and 24 960 genes at a significance level of p<0.001 (FDR ≤5.23%). Then, from the latter, we identified groups of genes that were highly correlated with each other (Kendall's $\tau > 0.9$) and chose the longest gene in each group to generate a statistically non-redundant set of 11 128 significant genes. Finally, we used mRMR method and identified an optimal set of 20 genes that were strongly associated with CRC status (see online supplementary figure S8 and table S13). PCA using these 20 genes showed good separation of patients with CRC from controls (figure 2A). PERMANOVA analysis showed that only CRC status, stage and fasting blood glucose explained the variation in the 20 marker gene abundances with statistical significance (p≤0.01; see online supplementary table S14). We computed a simple CRC index based on unweighed log relative abundance of these 20 markers, which clearly separated CRC patient microbiomes from control microbiomes, as well as from 490 faecal microbiomes from two previous studies on type 2 diabetes in Chinese individuals[25] and IBD in European individuals[38] (figure 2B; median CRC index for patients and controls in our study were 7.31 and −5.56, respectively; Wilcoxon rank-sum test, $q<6\times10^{-11}$ for all five comparisons; see online supplementary table S15).

Evaluating CRC biomarkers using targeted qPCR

Translating our gene markers into diagnostic biomarkers would require reliable measurement by simple, affordable and targeted methods such as qPCR. To verify whether gene abundances measured by metagenomics sequencing and qPCR are comparable, we randomly selected two case-enriched and two control-enriched gene markers and measured their abundances by qPCR in a subset of cohort C1 (51 cases and 45 controls). Quantification by metagenomic sequencing and qPCR platforms showed strong correlations (Spearman r=0.81–0.95; see online supplementary figure S9), suggesting that both measurements are reliable. Next, we measured the abundance of these four gene markers using qPCR in an independent Chinese cohort C2 (156 faecal samples; 47 cases and 109 controls; see online supplementary table S16). The two control-enriched genes did not show significant associations in C2 (p>0.31; see online supplementary table S17). On the other hand, CRC-enriched gene markers (m1704941, butyryl-CoA dehydrogenase from *F. nucleatum*; m482585, RNA-directed DNA polymerase from an unknown microbe) were also significantly enriched in CRC samples of C2 after adjusting for colonoscopy (p=0.0015 and 0.045, respectively, see online supplementary table S17). Among these, only the gene from *F. nucleatum* exhibited a significant OR after a Mantel–Haenszel test adjusted for colonoscopy (OR 18.5, p=0.0051; see online supplementary table S17). CRC index based on abundances of the four genes only moderately classified CRC microbiomes from control microbiomes in C2 (areas under the receiver-operating curve (AUC)=0.73; see online supplementary figure S10), suggesting that choosing randomly from the list of 20 biomarkers was not an effective strategy. Nevertheless, the gene from *F. nucleatum* was present only in 4 out of 109 control microbiomes, suggesting a potential for developing specific diagnostic tests for CRC using faecal samples.

Gene marker validation in independent metagenomic cohorts

To identify robust biomarkers that can have a more general applicability, we evaluated all 20 gene markers using faecal metagenomes from a cohort with different genetic background and lifestyle: 16 patients with CRC and 24 control individuals from Denmark (cohort D; see online supplementary table S18). When mapped to 4.3 million gut microbial genes, Danish metagenomes exhibited significantly higher gene richness and gene alpha diversity, both in cases (Wilcoxon rank-sum tests, gene count: $p=1.94\times10^{-5}$; Shannon's index: $p=5.85\times10^{-5}$) and controls (gene count: p=0.0017; Shannon's index: $p=9.34\times10^{-4}$; see online supplementary figure S11 and table S19), agreeing with a recent study and suggesting differences in gut microbial community structure between Chinese and Danish populations.[39] Among the 102 514 genes associated with CRC status in cohort C1, only 1498 genes could be validated in cohort D. However, CRC-enriched genes were shared significantly more between the two populations than control-enriched genes (1452 out of 35 735 CRC-enriched vs 46 out of 66 779 in control-enriched; two-tailed χ^2 test, $\chi^2=2576.57$, p<0.0001). Over half (53.6%) of the 1452 CRC-enriched genes were from just three species: *P. micra* (389 genes), *S. moorei* (204 genes) and *Clostridium symbiosum* (177 genes) (see online supplementary table S20). At the species level, *P. micra* was enriched in CRC microbiomes using all three methods, while *P. stomatis*, *Gemella morbillorum* and *S. moorei* were enriched according to two methods (Wilcoxon rank-sum test, q<0.05; see online supplementary table S21). Notably, all species that were validated by at least one method were CRC-enriched. These results suggest that changes in colorectal environment during CRC development and progression may facilitate growth of similar species across the two populations, potentially leading to the reduced microbial diversity observed in patients with CRC (see online supplementary figure S1C), in line with earlier observations by others.[40] CRC index using 20 gene markers discovered in cohort C1 marginally differentiated Danish patient microbiomes from controls (Wilcoxon rank-sum test, p=0.029) and exhibited moderate classification potential (area under ROC curve, AUC=0.71; see online supplementary figure S12). Only 4 out of 20 genes (two from *Peptostreptococcus anaerobius* and one each from *P. micra* and *F. nucleatum*) were associated with CRC status in cohort D (Wilcoxon rank-sum test, q≤0.05; all CRC-enriched; see online

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

supplementary table S22). Among the factors we had recorded, only CRC status could explain the variation in these four genes (PERMANOVA $p \leq 0.0001$; see online supplementary table S23).

For additional unbiased validation of the four gene markers, we used two recently published metagenomic datasets—an Austrian population (cohort A) consisting of 55 controls and 41 patients with CRC[31] and a French population (cohort F) consisting of 61 controls and 53 patients with CRC.[22] As our discovery cohort C1 only included carcinoma samples, we excluded all patients with adenoma and compared carcinoma patients with non-adenoma/non-carcinoma controls, contrary to the strategy used by the latter study[22] that included small adenomas in controls and excluded large adenomas. All four genes were significantly enriched in carcinoma faecal samples from both cohorts (Wilcoxon rank-sum test $q<0.0035$; see online supplementary table S24). CRC index using these four genes classified patients with CRC with AUC of 0.77 and 0.72 for cohorts A and F, respectively. When we checked association of all 20 markers, cohorts A and F each could validate an additional gene associated with CRC (see online supplementary table S25). Interestingly, one marker enriched in control samples in cohort C1 was enriched in CRC samples in cohort A.

Accurate classification of CRC using qPCR

Two of the four cross-ethnically validated gene markers were transposases from *P. anaerobius*. The third gene (m1704941, butyryl-CoA dehydrogenase from *F. nucleatum*) was incidentally among the two genes successfully validated using qPCR in cohort C2. The fourth gene from *P. micra* was the highly conserved *rpoB* gene encoding RNA polymerase subunit β, often used as a phylogenetic marker.[41] We performed additional qPCR measurements of *rpoB* from *P. micra* in cohort C2, which showed a significant enrichment in CRC patient microbiomes (Wilcoxon rank-sum test adjusted for colonoscopy, $p=8.97\times10^{-8}$). Mantel–Haenszel OR adjusted for colonoscopy was 20.17 (95% CI 4.59 to 88.6, $p=3.36\times10^{-7}$). Combined qPCR measurements of the two genes clearly separated CRC from control samples in cohort C2 (Wilcoxon rank-sum test adjusted for colonoscopy, $p=1.384\times10^{-8}$, figure 3A) and accurately classified CRC samples with an improved AUC of 0.84 (true-positive rate (TPR)=0.723; false-positive rate (FPR) =0.073; figure 3B). Accuracy was slightly better than that in a recent study (reporting AUC=0.836, TPR=0.58, FPR=0.08), even though they used a combination of abundances of 22 species using metagenomic sequencing.[22] Mantel–Haenszel OR, adjusted for colonoscopy, for detecting at least one of the two markers by qPCR in patients with CRC was 22.99 (95% CI 5.83 to 90.8, $p=5.79\times10^{-8}$). When stratifying cohort C2 into early-stage (stages I–II) and late-stage (stages III–IV) patients with cancer, classification potential and ORs were still significant (see online supplementary table S26). Abundance of these two genes was significantly higher compared with control samples starting from stage II of CRC (figure 3C, D), agreeing with our results from species abundances and providing proof- of principle that faecal metagenomes may harbour non-invasive biomarkers for identification of early-stage CRC.

DISCUSSION

We have reported the first successful cross-ethnic validation of metagenomic gene markers for CRC, notably including data from four countries. Two recent studies reported on potential CRC diagnosis using metagenomic sequencing of faecal microbiomes. The first study based on 16S ribosomal RNA gene used five operational taxonomic units to classify CRC from healthy samples in a cohort from the USA.[21] As they did not perform any independent validation, we cannot compare our validation accuracy with theirs. The second study based on shotgun metagenomic sequencing used 21 species discovered in a French cohort to accurately classify patients with CRC in a German cohort.[22] Higher accuracy in their external validation (AUC=0.85 compared with our AUC of 0.77 and 0.72) could be because the validation cohort comes from the same ethnic group. Indeed, when two gene markers discovered in Chinese cohort C1 were validated in the independent Chinese cohort C2 using qPCR, we also achieved a high accuracy (AUC=0.84) even though we moved to a different platform. By doing so, we have also demonstrated, for the first time, the potential for CRC diagnosis through affordable targeted detection methods for microbial biomarkers in faecal samples. Significant improvement in the qPCR classification potential (from AUC=0.73 to AUC=0.84) by using a gene (*rpoB* gene from *P. micra*) validated in cohorts D, F and A reiterates the importance of validating newly discovered biomarkers in independent cohorts with different genetic and environmental background. Further work performing biomarker discovery in high-diversity cohorts or a meta-analysis of published cohorts could reveal whether it leads to increased predictive power. Combining metagenomic markers with the current clinical standard test (faecal occult blood test (FOBT)) has been shown to improve TPR from 49% to 72%.[22] The two markers reported here have reached a comparable TPR without using FOBT. It remains to be seen whether combining FOBT with these markers will further improve accuracy.

Gene markers shared between cohorts from China, Denmark, Austria and France suggest that even though different populations may have different microbial community structures, signatures of CRC-associated microbial dysbiosis could have universal features. Several important observations should be noted: (i) CRC-enriched gene markers had higher correlation between metagenomic and qPCR abundances (r=0.93 and r=0.95) compared with control-enriched genes (r=0.81 and 0.85) in cohort C1; (ii) among four gene markers randomly tested using qPCR in cohort C2, only CRC-enriched genes were validated; (iii) all four gene markers validated in cohort D, all five markers validated in cohort A and four out of five markers validated in cohort F were CRC-enriched (see online supplementary table S25), even though there were 12 control-enriched markers compared with only 8 CRC-enriched markers; (iv) the only marker that switched enrichment during validation in different cohorts was control-enriched; (v) cohort D shared significantly more CRC-enriched genes than control-enriched genes with cohort C1; and (vi) all CRC-associated species from cohort C1 validated in cohort D were CRC-enriched. These features suggest that CRC-enriched biomarkers have a higher chance to be shared across populations and have better diagnostic potential than control-enriched biomarkers. One explanation could be that biomarkers for being healthy are harder to find than biomarkers for a specific disease, which goes against the Anna Karenina principle applied to gut microbiome that predicts higher number of disease-specific disturbed states than undisturbed states.[42] Although it is mandatory to have further validation for all biomarkers in larger cohorts across different populations, our results provide a proof of principle that development of an affordable diagnostic test using faecal microbial gene markers to identify patients with CRC may indeed be possible.

The finding that only two microbial metabolic modules associated with CRC status suggests that the role of microbial pathogens may be more important in disease development than that

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

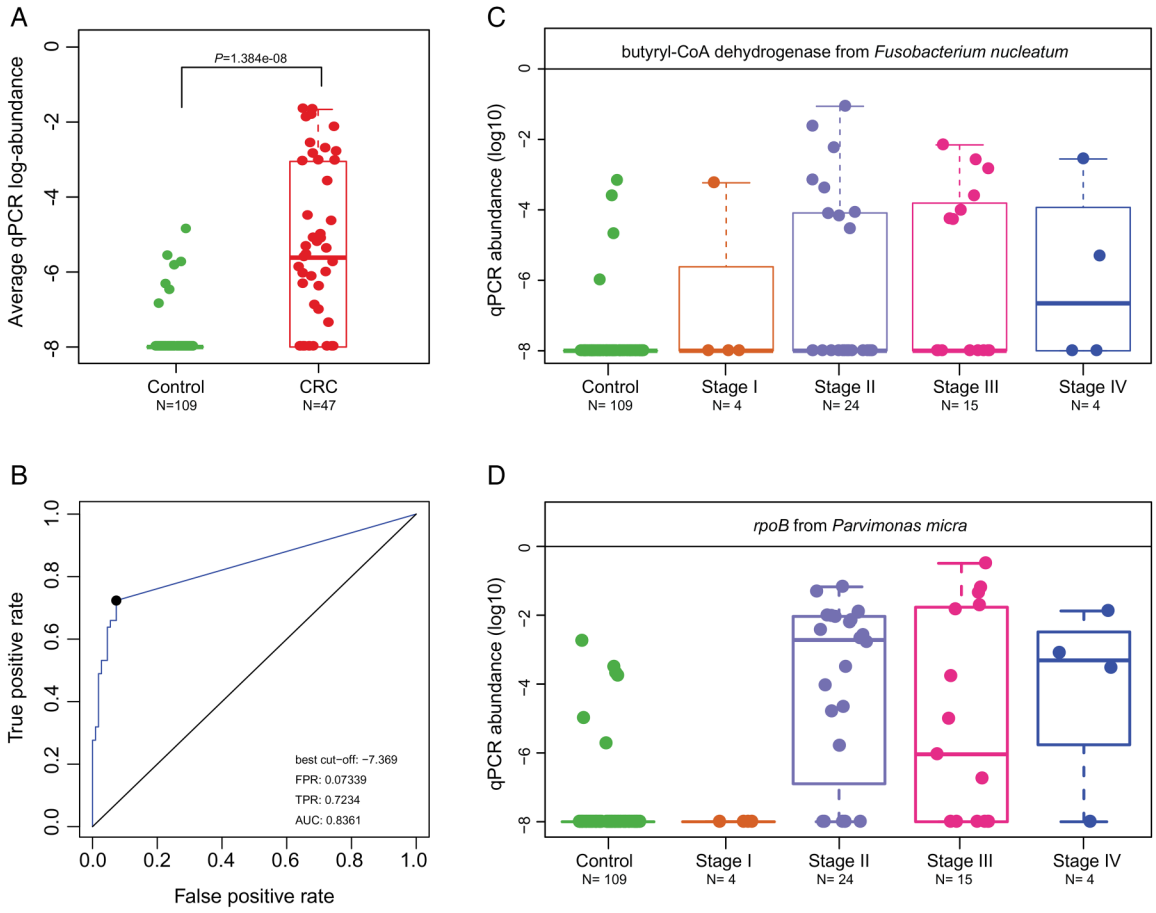


Figure 3 Validating robust gene markers associated with colorectal cancer (CRC). Quantitative PCR (qPCR) abundance of two gene markers (m1704941: butyryl-CoA dehydrogenase from *Fusobacterium nucleatum*, m1696299: RNA polymerase subunit β, *rpoB*, from *Parvimonas micra*) were measured in cohort C2 consisting of 47 cases and 109 healthy controls. Combined log-abundance of the two genes clearly separates CRC microbiomes from controls (A) and classifies CRC microbiomes with an area under the receiver operating characteristic curve of 0.84 (B). The two marker genes show relatively higher incidence and abundance in CRC stages II and III compared with control and stage I microbiomes (C and D). Abundances are plotted in log10 scale, and zero abundance is plotted as −8. AUC, areas under the receiver-operating curve; FPR, false-positive rate; TPR, true-positive rate.

of functional abnormalities of the gut microbiome. Alternatively, expression levels of microbial genes may be more important than functional potential. Further research employing metatranscriptomic studies of microbial gene expression levels will clarify this.

The fact that only CRC-enriched genes and species could be validated across cohorts limits our conclusions on species depleted in CRC-associated microbiomes. We observed significant over-representations of several oral pathogens—*P. micra*, *P. stomatis*, *S. moorei* and *F. nucleatum* in the stool from patients with CRC, suggesting an oral–gut translocation route associated with CRC. Even though we cannot prove this route without further experiments, a recent study based on 300 healthy individuals reported that oral and gut microbiomes were predictive of each other, supporting this view.[43] While some of these species have been statistically associated with oral cancer in earlier studies,[21 22 40] only *F. nucleatum* has been shown to promote a proinflammatory environment leading to tumorigenesis.[19] Our study now introduces *P. micra* as a novel bacterial candidate involved in CRC-associated dysbiosis showing stronger associations with CRC across all five cohorts we investigated. Strong co-occurrence pattern between *P. micra* and the Gram-negative *F. nucleatum*,[44] and the former's ability to increase its capacity to induce inflammatory responses by binding to lipopolysaccharides from Gram-negative bacteria,[45] could mean cooperation between the two, both in terms of colonisation strategies and in promoting a proinflammatory tumorigenic microenvironment. Enrichment of these species starts as early as in stage II of CRC, suggesting that they may play a role in the progression of CRC. Further work characterising *P. micra* could elucidate its role in CRC.

We have demonstrated consistent faecal microbial changes in CRC across four cohorts, identified novel bacterial candidates that may be involved in the development and progression of CRC, validated gene markers in three cohorts from three different countries and reported two bacterial genes that could serve as effective diagnostic biomarkers of CRC. Systematic investigation of key species and gene markers identified here might reveal further candidates. Additional work will be imperative (i) to benchmark these observations against currently used diagnostic approaches, (ii) to identify additional markers with improved predictive value and (iii) to eventually validate them in much larger cohorts. The ultimate goal would be to identify faecal metagenomic markers with strong predictive power to detect early stages of CRC, which would significantly reduce CRC-associated mortality.

Author affiliations
[1]Department of Medicine & Therapeutics, State Key Laboratory of Digestive Disease, Institute of Digestive Disease, LKS Institute of Health Sciences, CUHK Shenzhen Research Institute, The Chinese University of Hong Kong, Hong Kong

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

[2]BGI-Shenzhen, Shenzhen, China
[3]Department of Biology, University of Copenhagen, Copenhagen, Denmark
[4]Department of Veterinary Disease Biology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[5]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia
[6]Department of Surgical Gastroenterology, Hvidovre Hospital, Hvidovre, Denmark
[7]National Institute of Nutrition and Seafood Research, Bergen, Norway
[8]Department of Internal Medicine, Hospital Oberndorf, Q3 Teaching Hospital of the Paracelsus Private University of Salzburg, Oberndorf, Austria
[9]First Department of Internal Medicine, Medical University Innsbruck, Innsbruck, Austria
[10]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[11]Macau University of Science and Technology, Macau, China

Correction notice This article has been corrected since it published Online First. The data sharing statement has been corrected.

Contributors JuY, QF, SHW, DZ and QL contributed equally. The project was designed by JuW, JoJS, JuY, NB and MA. JuY, QF, JoJS and JuW managed the project. JuY, TOY, JS, HJN, TYTL, SCN, QL, ASLC, VW-SW, WKKW and FKLC contributed to acquisition of clinical samples, patients' information and clinical data analyses. QL, SHW, JuY, ZL, PK and BAHJ performed DNA experiments. JuW, JuY, SHW, MA, KK, QF and DYZ designed the analysis. DYZ, MA, QF, YWQ, LQT, YLL, YL, NC, HJJ, JHL, LX and ZL performed the data analysis. DYZ, MA, QF, YWQ, LQT, YLL, YL, NC, HJJ, JHL, LX and ZL worked on metagenomic-wide association study. QL, JuY and T OY did the experimental validation. MA, JuY, SH W, QF, DY Z and LQT wrote the paper. JuW, KK, LM, JoJS, JuY, NB, JiW, HMY, HJJ, JA-A and XX revised the paper.

Funding The project was supported by the National Basic Research Program of China (973 Program, 2011CB809203, 2013CB531401), SHHO foundation Hong Kong, theme-based Research Scheme of the Hong Kong Research Grants Council (T12-403-11), the introduction of innovative R&D team programme of Guangdong Province (no. 2009010016), the Danish Cancer Society (R72-A4659-13-S2) and the Shenzhen Municipal Government of China (CXB201108250098A).

Competing interests None declared.

Patient consent Obtained.

Ethics approval Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (CUHK-NTEC CREC), Ethics Committee of the Capital Region of Denmark and Danish Data Protection Agency.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Metagenomic sequence data set has been deposited in European Nucleotide Archive with accession number PRJEB10878.

REFERENCES

1 Ferlay J, Soerjomataram I, Ervik M, *et al*. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide. IARC CancerBase*. Lyon, France: International Agency for Research on Cancer, 2013.
2 Galvan A, Ioannidis JP, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet* 2010;26:132–41.
3 Lichtenstein P, Holm NV, Verkasalo PK, *et al*. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343:78–85.
4 Foulkes WD. Inherited susceptibility to common cancers. *N Engl J Med* 2008;359:2143–53.
5 Dove WF, Clipson L, Gould KA, *et al*. Intestinal neoplasia in the ApcMin mouse: independence from the microbial and natural killer (beige locus) status. *Cancer Res* 1997;57:812–14.
6 Arthur JC, Perez-Chanona E, Muhlbauer M, *et al*. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* 2012;338:120–3.
7 Cuevas-Ramos G, Petit CR, Marcq I, *et al*. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc Natl Acad Sci USA* 2010;107:11537–42.
8 Grivennikov SI, Wang K, Mucida D, *et al*. Adenoma-linked barrier defects and microbial products drive IL-23/IL-17-mediated tumour growth. *Nature* 2012;491:254–8.
9 Toprak NU, Yagci A, Gulluoglu BM, *et al*. A possible role of Bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clin Microbiol Infect* 2006;12:782–6.
10 Uronis JM, Muhlbauer M, Herfarth HH, *et al*. Modulation of the intestinal microbiota alters colitis-associated colorectal cancer susceptibility. *PLoS ONE* 2009;4:e6026.
11 Wu S, Rhee KJ, Albesiano E, *et al*. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat Med* 2009;15:1016–22.
12 Boleij A, Schaeps RM, Tjalsma H. Association between Streptococcus bovis and colon cancer. *J Clin Microbiol* 2009;47:516.
13 Seder CW, Kramer M, Long G, *et al*. Clostridium septicum aortitis: Report of two cases and review of the literature. *J Vasc Surg* 2009;49:1304–9.
14 Scanlan PD, Shanahan F, Clune Y, *et al*. Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis. *Environ Microbiol* 2008;10:789–98.
15 Sobhani I, Tap J, Roudot-Thoraval F, *et al*. Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS ONE* 2011;6:e16393.
16 Chen W, Liu F, Ling Z, *et al*. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS ONE* 2012;7:e39743.
17 Castellarin M, Warren RL, Freeman JD, *et al*. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res* 2012;22:299–306.
18 Kostic AD, Gevers D, Pedamallu CS, *et al*. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* 2012;22:292–8.
19 Kostic AD, Chun E, Robertson L, *et al*. Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment. *Cell Host Microbe* 2013;14:207–15.
20 Rubinstein MR, Wang X, Liu W, *et al*. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. *Cell Host Microbe* 2013;14:195–206.
21 Zackular JP, Rogers MA, Ruffin MTt, *et al*. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev Res (Phila)* 2014;7:1112–21.
22 Zeller G, Tap J, Voigt AY, *et al*. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 2014;10:766.
23 Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol* 2010;17:1471–4.
24 Godon JJ, Zumstein E, Dabert P, *et al*. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl Environ Microbiol* 1997;63:2802–13.
25 Qin J, Li Y, Cai Z, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;490:55–60.
26 Sunagawa S, Mende DR, Zeller G, *et al*. Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* 2013;10:1196–9.
27 Markowitz VM, Chen IM, Palaniappan K, *et al*. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 2012;40:D115–22.
28 Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 2005;27:1226–38.
29 Kabat GC, Kim MY, Strickler HD, *et al*. A longitudinal study of serum insulin and glucose levels in relation to colorectal cancer risk among postmenopausal women. *Br J Cancer* 2012;106:227–32.
30 van Duijnhoven FJ, Bueno-De-Mesquita HB, Calligaro M, *et al*. Blood lipid and lipoprotein concentrations and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition. *Gut* 2011;60:1094–102.
31 Feng Q, Liang S, Jia H, *et al*. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* 2015;6:6528.
32 Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;40:D109–14.
33 Baracos VE, Mackenzie ML. Investigations of branched-chain amino acids and their metabolites in animal models of cancer. *J Nutr* 2006;136:237S–42S.
34 Gonçalves EM, Salomão EM, Gomes-Marcondes MCC. Leucine modulates the effect of Walker factor, a proteolysis-inducing factor-like protein from Walker tumours, on gene expression and cellular activity in C2C12 myotubes. *Cytokine* 2013;64:343–50.
35 Pedersen RM, Holt HM, Justesen US. Solobacterium moorei bacteremia: identification, antimicrobial susceptibility, and clinical characteristics. *J Clin Microbiol* 2011;49:2766–8.
36 Sundqvist G. Taxonomy, ecology, and pathogenicity of the root canal flora. *Oral Surg Oral Med Oral Pathol* 1994;78:522–30.
37 Knights D, Costello EK, Knight R. Supervised classification of human microbiota. *FEMS Microbiol Rev* 2011;35:343–59.
38 Qin J, Li R, Raes J, *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464:59–65.
39 Li J, Jia H, Cai X, *et al*. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* 2014;32:834–41.
40 Ahn J, Sinha R, Pei Z, *et al*. Human gut microbiome and risk for colorectal cancer. *J Natl Cancer Inst* 2013;105:1907–11.

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

41 Ciccarelli FD, Doerks T, von Mering C, *et al*. Toward automatic reconstruction of a highly resolved tree of life. *Science* 2006;311:1283–7.
42 Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* 2012;7:e30126.
43 Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. *Nature* 2014;509:357–60.
44 Kremer BH, van Steenbergen TJ. Peptostreptococcus micros coaggregates with Fusobacterium nucleatum and non-encapsulated Porphyromonas gingivalis. *FEMS Microbiol Lett* 2000;182:57–62.
45 Yoshioka M, Grenier D, Mayrand D. Binding of Actinobacillus actinomycetemcomitans lipopolysaccharides to Peptostreptococcus micros stimulates tumor necrosis factor alpha production by macrophage-like cells. *Oral Microbiol Immunol* 2005;20:118–21.

Gut: first published as 10.1136/gutjnl-2015-309800 on 25 September 2015. Downloaded from http://gut.bmj.com/ on April 27, 2024 by guest. Protected by copyright.

1. Supplementary Methods

Contents

1.1. Sample collection and DNA preparation 2
1.1.1. Sample collection in China 2
1.1.2. Sample collection in Denmark 2
1.1.3. DNA extraction 3
1.1.4. DNA library construction and sequencing 3
1.2. Gene profile analysis 3
1.2.1. Generating gene profiles 3
1.2.2. Bio-diversity analysis 3
1.2.3. Rarefaction analysis based on gene profile 4
1.2.4. Analysis of factors influencing gut microbial gene profile 4
1.2.5. Identification of CRC associated genes 4
1.2.6. Estimating the false discovery rate (FDR) 4
1.3. Taxonomic annotation of genes 5
1.3.1. Creating IMG genome database and species annotation of IMG genomes 5
1.3.2. Identification of CRC associated metagenomic linkage group (MLG) species 5
1.4. Data profile construction 5
1.4.1. Functional profiles based on KEGG database 5
1.4.2. Molecular operational taxonomic unit (mOTU) profiles 5
1.4.3. IMG-species and IMG-genus profiles 5
1.4.4. MLG-species and MLG-genus profiles 6
1.5. Biomarker discovery analysis 6
1.5.1. Minimum Redundancy Maximum Relevance (mRMR) framework 6
1.5.2. Definition of CRC index 6
1.5.3. Receiver Operator Characteristic (ROC) analysis 7
1.5.4. Functional signatures associated with CRC 7
1.5.5. Gut microbial species associated with CRC 7
1.5.6. Identifying gut microbial species that can classify CRC microbiomes 7
1.5.7. Species co-occurrence network construction 7
1.6. References 8

1.1. Sample collection and DNA preparation

1.1.1. Sample collection in China

The study included adult individuals undergoing colonoscopy at the Shaw Endoscopy Centre at the Prince of Wales Hospital, the Chinese University of Hong Kong. The Chinese cohorts C1 (**Table S1**) and C2 (**Table S16)** included individuals presenting symptoms such as change of bowel habit, rectal bleeding, abdominal pain or anaemia, and asymptomatic individuals aged 50 or above undergoing screening colonoscopy. The exclusion criteria were: 1) use of antibiotics within the past 3 months; 2) on a vegetarian diet; 3) had an invasive medical intervention within the past 3 months; 4) had a past history of any cancer, or inflammatory disease of the intestine. Subjects were asked to collect stool samples in standardized containers at home, and store the samples in their home freezer immediately. Frozen samples were then delivered to the hospital in insulating polystyrene foam containers and stored at -80ºC immediately until further analysis. The study protocol in Hong Kong was approved by the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (CUHK-NTEC CREC).

1.1.2. Sample collection in Denmark

Cohort D: Stool samples were collected from individuals referred to colonoscopy due to symptoms associated with CRC or from patients who had been diagnosed with CRC and referred to large bowel resection for their primary cancer disease (See **Table S18**). All individuals were included at their visit to the out-patient clinic either before colonoscopy or before the operation and always before bowel evacuation. The individuals received a stool collection set including a tube without stabilizing buffer and were instructed to collect a stool sample at home one or two days before initiation of large bowel evacuation. Every included individual kept the sample refrigerated at -18ºC and contacted a research nurse who collected the sample. At the laboratory stool samples were immediately snap frozen in liquid nitrogen and subsequently stored at -80ºC under 24/7 electronic surveillance until analysis.

All included individuals thus underwent complete colonoscopy either as the primary examination or after the subsequent operation. Exclusion criteria were previous adenoma, previous CRC and previous or present other malignant diseases.

The recording of data from the included individuals was performed according to the Helsinki II declaration. The protocol was approved by the Ethics Committee of the Capital Region of Denmark (H-3-2009-110) and the Danish Data Protection Agency (2008-41-2252).

1.1.3. DNA extraction

Chinese samples: Stool samples were thawed on ice and DNA extraction was performed using the Qiagen QIAamp DNA Stool Mini Kit (Qiagen) according to manufacturer's instructions. Extracts were treated with DNase-free RNase to eliminate RNA contamination. DNA quantity was determined using NanoDrop spectrophotometer, Qubit Fluorometer (with the Quant-iTTMdsDNA BR Assay Kit) and gel electrophoresis.

Danish samples: A frozen aliquot (200 mg) of each fecal sample was suspended in 250 µl of 4 M guanidine thiocyanate– 0.1 M Tris (pH 7.5) and 40 µl of 10% N-lauroyl sarcosine. Then, DNA extraction was conducted using bead beating method as previously described[24]. The DNA concentration and its molecular size were estimated by nanodrop (Thermo Scientific) and agarose gel electrophoresis.

1.1.4. DNA library construction and sequencing

DNA library construction for samples from cohort C1, C2 and D was performed following the manufacturer's instruction (Illumina) at the same facility. We used a previously described workflow to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturation, and hybridization of the sequencing primers[25].

We constructed one paired-end (PE) library with insert size of 350bp for each sample, followed by high-throughput sequencing to obtain around 30 million PE reads of length 2x100bp. High-quality reads were obtained by filtering low-quality reads with ambiguous 'N' bases, adapter contamination and human DNA contamination from the Illumina raw reads, and by trimming low-quality terminal bases of reads simultaneously.

1.2. Gene profile analysis

1.2.1. Generating gene profiles

We mapped our high-quality reads to a published reference gut microbial gene catalogue derived from European and Chinese adults[25] (using sequence identity >= 90%). We then derived the gene profiles using previously described procedures[25].

1.2.2. Bio-diversity analysis

Based on the gene profiles, we calculated the within-sample (alpha) diversity to estimate the gene richness using Shannon index and Simpson index of alpha diversity[25], where larger value indicates a higher degree of diversity in the sample. To analyse effects of different phenotype factors, including

age, BMI, eGFR, TCHO, LDL, HDL, and TG, on gut microbial diversity, Pearson correlation coefficients between each factor and Shannon index were also calculated.

1.2.3. Rarefaction analysis based on gene profile

Estimation of total gene richness in a set of metagenomics samples was performed by randomized sampling with replacement. This was done independently for cohort C1, CRC patients group in C1, and non-CRC control group in C1. In each set of size *n*, we randomly sampled *n* individual samples with replacement and then calculated the total number of genes that could be identified from these samples. Only genes with ≥1 mapping reads were considered to be present. This was repeated 100 times. The result showed that the control group had higher gene richness than the case group.

1.2.4. Analysis of factors influencing gut microbial gene profile

From the reference gene catalogue[25], we derived a subset of 2.1M genes that appeared in at least 6 samples in all 128 samples from cohort C1 (74 CRC and 54 control). We used the permutational multivariate analysis of variance (PERMANOVA) test to assess the effect of different characteristics, including age, BMI, eGFR, TCHO, LDL, HDL, TG, gender, DM, CRC status and location, on gene profiles of 2.1M genes (see Supplementary Table S1 for explanation of these factors). We performed the analysis using the implementation in the "vegan" package in R, and the permuted p-value was obtained by performing 10,000 permutations. We also adjusted for multiple testing using the function "p.adjust" in R with Benjamini-Hochberg method to get the corresponding q-values.

1.2.5. Identification of CRC associated genes

To identify the association between the metagenomic gene profiles and CRC, a two-tailed Wilcoxon rank-sum test was performed for each of the 2.1M genes. We obtained 140,455 gene markers which were enriched in either case or control with $P<0.01$. To control for colonoscopy as a confounding factor, we performed the independence test after stratifying by colonoscopy status, using the *kruskal_test* function implemented in COIN package in R.

1.2.6. Estimating the false discovery rate (FDR)

Instead of a sequential p-value rejection method, we applied the "qvalue" method proposed in a previous study[46] to estimate the FDR.

1.3. Taxonomic annotation of genes

1.3.1. Creating IMG genome database and species annotation of IMG genomes

Bacterial, archaeal and fungal genome sequences were extracted from IMG v400 reference database[27] downloaded from http://ftp.jgi-psf.org. In total, 522,093 sequences were obtained. For each IMG genome, using the NCBI taxonomy identifier provided by IMG, we identified the corresponding NCBI taxonomic classification at species and genus levels using NCBI taxonomy dump files. The genomes without corresponding NCBI species names were left with their original IMG names, most of which were unclassified at the genus and species levels.

1.3.2. Identification of CRC associated metagenomic linkage group (MLG) species

Based on the identified 140,455 CRC associated marker genes, we constructed the CRC associated MLGs using the method described in our previous study on type 2 diabetes[25]. All the above genes were aligned to the reference genomes of IMG database v400 to get genome level annotation. An MLG was assigned to a genome if >50% constituent genes were annotated to that genome, otherwise it was termed as unclassified. 86 MLGs consisting over 100 genes were selected as CRC associated MLGs. These MLGs were grouped based on the species annotation of these genomes to construct MLG species.

1.4. Data profile construction

1.4.1. Functional profiles based on KEGG database

Based on the gene profiles, we derived the KO profiles using previously described procedures[25]. Functional analysis was performed based on KEGG orthologous group (KO) abundance profiles. KEGG module and pathway (the KEGG Class Level 2) abundance profiles were calculated by summing the abundances of KOs belonging to each functional category.

1.4.2. Molecular operational taxonomic unit (mOTU) profiles

Clean reads were aligned to mOTU reference database (total 79268 sequences) with default parameters[26]. 549 species level mOTUs were identified, including 307 annotated species and 242 mOTU linkage groups (not to be confused with metagenomics linkage groups) without representative genomes. Most of the mOTU linkage groups were putatively Firmicutes or Bacteroidetes.

1.4.3. IMG-species and IMG-genus profiles

SOAP reference index was constructed for the IMG genome database based on 7 equal size chunks of the original file. Clean reads were aligned to reference using SOAP aligner[47] version 2.22, with parameters “-m 4 -s 32 -r 2 -n 100 -x 600 -v 8 -c 0.9 -p 3”. Then, SOAP coverage software was used to calculate read coverage of each genome, normalized with genome length, and further normalized to

relative abundance for each individual sample. The profile was generated based on uniquely mapped reads only.

1.4.4. MLG-species and MLG-genus profiles

To estimate the relative abundance of an MLG species, we estimated the average abundance of the genes of the MLG species, after removing the 5% lowest and 5% highest abundant genes. Relative abundance of IMG species was estimated by summing the abundance of IMG genomes belonging to that species. Genus abundances were estimated by analogously summing species abundances.

1.5. Biomarker discovery analysis

1.5.1. Minimum Redundancy Maximum Relevance (mRMR) framework

To establish CRC classification only using gut metagenomic markers, we adopted the mRMR method[28] to perform feature selection. We used the "sideChannelAttack" package from R to perform an incremental search and found 128 sequential marker sets. For each sequential set, we estimated the error rate by leave-one-out cross-validation (LOOCV) of a linear discrimination classifier. The optimal selection of marker sets was the one corresponding to the lowest error rate. In the present study, we made the feature selection on a set of 102,514 CRC associated gene markers. Since it was computationally prohibitive to perform mRMR using all genes, we derived a statistically non-redundant gene set. Firstly, we pre-grouped the 102,514 CRC associated genes that are highly correlated with each other (Kendall correlation > 0.9). Then we chose the longest gene as representative gene for the group, since longer genes have a higher chance of being functionally annotated, and will attract more reads during the mapping procedure. This generated a non-redundant set of 11,128 significant genes. Subsequently, we applied the mRMR feature selection method[28] to the 11,128 significant genes and identified an optimal set of 20 gene biomarkers that are strongly associated with CRC for classification.

1.5.2. Definition of CRC index

To evaluate the risk of CRC from the gut metagenome, we defined and computed a CRC index for each individual on the basis of the 20 gene markers identified by mRMR procedure. For each individual sample, the CRC index of sample *j* that denoted by I_j was computed by the formula below:

$$I_j = \left[\frac{\sum_{i\in N}\log_{10}\left(A_{ij}+10^{-20}\right)}{|N|} - \frac{\sum_{i\in M}\log_{10}\left(A_{ij}+10^{-20}\right)}{|M|}\right]$$

where A_{ij} is the relative abundance of marker i in sample j. N is a subset of all CRC-enriched markers in these 20 genes. M is a subset of all control-enriched markers in these 20 genes. And $|N|$ and $|M|$ are

the sizes of these two sets. The ability of the CRC index to distinguish CRC patient microbiomes from non-CRC microbiomes was examined using Wilcoxon rank-sum test. P-values estimated by these tests were adjusted for multiple testing using Benjamini-Hochberg method, when comparing CRC samples in cohort C1 with several other sample sets.

1.5.3. Receiver Operator Characteristic (ROC) analysis

We applied the ROC analysis to assess the performance of CRC classification based on metagenomic markers. We used the "Daim" package in R to draw the ROC curve.

1.5.4. Functional signatures associated with CRC

Wilcoxon rank-sum test with Benjamini-Hochberg adjustment was employed to identify KEGG KOs, modules and pathways associated with CRC.

1.5.5. Gut microbial species associated with CRC

Out of the 86 MLG species consisting over 100 genes, 85 MLGs were associated with CRC at a significance level of $q<0.05$ according to Wilcoxon rank-sum tests with Benjamini-Hochberg adjustment. This higher number is expected as the MLGs were constructed with genes that are associated with CRC in the first place. Using the same procedure at the same significance level, 28 IMG species and 21 mOTU species were associated with CRC.

1.5.6. Identifying gut microbial species that can classify CRC microbiomes

To evaluate the classification potential of the gut microbial species associated with CRC (identified by three methods: 85 MLG-species, 28 IMG species, and 21 mOTU species), we used "randomForest 4.5-36" package in R vision 2.10 based on these species profiles. For each method, firstly, we sorted all the *N* species by the importance given by the "randomForest" method. Then we created incremental marker sets by creating subsets of the top ranked species, starting from top 1 species and ending at *N* species. For each marker set, we calculated the false prediction ratio in Chinese cohort C1. Species from the marker set with lowest false prediction ratio were considered to have high potential for classification of CRC microbiomes from control microbiomes. Furthermore, we drew the ROC curve using the probability of illness based on these selected species markers.

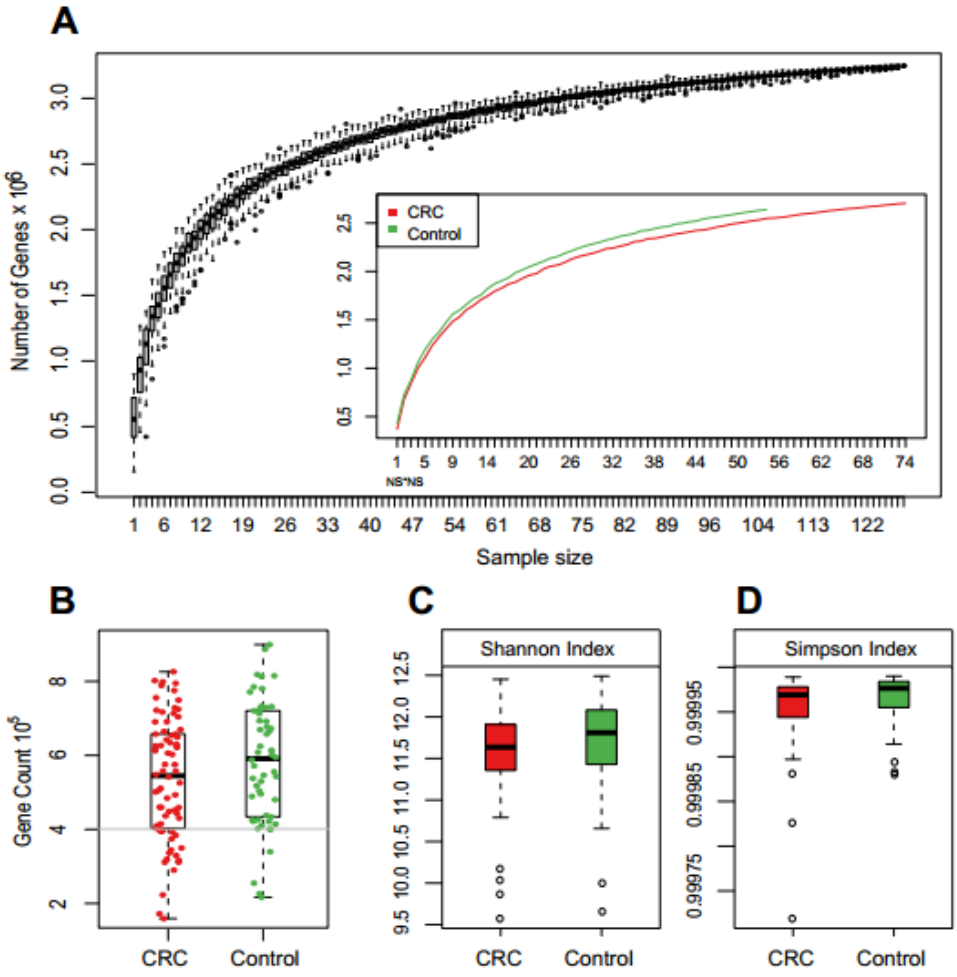
1.5.7. Species co-occurrence network construction

Co-occurrence networks were constructed for the 85 MLGs, 28 IMG species and 21 mOTUs associated with CRC ($q<0.05$) using Spearman's correlation coefficient (>0.5 or <-0.5), as described previously[25]. Cytoscape[48] v3.0.2 was used to construct the three networks.

1.6. References

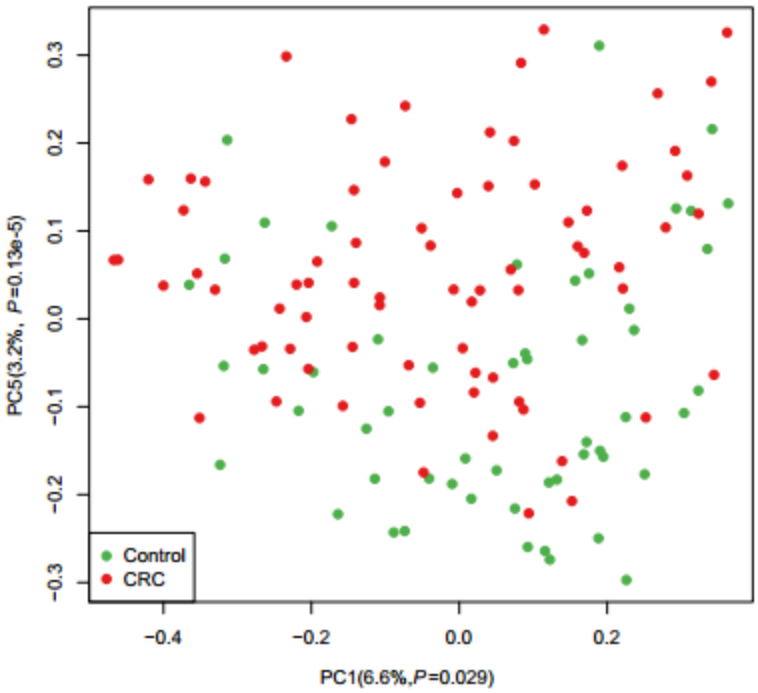
46 Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 2003;**100**:9440-5.
47 Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, *et al.* SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009;**25**:1966-7.
48 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 2003;**13**:2498-504.

1. Supplementary Figures



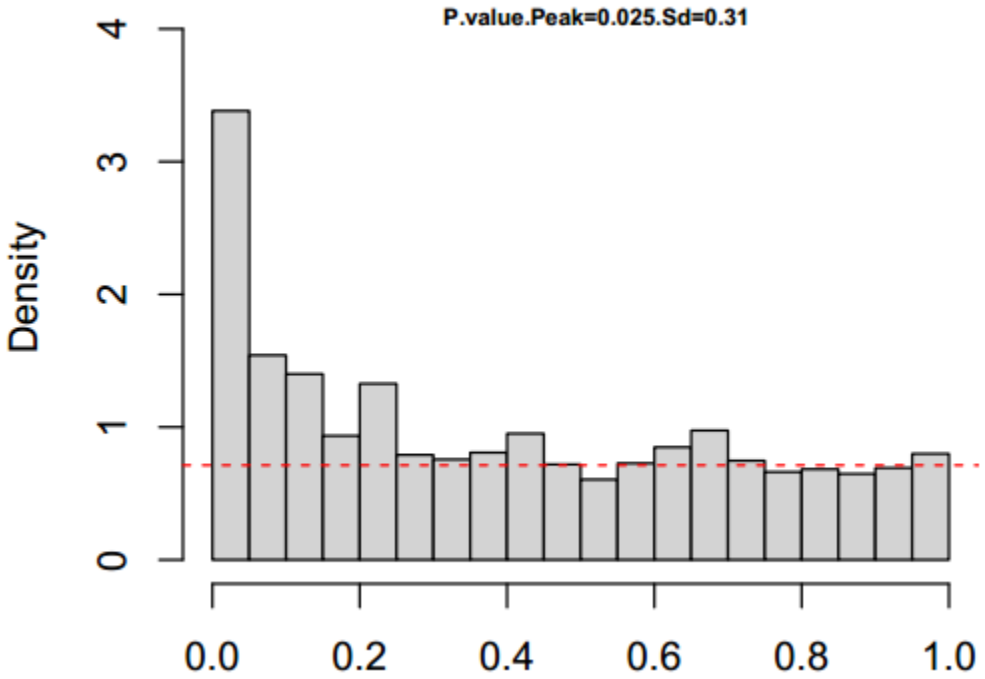
Supplementary Figure 1. Microbial dysbiosis in colorectal cancer.

(A) Rarefaction curves showing cumulative number of genes sequenced with increasing sample size. The numbers are close to saturation given the current sample sizes for all 128 samples (grey). Inset: CRC samples (red) have significantly lower gene counts compared to healthy controls (green). Only the first three points corresponding to sizes of up to 3 samples each are not significant (NS). **(B)** Gene count distribution of 128 CRC case and healthy control individuals. The grey line shown corresponds to 400,000 genes, below which 80% individuals had cancer. **(C)**The Shannon index of the CRC case and healthy control microbiomes from gene abundances. **(D)** Simpson index of the CRC case and healthy control microbiomes from gene abundances.



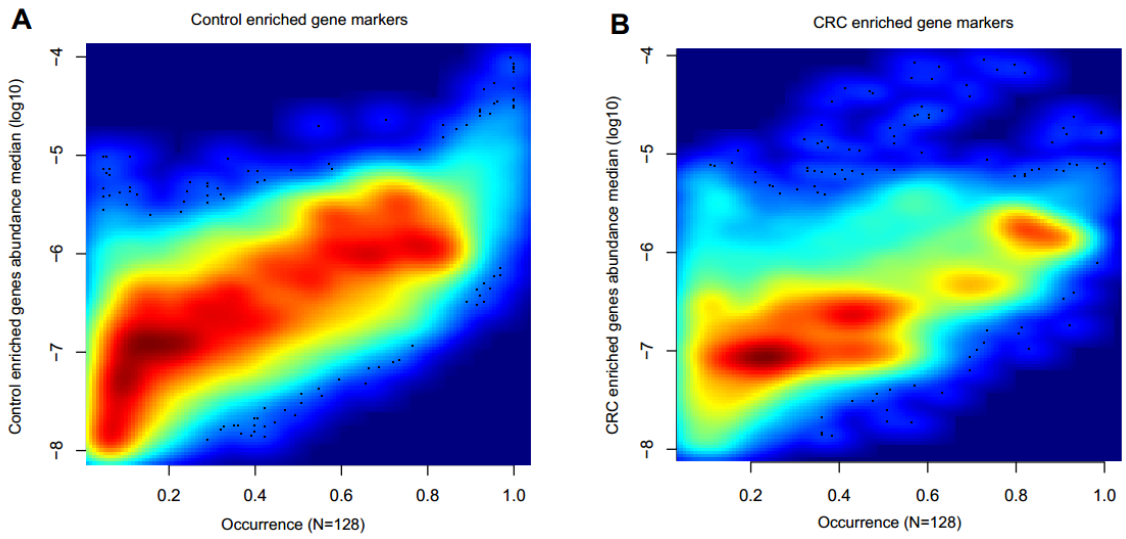
Supplementary Figure 2. Principal component analysis using 2,110,489 genes identified in cohort C1.

PC1 and PC5 are associated with CRC status (see Supplementary Table S5), but do not show separation between gut microbiomes of CRC patients and control individuals based on gene profiles. See **Fig. 2A** where a strong separation is observed.



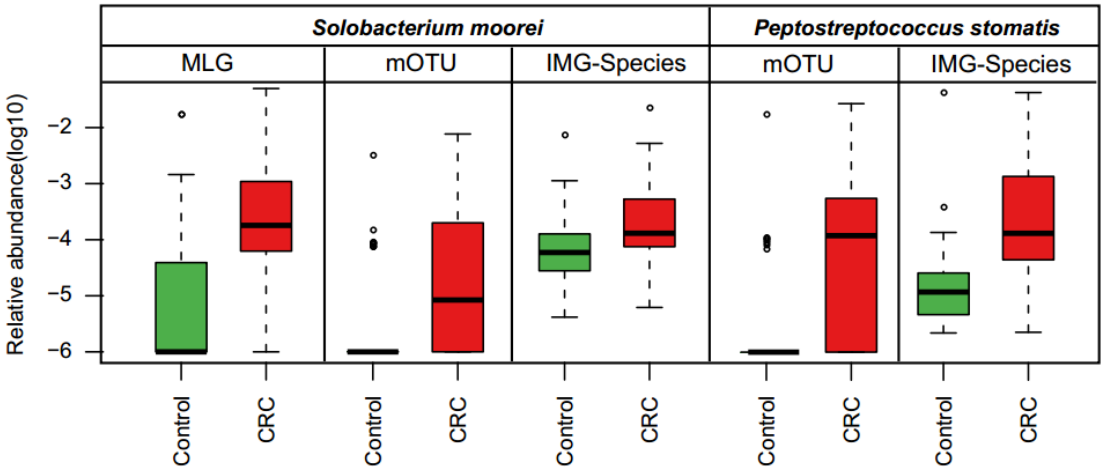
Supplementary Figure 3. Distribution of *P*-value association statistics of all microbial genes in cohort C1.

The association analysis of CRC p-value distribution identified a disproportionate over-representation of strongly associated markers at lower P-values, with the majority of genes following the expected P-value distribution under the null hypothesis. This suggests that the significant markers likely represent true rather than spurious associations.

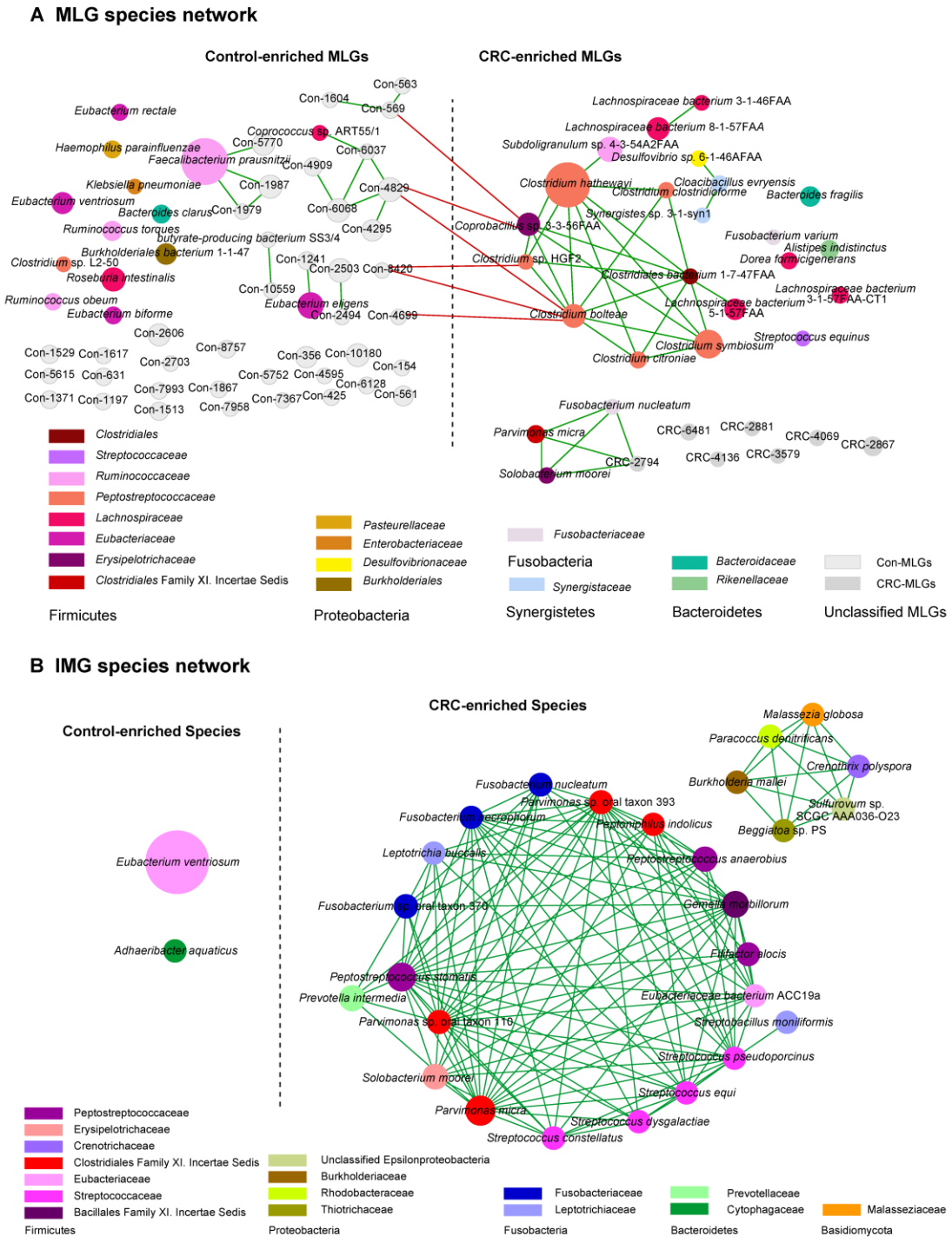


Supplementary Figure 4. 2-dimensional histogram of abundance-vs-occurrence rate of CRC-associated gene markers.

The CRC-associated gene markers selection were based on the significant enrichment in CRC case or healthy control. We computed the occurrence rate and median relative abundance for the CRC-enriched gene markers and control-enriched gene markers in all 128 samples from C1, and generated a 2-dimensional histogram following previously described methods[25] to show the distribution of all marker genes. **(A)** Control-enriched gene markers exhibit a wider range of occurrence rate and relative abundance. **(B)** CRC-enriched gene markers are mostly present in low occurrence rate and low relative abundance.

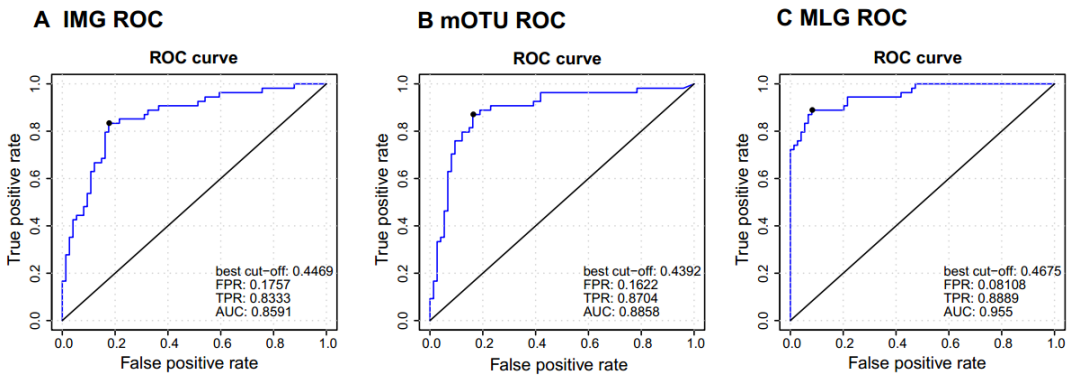


Supplementary Figure 5. Enrichment of *Solobacterium moorei* and *Peptostreptococcus stomatis* in CRC patient microbiomes in cohort C1.



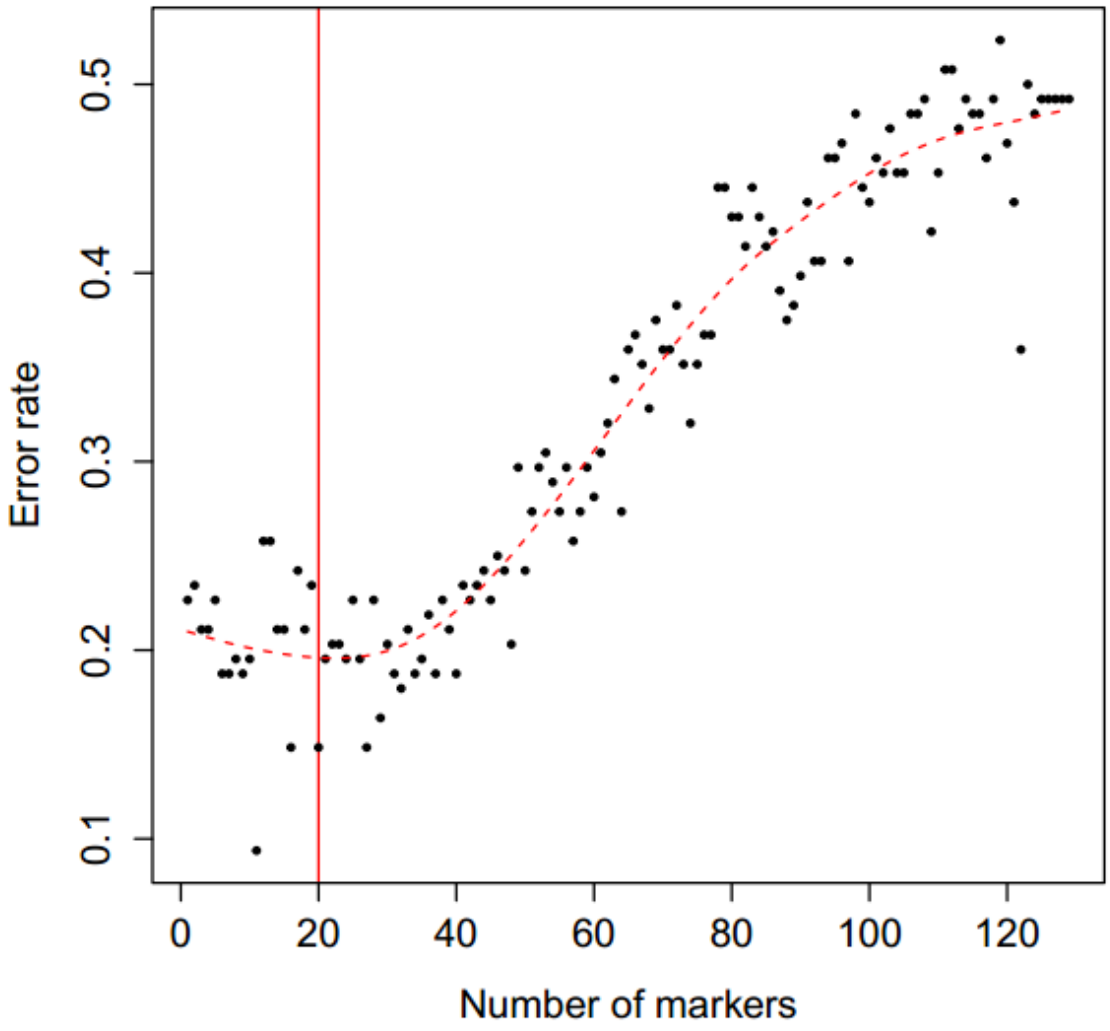
Supplementary Figure 6. Significant species network using IMG version 400 and MLG annotation with q-value<0.05 in cohort C1.

(A) A co-occurrence network deduced from 85 MLGs significantly associated with CRC. Each MLG with more than 100 genes and at least 50% genes annotated to a single species was annotated with species name. The remaining MLGs were named Con or CRC MLGs according to their enrichment in control and CRC samples, respectively. Species are rearranged in two sides based on their enrichment in CRC or healthy microbiomes. Spearman correlation coefficient values lower than -0.5 (negative correlation) are indicated as red edges, and coefficient values higher than 0.5 (positive correlation) are indicated as green edges. Node size indicates the number of genes within the MLG, and node color shows their taxonomic annotation. **(B)** A co-occurrence network deduced from 28 IMG species significantly associated with CRC. Node size indicates the average of relative abundance for each species. See legend for panel **A** for other details.



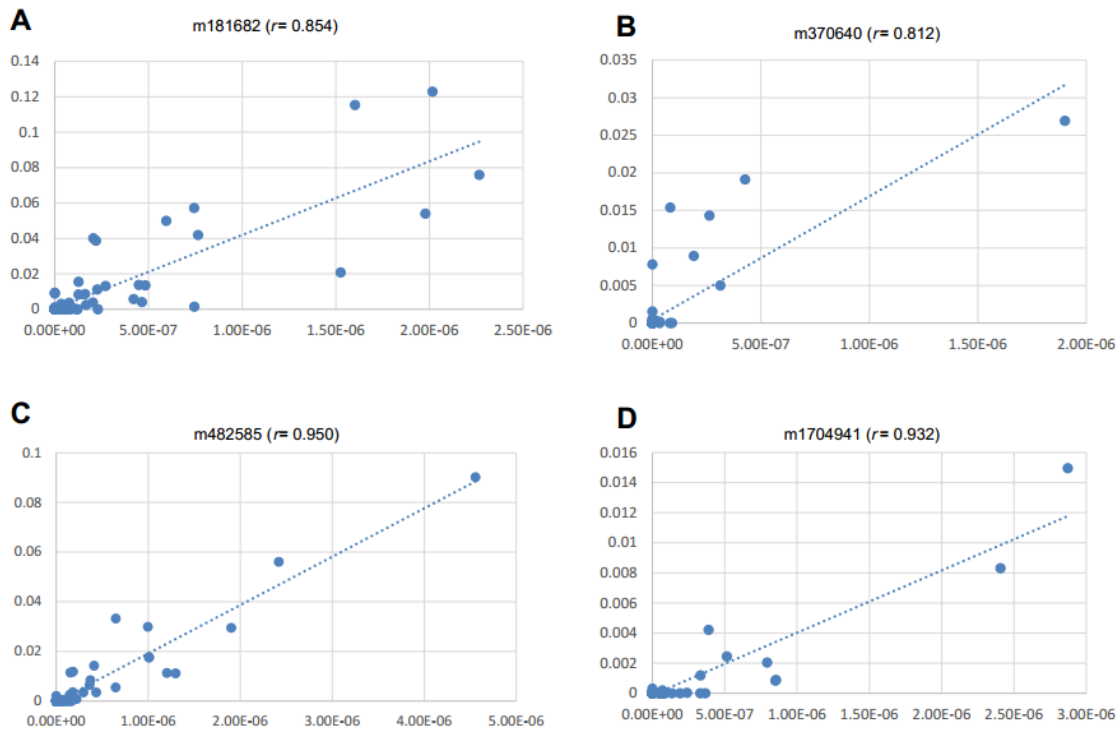
Supplementary Figure 7. The Receive-Operator-Curve of CRC specific species marker selection using random forest method and three different species annotation methods.

A, IMG species annotation using clean reads to IMG version 400. **B,** mOTU species using published methods[26], **C**, All significant genes clustered using MLG methods[25] and the species annotation using IMG version 400.

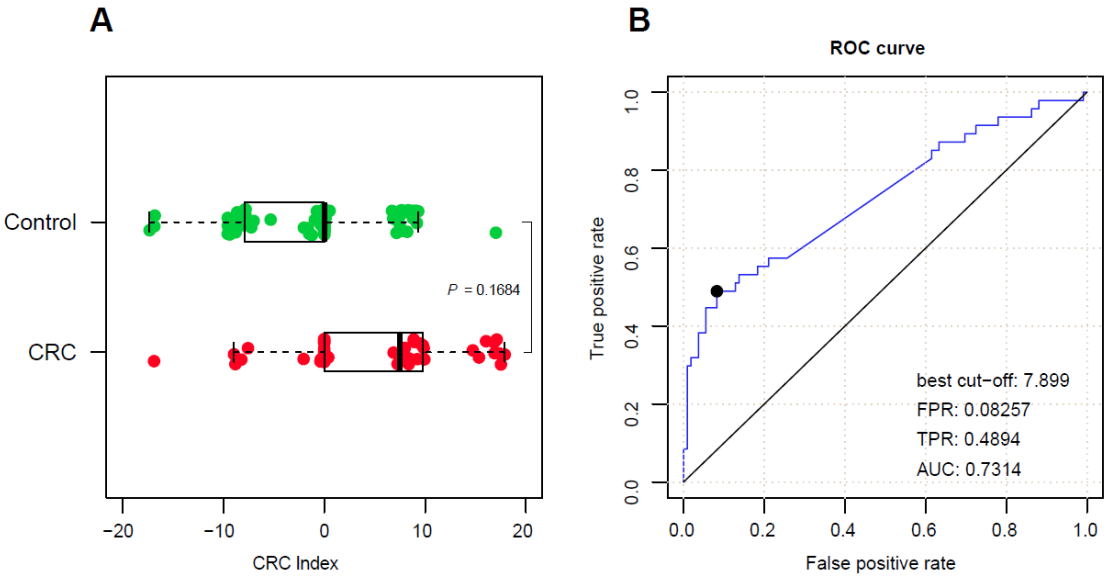


Supplementary Figure 8. Minimum redundancy maximum relevance (mRMR) method to identify 20 gene markers that differentiate CRC cases from controls in cohort C1.

Incremental search was performed using the mRMR method which generated a sequential number of subsets. For each subset, the error rate was estimated by a leave-one-out cross-validation (LOOCV) of a linear discrimination classifier. The optimum subset with the lowest error rate contained 20 gene markers.



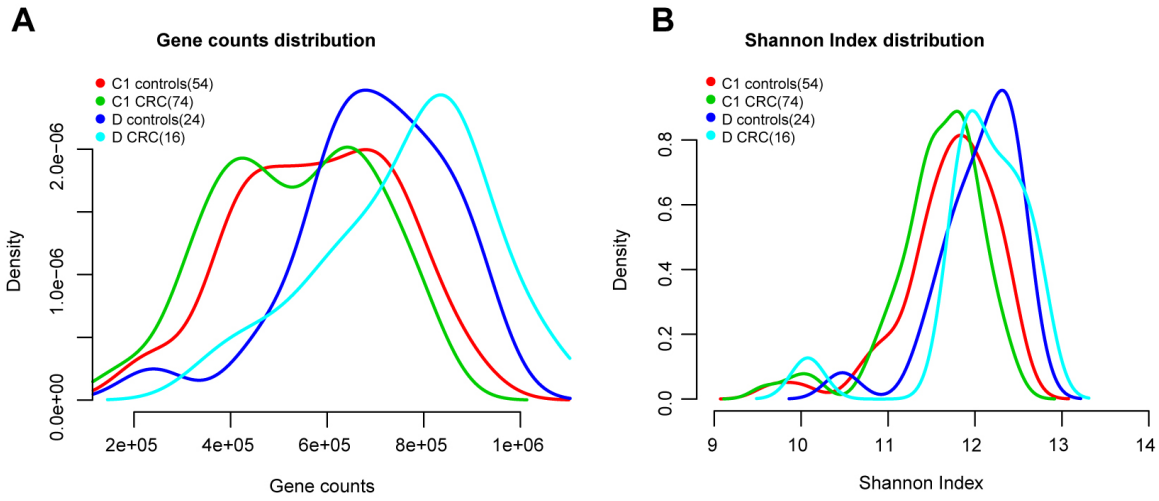
Supplementary Figure 9. Correlation between quantification by the metagenomic approach versus quantitative polymerase chain reaction (qPCR) for four gene markers.



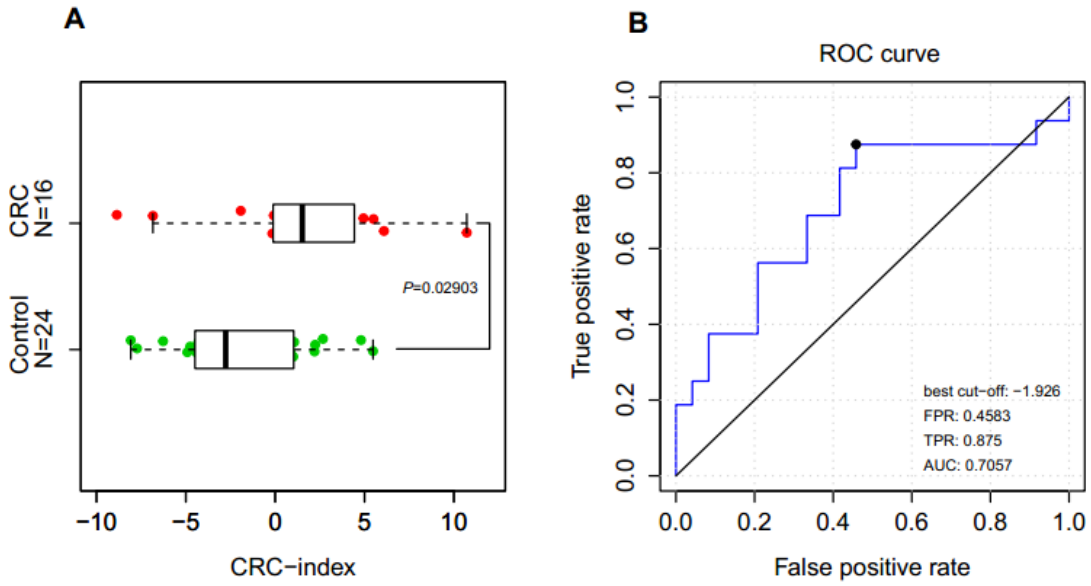
Supplementary Figure 10. Evaluating CRC index from four markers in Chinese

cohort C2 of 156 individuals.

(A) CRC index based on qPCR abundance of 4 gene markers shows marginal separation of CRC and control microbiomes. **(B)** ROC analysis reveals moderate potential for classification using CRC index, with an area under the curve of 0.73.



Supplementary Figure 11. Comparison of (A) gene richness (gene count) and (B) alpha-diversity (Shannon index) distribution in cohorts C1 and D.



Supplementary Figure 12. Evaluating CRC index in cohort D consisting of 40 individuals.

(A) CRC index based on 20 gene markers shows marginal separation of CRC and control microbiomes. **(B)** ROC analysis reveals moderate potential for classification using CRC index, with an area under the curve of 0.71.

Supplementary Tables

Supplementary Table S1. Baseline characteristics of colorectal cancer (CRC) cases and controls in cohort C1. For quantitative traits, the median, minimum and maximum are shown. FBG: fasting blood glucose; ALT/GPT: alanine transaminase/glutamate pyruvated transaminase; BMI: body mass index; DM: diabetes mellitus type 2; HDL: high density lipoprotein; TG: triglyceride; eGFR: epidermal growth factor receptor; TCHO: total cholesterol; Cr: creatinine; LDL; low density lipoprotein; TNM: tumor node metastasis staging system; Statistical tests used for identifying associations between metadata and CRC: † - Wilcoxon test, ‡ - Fisher's exact test. * - Information missing in one CRC patient sample.

Parameter	**Controls (n=54)**	**Cases (n=74)**	***P-value***	**q-value**
Age	63(50,73)	67(34,89)	0.007373932†	0.027652245
Gender (M:F)	33:21	48:26	0.7124‡	0.875363308
BMI	22.86341 (17.08744,35.08618)	23.89549 (17.28882,31.25)	0.1107815†	0.237388929
DM (%)	16 (29.6%)	29 (39.2%)	0.3488‡	0.503844136
Stage of CRC (1:2:3:4)	n.a	18:22:26:8	n.a	n.a
Distribution of detailed TNM stages (T1N0:T2N0:T3N0:T4N0:T2N1:T3N1:T3N2:T3N+:T4N2:T3N1M1:T3N3M1:T4N1M1:T4N2M1:M:M1:multiple liver met)	n.a	12:6:21:1:3:14:5:2:2:2:1:1:1:1:1	n.a	n.a

Leison location(1:2:NA)	n.a	11:54:9	n.a	n.a
Leison specific location（1:2:3:4:6:7:8:9）	n.a	3:3:3:2:6:14:5:29	n.a	n.a
Fecal sampling before or after colonoscopy* (before:after)	30:24 (56%:44%)	21:52 (29%:71%)	0.003295‡	0.016475000
Duration between colonoscopy and fecal sample collection* (days)	-1.5 (-34,106.3438)	19.89097 (-110.7083,247)	0.7586482†	0.875363308
Duration of frozen storage of fecal samples (days)	185.1076 (86.67708,2032)	149 (6.6875,1280)	0.3694857†	0.503844136
FBG	5.1 (4.3,6.9)	5.75 (4.3,13.2)	0.000131342†	0.001407555
TCHO	5 (3.2,6.7)	4.9 (2.6,8.6)	0.299775†	0.503844136
LDL	2.65 (1.4,5.2)	2.9 (0.7,5)	0.9413451†	0.989788600
HDL	1.8 (0.8,3.5)	1.3 (0.5,2.2)	0.000187674†	0.001407555
TG	1 (0.37,2.9)	1.2 (0.5,5.1)	0.01991682†	0.059750460
Cr	71.5 (43,101)	74 (41,202)	0.3257186†	0.503844136
ALT GPT	21 (9,68)	18 (10,69)	0.05068182†	0.126704550
eGFR	69.51 (50.82,115.04)	71.13 (16.81,136.52)	0.9897886†	0.989788600

Supplementary Table S2. Summary of metagenomic data from C1 and mapping to reference gene catalogue. Fourth column reports results from Wilcoxon rank-sum tests.

Parameter	Controls	Cases	*P-value*
Average raw reads	60162577	60496561	0.8082
After removing low quality reads	59423292 (98.77%)	59715967 (98.71%)	0.831
After removing human reads	59380535 ± 7378751	58112890 ± 10324458	0.419
Mapping rate	66.82%	66.27%	0.252

Supplementary Table S3. Gene number and gene alpha diversity of CRC and healthy microbiomes in cohort C1. Diversity was represented by Shannon and Simpson indices.

Parameter	Controls		Cases		P-value
	mean	sd	mean	sd	
Gene number	581635	165527	534440	164184	0.1127
Shannon index	11.699	0.565	11.558	0.546	0.0746
Simpson index	0.99997	2.71E-05	0.99996	3.94E-05	0.0276

Supplementary Table S4. PERMANOVA analysis of microbial gene profiles in cohort C1. The analysis was conducted to test whether clinical parameters and CRC status have significant impact on the gut microbiota with q<0.05. BMI: body mass index; DM: diabetes mellitus type 2; FBG: fasting blood glucose; HDL: high density lipoprotein; TG: triglyceride; eGFR: epidermal growth factor receptor; TNM: tumor node metastasis staging system; TCHO: total cholesterol; Cr: creatinine; LDL; low density lipoprotein. ALT/GPT: alanine transaminase/glutamate pyruvated transaminase.

Parameter	**Df**	**SumsOfSqs**	**MeanSqs**	**F.Model**	**R^2**	**Pr(>F)**	**q-value**
CRC Status	1	0.6792933	0.6792933	1.9596297	0.0153144	0.0004	0.0076
Stage of CRC	4	1.7697175	0.4424294	1.2778364	0.0398977	0.0058	0.0551
Lesion location(1:2:NA)	1	0.464298	0.464298	1.31427	0.020435	0.0536	0.2717
BMI	1	0.4600024	0.4600024	1.3200099	0.0104497	0.0572	0.2717
DM	1	0.4383585	0.4383585	1.257642	0.0098826	0.084	0.285
FBG	1	0.4319269	0.4319269	1.2300105	0.0123955	0.09	0.285
Lesion specific location （1:2:3:4:6:7:8:9:NA）	1	0.421307	0.421307	1.190278	0.018543	0.1369	0.371586
Age	1	0.3972817	0.3972817	1.1387282	0.0089566	0.1923	0.456713
HDL	1	0.3641778	0.3641778	1.0352042	0.010246	0.3578	0.722
eGFR	1	0.3585266	0.3585266	1.0231375	0.0094715	0.38	0.722
TG	1	0.3522642	0.3522642	1.001382	0.0099145	0.4329	0.747736

Duration between colonoscopy and fecal sample collection	1	0.3397823	0.3397823	0.9722612	0.0077181	0.5036	0.761608
Fecal sampling before or after colonoscopy	1	0.3378151	0.3378151	0.9665887	0.0076734	0.5211	0.761608
TNM	15	5.3000663	0.3533378	0.9890377	0.2036857	0.5781	0.766587
Cr	1	0.3281613	0.3281613	0.9330291	0.0088077	0.6052	0.766587
TCHO	1	0.3127842	0.3127842	0.8878167	0.0088	0.7198	0.854763
LDL	1	0.2994855	0.2994855	0.8502487	0.0084308	0.8146	0.863233
ALT/GPT	1	0.2976508	0.2976508	0.847193	0.007929	0.8178	0.863233
Gender	1	0.2677377	0.2677377	0.7651615	0.0060361	0.9528	0.9528

Supplementary Table S5. Principal component analysis (PCA) using microbial gene profiles from cohort C1 using 2.1 million genes. Association tests of the first five principal components (PC) with 17 parameters are reported.

	Gene level *P-value*					
Parameter	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**Statistical test for differences**
CRC Status	**0.029375**	0.7556469	0.0908164	0.2458964	**1.29E-06**	Wilcoxon rank-sum tests

Age	0.1107786	0.7187803	0.9579642	0.3323753	0.1740341	Pearson correlation test
BMI	0.1666538	0.349701	0.2799689	0.9666352	0.6664927	Pearson correlation test
Duration between colonoscopy and fecal sample collection	0.2612967	0.3677261	0.5027833	0.8471867	0.985353	Pearson correlation test
Fecal sampling before or after colonoscopy	0.3051672	0.3564576	0.6633822	0.998038	0.2695479	Wilcoxon rank-sum tests
DM	0.3304729	0.7684188	0.192732	0.4910126	**0.025695**	Wilcoxon rank-sum tests
TNM	0.3587305	0.7179382	0.4123964	0.6422653	0.2646984	Kruskal-Wallis tests
Gender	0.3762511	0.692509	0.0652127	0.6280812	0.4261203	Wilcoxon rank-sum tests
TCHO	0.3918745	0.6337139	0.8437887	0.8920492	0.8586685	Pearson correlation test
LDL	0.3996362	0.3026439	0.2289333	0.8912377	0.5012763	Pearson correlation test
eGFR	0.4185351	0.6904019	0.4945847	0.3171986	0.5644339	Pearson correlation test
Stage of CRC	0.4785966	0.4859963	0.3868685	0.6319499	0.1400903	Kruskal-Wallis tests
HDL	0.4855939	0.265294	0.9413181	0.8985499	0.1237575	Pearson correlation test
TG	0.5435062	0.7623276	0.4072886	0.1106054	0.0247417	Pearson correlation test
ALT/GPT	0.6686028	0.819014	0.5737057	0.3283116	0.6117176	Pearson correlation test
Cr	0.8059999	0.5986523	0.743814	0.7723353	0.3177772	Pearson correlation test
FBG	0.8765164	0.6637887	**0.037985**	0.8754288	0.0596181	Pearson correlation test

Supplementary Table S6. List of KEGG modules and pathways associated with CRC status at P-value<0.01 in cohort C1.

KEGG Modules						
Module ID	**Control rank mean**	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**	**Definition**
M00036	48.72222222	76.01351351	0	4.36E-05	0.014810703	Leucine degradation, leucine => acetoacetate + acetyl-CoA
M00050	50.2962963	74.86486486	0	0.000141552	0.02406378	Guanine nucleotide biosynthesis, IMP => GDP/dGDP,GTP/dGTP
M00037	51.72222222	73.82432432	0	0.000655997	0.056363614	Melatonin biosynthesis, tryptophan => serotonin => melatonin
M00042	52.05555556	73.58108108	0	0.000663101	0.056363614	Catecholamine biosynthesis, tyrosine => dopamine => noradrenaline => adrenaline
M00020	53.07407407	72.83783784	0	0.002447934	0.142598694	Serine biosynthesis, glycerate-3P => serine
M00046	54.81481481	71.56756757	0	0.003088118	0.142598694	beta-Alanine biosynthesis, cytosine / uracil => beta-alanine
M00055	77.38888889	55.09459459	1	0.003404079	0.142598694	N-glycan precursor biosynthesis
M00250	76.40740741	55.81081081	1	0.003804616	0.142598694	Lipopolysaccharide transport system
M00135	53.09259259	72.82432432	0	0.004435999	0.142598694	GABA biosynthesis, eukaryotes, putrescine => GABA
M00144	52.27777778	73.41891892	0	0.004478832	0.142598694	Complex I (NADH dehydrogenase), NADH dehydrogenase I
M00267	76.2962963	55.89189189	1	0.004613487	0.142598694	PTS system, N-acetylglucosamine-specific II component

M00117	53.53703704	72.5	0	0.00506723	0.143571525	Ubiquinone biosynthesis, prokaryotes, chorismate => ubiquinone
M00319	55.03703704	71.40540541	0	0.005986981	0.156582578	Manganese/zinc/iron transport system
M00045	53.81481481	72.2972973	0	0.006759204	0.164152094	Histidine degradation, histidine => N-formiminoglutamate => glutamate
M00318	56.85185185	70.08108108	0	0.007585097	0.171928858	Iron/zinc/copper transport system
M00209	53.72222222	72.36486486	0	0.009495418	0.20177763	Osmoprotectant transport system
KEGG Pathways						
Map ID	**Control rank mean**	**Case rank mean**	**Enrichment(0:case/ 1:control)**	**P-value**	**q-value**	**Definition**
map00901	51.72222222	73.82432432	0	0.000655997	0.093807599	Indole alkaloid biosynthesis
map00965	51.72222222	73.82432432	0	0.000655997	0.093807599	Betalain biosynthesis
map00943	76.35185185	55.85135135	1	0.001077093	0.102682879	Isoflavonoid biosynthesis
map00253	76.75925926	55.55405405	1	0.002345148	0.167678104	Tetracycline biosynthesis
map00190	52.53703704	73.22972973	0	0.003379124	0.177259289	Oxidative phosphorylation
map00430	52.7962963	73.04054054	0	0.003718726	0.177259289	Taurine and hypotaurine metabolism
map00280	54.22222222	72	0	0.006119571	0.222185359	Valine, leucine and isoleucine degradation

map04724	54.68518519	71.66216216	0	0.006639056	0.222185359	Glutamatergic synapse
map00562	54.07407407	72.10810811	0	0.008048415	0.222185359	Inositol phosphate metabolism
map00061	74.27777778	57.36486486	1	0.008239217	0.222185359	Fatty acid biosynthesis
map00910	55.16666667	71.31081081	0	0.008545591	0.222185359	Nitrogen metabolism
map04940	54.53703704	71.77027027	0	0.009490601	0.226192666	Type I diabetes mellitus

Supplementary Table S7. List of KEGG orthologous groups (KOs) associated with CRC status at q-value <0.05 in cohort C1.

KO ID	Control rank mean	Case rank mean	Enrichment (0:case/1:control)	P-value	q-value	Definition
K09778	46.68519	77.5	0	2.91E-06	0.017649179	Hypothetical protein
K10670	46.44444	77.67568	0	6.94E-06	0.020377011	Glycine reductase
K09065	49.01852	75.7973	0	1.15E-05	0.020377011	N-acetylornithine carbamoyltransferase
K13772	47.46296	76.93243	0	1.34E-05	0.020377011	Rrf2 family transcriptional regulator, iron-responsive regulator
K01464	49.27778	75.60811	0	2.35E-05	0.022105986	Dihydropyrimidinase
K02656	81.26852	52.26351	1	2.51E-05	0.022105986	Type IV pilus assembly protein PilF
K08286	81.05556	52.41892	1	2.55E-05	0.022105986	Protein-serine/threonine kinase

K01096	80.68519	52.68919	1	4.04E-05	0.029007909	Phosphatidylglycerophosphatase B
K00087	49.61111	75.36486	0	4.56E-05	0.029007909	Xanthine dehydrogenase molybdenum-binding subunit
K05020	48.07407	76.48649	0	5.26E-05	0.029007909	Glycine betaine transporter
K07301	81.09259	52.39189	1	5.43E-05	0.029007909	Inner membrane protein
K01318	79.22222	53.75676	1	5.73E-05	0.029007909	Glutamyl endopeptidase
K11786	80.85185	52.56757	1	6.74E-05	0.031479285	ATP-dependent helicase STH1/SNF2
K01951	49.25926	75.62162	0	7.32E-05	0.031758681	GMP synthase (glutamine-hydrolysing)
K01459	78.51852	54.27027	1	1.31E-04	0.049313518	NA
K00132	50.64815	74.60811	0	1.38E-04	0.049313518	Acetaldehyde dehydrogenase (acetylating)
K04835	51.12963	74.25676	0	1.44E-04	0.049313518	Methylaspartate ammonia-lyase
K11337	49.16667	75.68919	0	1.51E-04	0.049313518	3-hydroxyethyl bacteriochlorophyllide a dehydrogenase
K04749	49.11111	75.72973	0	1.54E-04	0.049313518	Anti-sigma B factor antagonist

Supplementary Table S8. IMG, mOTU and MLG species associated with CRC with q-value < 0.05 in cohort C1. 86 MLG species were formed after grouping 106 MLGs with more than 100 genes using species annotation when available. MLG species identifiers starting with "Con_" are enriched in control samples, and those starting with "CRC_" are enriched in CRC samples.

28 IMG species					
	Control rank mean	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**
Peptostreptococcus stomatis	37.25926	84.37838	0	5.11E-12	1.32E-08
Parvimonas micra	38.43519	83.52027	0	4.21E-11	5.43E-08
Parvimonas sp. oral taxon 393	39.81481	82.51351	0	2.79E-10	2.40E-07
Parvimonas sp. oral taxon 110	43.52778	79.80405	0	6.17E-08	3.98E-05
Gemella morbillorum	43.87037	79.55405	0	1.53E-07	7.88E-05
Fusobacterium nucleatum	45.09259	78.66216	0	3.86E-07	1.56E-04
Leptotrichia buccalis	45.60185	78.29054	0	4.44E-07	1.56E-04
Fusobacterium sp. oral taxon 370	45.02778	78.70946	0	4.83E-07	1.56E-04
Burkholderia mallei	45.19444	78.58784	0	7.93E-07	2.27E-04
Prevotella intermedia	46.47222	77.65541	0	1.92E-06	4.95E-04

Streptococcus pseudoporcinus	47.5	76.90541	0	4.03E-06	8.99E-04
Streptococcus dysgalactiae	47.06481	77.22297	0	4.18E-06	8.99E-04
Beggiatoa sp. PS	46.53704	77.60811	0	5.03E-06	9.97E-04
Malassezia globosa	46.35185	77.74324	0	8.71E-06	1.60E-03
Paracoccus denitrificans	47.48148	76.91892	0	1.18E-05	2.02E-03
Eubacterium ventriosum	80.98148	52.47297	1	1.27E-05	2.05E-03
Streptococcus constellatus	48.2037	76.39189	0	1.66E-05	2.52E-03
Filifactor alocis	49.06481	75.76351	0	3.94E-05	5.65E-03
Peptoniphilus indolicus	51.2963	74.13514	0	4.53E-05	6.14E-03
Crenothrix polyspora	48.76852	75.97973	0	5.14E-05	6.63E-03
Peptostreptococcus anaerobius	50.14815	74.97297	0	5.88E-05	7.22E-03
Streptococcus equi	50.58333	74.65541	0	6.91E-05	8.10E-03
Solobacterium moorei	47.66667	76.78378	0	8.79E-05	9.85E-03
Sulfurovum sp. SCGC AAA036-O23	52.12037	73.53378	0	1.28E-04	1.37E-02
Streptobacillus moniliformis	52.35185	73.36486	0	1.44E-04	1.49E-02

Eubacteriaceae bacterium ACC19a	51.87037	73.71622	0	1.93E-04	1.92E-02
Fusobacterium necrophorum	52.37037	73.35135	0	3.72E-04	3.55E-02
Adhaeribacter aquaticus	77.06481	55.33108	1	4.79E-04	4.41E-02
21 mOTU species					
	Control rank mean	Case rank mean	Enrichment(0:case /1:control)	P-value	q-value
Parvimonas micra	46.2963	77.78378	0	2.31E-08	7.73E-06
Peptostreptococcus stomatis	46.25	77.81757	0	2.81E-08	7.73E-06
motu_linkage_group_731	50.42593	74.77027	0	2.91E-07	5.33E-05
Gemella morbillorum	47.93519	76.58784	0	8.63E-07	1.18E-04
motu_linkage_group_407	81.13889	52.35811	1	8.51E-06	9.34E-04
motu_linkage_group_490	80.46296	52.85135	1	3.04E-05	2.78E-03
Fusobacterium nucleatum	54.62037	71.70946	0	3.56E-05	2.79E-03
Clostridium symbiosum	48.66667	76.05405	0	4.50E-05	2.99E-03
motu_linkage_group_443	79.66667	53.43243	1	4.91E-05	2.99E-03
motu_linkage_group_316	79.61111	53.47297	1	7.03E-05	3.86E-03

Eubacterium ventriosum	78.09259	54.58108	1	9.82E-05	4.90E-03
Solobacterium moorei	51.22222	74.18919	0	2.49E-04	1.14E-02
Bacteroides fragilis	51.09259	74.28378	0	3.75E-04	1.58E-02
unclassified Fusobacterium	54.22222	72	0	4.20E-04	1.59E-02
Clostridiales bacterium 1_7_47FAA	51.27778	74.14865	0	4.34E-04	1.59E-02
Clostridium ramosum	50.92593	74.40541	0	5.21E-04	1.75E-02
motu_linkage_group_611	77.2963	55.16216	1	5.50E-04	1.75E-02
Prevotella nigrescens	58.09259	69.17568	0	5.72E-04	1.75E-02
motu_linkage_group_624	51.01852	74.33784	0	1.33E-03	3.69E-02
motu_linkage_group_510	77.84259	54.76351	1	1.35E-03	3.69E-02
Clostridium bolteae	51.81481	73.75676	0	1.41E-03	3.69E-02
85 MLG species					
	Control rank mean	Case rank mean	Enrichment(0:case /1:control)	P-value	q-value
Parvimonas micra	38.40741	83.54054	0	5.56E-12	4.84E-10
Fusobacterium nucleatum	40.32407	82.14189	0	1.72E-10	7.48E-09

Solobacterium moorei	42.2037	80.77027	0	4.01E-08	1.16E-06
Clostridium symbiosum	46.31481	77.77027	0	2.67E-06	5.80E-05
Con 10180	82.03704	51.7027	1	6.06E-06	1.05E-04
CRC 2881	51.25926	74.16216	0	7.57E-06	1.10E-04
CRC 2794	51.03704	74.32432	0	1.04E-05	1.30E-04
Coprococcus sp. ART55/1	80.85185	52.56757	1	2.09E-05	2.05E-04
Clostridium hathewayi	46.77778	77.43243	0	2.12E-05	2.05E-04
Clostridiales bacterium 1_7_47FAA	48.16667	76.41892	0	2.49E-05	2.17E-04
CRC 4136	50.99074	74.35811	0	2.97E-05	2.32E-04
butyrate-producing bacterium SS3/4	80.57407	52.77027	1	3.19E-05	2.32E-04
Haemophilus parainfluenzae	80.49074	52.83108	1	4.18E-05	2.69E-04
Con 154	80.35185	52.93243	1	4.45E-05	2.69E-04
Clostridium clostridioforme	50.2037	74.93243	0	4.64E-05	2.69E-04
Bacteroides fragilis	49.09259	75.74324	0	5.56E-05	3.02E-04
Con 1979	79.94444	53.22973	1	6.03E-05	3.09E-04

Eubacterium ventriosum	78.62963	54.18919	1	6.88E-05	3.33E-04
Con 7958	75.27778	56.63514	1	7.40E-05	3.33E-04
Con 5770	79.39815	53.62838	1	7.66E-05	3.33E-04
Clostridium sp. HGF2	48.27778	76.33784	0	8.28E-05	3.43E-04
CRC 6481	52.09259	73.55405	0	9.87E-05	3.90E-04
Cloacibacillus evryensis	52.73148	73.08784	0	1.13E-04	4.23E-04
Con 1987	79.42593	53.60811	1	1.17E-04	4.23E-04
Con 4595	77.21296	55.22297	1	1.38E-04	4.81E-04
Con 1617	76.12963	56.01351	1	1.50E-04	5.03E-04
Con 1371	78.46296	54.31081	1	2.05E-04	6.60E-04
Lachnospiraceae bacterium 5_1_57FAA	49.96296	75.10811	0	2.49E-04	7.73E-04
Eubacterium biforme	74.68519	57.06757	1	3.00E-04	8.70E-04
Faecalibacterium prausnitzii	78.25926	54.45946	1	3.00E-04	8.70E-04
Con 4699	78.78704	54.07432	1	3.13E-04	8.79E-04
Desulfovibrio sp. 6_1_46AFAA	53.33333	72.64865	0	3.70E-04	9.87E-04

Con 1529	75.05556	56.7973	1	3.74E-04	9.87E-04
Ruminococcus torques	76.92593	55.43243	1	5.28E-04	1.35E-03
Coprobacillus sp. 3_3_56FAA	50.53704	74.68919	0	6.01E-04	1.46E-03
Streptococcus equinus	54.52778	71.77703	0	6.02E-04	1.46E-03
Synergistes sp. 3_1_syn1	54.37963	71.88514	0	6.89E-04	1.62E-03
Lachnospiraceae bacterium 8_1_57FAA	51.88889	73.7027	0	7.91E-04	1.81E-03
Klebsiella pneumoniae	74.7037	57.05405	1	8.33E-04	1.86E-03
Eubacterium eligens	79.53704	53.52703	1	9.07E-04	1.97E-03
Clostridium bolteae	51.39815	74.06081	0	9.27E-04	1.97E-03
Con 1513	76.59259	55.67568	1	1.02E-03	2.11E-03
Clostridium citroniae	51.71296	73.83108	0	1.08E-03	2.19E-03
Fusobacterium varium	54.57407	71.74324	0	1.15E-03	2.28E-03
Bacteroides clarus	75.55556	56.43243	1	1.29E-03	2.50E-03
Ruminococcus obeum	77.53704	54.98649	1	1.34E-03	2.54E-03
Con 2606	77.5	55.01351	1	1.42E-03	2.59E-03

Lachnospiraceae bacterium 3_1_46FAA	52.53704	73.22973	0	1.44E-03	2.59E-03
CRC 2867	52.31481	73.39189	0	1.46E-03	2.59E-03
Con 6037	77.5463	54.97973	1	1.56E-03	2.71E-03
Clostridium sp. L2-50	76.37963	55.83108	1	1.61E-03	2.75E-03
Con 1867	76.38889	55.82432	1	2.13E-03	3.57E-03
Roseburia intestinalis	76.99074	55.38514	1	2.20E-03	3.58E-03
Subdoligranulum sp. 4_3_54A2FAA	51.56481	73.93919	0	2.24E-03	3.58E-03
Con 1197	75.42593	56.52703	1	2.26E-03	3.58E-03
CRC 4069	53.7963	72.31081	0	2.56E-03	3.96E-03
Con 8757	77.17593	55.25	1	2.60E-03	3.96E-03
Con 5752	73.65741	57.81757	1	2.71E-03	4.07E-03
Con 4295	74.98148	56.85135	1	2.95E-03	4.34E-03
Eubacterium rectale	75.90741	56.17568	1	3.21E-03	4.60E-03
Con 2494	74.35185	57.31081	1	3.22E-03	4.60E-03
Con 7367	76.23148	55.93919	1	3.63E-03	5.09E-03

Con 4829	76.7963	55.52703	1	3.88E-03	5.35E-03
Con 356	75.94444	56.14865	1	3.95E-03	5.37E-03
Dorea formicigenerans	52.98148	72.90541	0	4.36E-03	5.84E-03
Con 10559	76.59259	55.67568	1	4.52E-03	5.91E-03
Con 563	72.7037	58.51351	1	4.55E-03	5.91E-03
Con 4909	75.72222	56.31081	1	4.79E-03	6.12E-03
Con 6128	76.22222	55.94595	1	4.86E-03	6.13E-03
Con 2503	74.14815	57.45946	1	6.02E-03	7.46E-03
CRC 3579	54.05556	72.12162	0	6.09E-03	7.46E-03
Con 2703	74.55556	57.16216	1	7.67E-03	9.15E-03
Con 6068	75.74074	56.2973	1	7.67E-03	9.15E-03
Con 1604	71.92593	59.08108	1	8.96E-03	1.05E-02
Con 5615	76.07407	56.05405	1	9.70E-03	1.12E-02
Lachnospiraceae bacterium 3_1_57FAA_CT1	54.07407	72.10811	0	1.04E-02	1.19E-02
Con 569	73.41667	57.99324	1	1.30E-02	1.46E-02

Con 631	70.01852	60.47297	1	1.31E-02	1.46E-02
Con 1241	76.27778	55.90541	1	1.46E-02	1.61E-02
Alistipes indistinctus	54.50926	71.79054	0	1.59E-02	1.72E-02
Con 8420	72.64815	58.55405	1	2.32E-02	2.48E-02
Burkholderiales bacterium 1_1_47	72.37963	58.75	1	2.34E-02	2.48E-02
Con 7993	73.74074	57.75676	1	3.01E-02	3.16E-02
Con 425	73.19444	58.15541	1	3.87E-02	4.01E-02
Con 561	70.5	60.12162	1	4.81E-02	4.92E-02

Supplementary Table S9. PERMANOVA analysis of variation in three CRC-enriched species measured by three different methods in cohort C1. CRC- and colonoscopy-related factors explain the variation in these three species.

Parameter	**Df**	**mOTU species**					**IMG species**					**MLG species**				
		SumsOf Sqs	**MeanSqs**	**F. Model**	**R^2**	**Pr(>F)**	**SumsOf Sqs**	**MeanSqs**	**F.Model**	**R^2**	**Pr(>F)**	**SumsOf Sqs**	**MeanSqs**	**F.Model**	**R^2**	**Pr(>F)**
CRC Status	1	5.85E-05	5.85E-05	5.1835238	0.0395135	**0.0076**	2.42E-04	2.42E-04	4.2189512	0.0323989	**0.0127**	7.02E-03	7.02E-03	5.9492807	0.0450876	**0.0072**
Duration between colonoscopy and fecal	1	4.05E-05	4.05E-05	3.5159771	0.0273583	**0.0523**	1.57E-04	1.57E-04	2.6787139	0.0209801	**0.0777**	4.25E-03	4.25E-03	3.5265637	0.0274384	**0.0569**

sample collection																
Fecal sampling before or after colonoscopy	1	3.21E-05	3.21E-05	2.7722393	0.0216967	**0.0799**	1.12E-04	1.12E-04	1.8992995	0.0149670	0.163	3.54E-03	3.54E-03	2.9217093	0.0228398	**0.0799**
Stage of CRC	4	8.38E-05	2.09E-05	1.8432688	0.0565537	0.1262	4.44E-04	1.11E-04	1.9437773	0.0594540	0.1157	1.27E-02	3.17E-03	2.7293564	0.0815236	**0.0354**
Lesion location	1	3.02E-05	3.02E-05	1.5272855	0.0236688	0.1846	1.28E-04	1.28E-04	1.2152307	0.0189243	0.1988	2.27E-03	2.27E-03	1.0493068	0.0163828	0.3215
LDL	1	2.03E-05	2.03E-05	1.4217908	0.0140186	0.2414	2.52E-05	2.52E-05	0.3436566	0.0034248	0.5793	6.77E-04	6.77E-04	0.4524804	0.0045044	0.5249
eGFR	1	5.78E-06	5.78E-06	0.4256440	0.0039622	0.5138	4.77E-06	4.77E-06	0.0692402	0.0006467	0.8438	3.31E-04	3.31E-04	0.2318740	0.0021624	0.6453
TCHO	1	1.24E-05	1.24E-05	0.8618039	0.0085444	0.3454	7.84E-06	7.84E-06	0.1067080	0.0010659	0.7915	2.81E-04	2.81E-04	0.1872153	0.0018687	0.6821
Lesion specific location	1	4.15E-06	4.15E-06	0.2052181	0.0032469	0.6648	1.41E-06	1.41E-06	0.0131386	0.0002085	0.9754	8.14E-05	8.14E-05	0.0370280	0.0005874	0.9353
HDL	1	3.24E-07	3.24E-07	0.0222985	0.0002229	0.9401	4.69E-06	4.69E-06	0.0638119	0.0006377	0.8687	3.50E-05	3.50E-05	0.0232691	0.0002326	0.955
Age	1	1.75E-07	1.75E-07	0.0148715	0.0001180	0.9652	3.05E-06	3.05E-06	0.0515304	0.0004088	0.8841	3.47E-05	3.47E-05	0.0280829	0.0002228	0.9507
FBG	1	4.03E-06	4.03E-06	0.2850014	0.0028997	0.5725	1.73E-05	1.73E-05	0.2322323	0.0023641	0.6205	1.70E-03	1.70E-03	1.1175736	0.0112752	0.2544
BMI	1	1.41E-06	1.41E-06	0.1195008	0.0009551	0.749	1.07E-05	1.07E-05	0.1801544	0.0014392	0.6958	8.11E-05	8.11E-05	0.0651803	0.0005212	0.8618
Cr	1	2.32E-06	2.32E-06	0.1668589	0.0015866	0.6698	3.16E-06	3.16E-06	0.0449746	0.0004281	0.8759	1.61E-04	1.61E-04	0.1103230	0.0010496	0.7615
ALT/GPT	1	8.01E-07	8.01E-07	0.0625344	0.0005896	0.8156	6.22E-06	6.22E-06	0.0929296	0.0008759	0.7813	5.69E-04	5.69E-04	0.4106836	0.0038594	0.4907
TNM	15	5.83E-05	3.89E-06	0.1815751	0.0448528	0.9841	3.68E-04	2.46E-05	0.2193220	0.0536766	0.9134	1.15E-02	7.68E-04	0.3435946	0.0816089	0.8323

TG	1	3.80E-07	3.80E-07	0.0261886	0.0002618	0.9144	6.05E-07	6.05E-07	0.0082320	0.0000823	0.9827	1.39E-04	1.39E-04	0.0922060	0.0009212	0.7912
Gender	1	1.07E-06	1.07E-06	0.0908585	0.0007206	0.8475	9.10E-06	9.10E-06	0.1537437	0.0012187	0.8233	1.65E-04	1.65E-04	0.1336220	0.0010594	0.7801
DM	1	5.19E-07	5.19E-07	0.0441774	0.0003505	0.9158	4.74E-06	4.74E-06	0.0800697	0.0006351	0.8975	2.34E-04	2.34E-04	0.1895356	0.0015020	0.7209

Supplementary Table S10. List of 13 genera associated with CRC status in cohort C1.

	Control rank mean	**Case rank mean**	**Enrichment(0:case/1:control)**	**P-value**	**q-value**
Parvimonas	38.55556	83.43243	0	3.97E-11	3.86E-08
Peptostreptococcus	40.55556	81.97297	0	5.49E-10	2.67E-07
Fusobacterium	45.51852	78.35135	0	6.90E-07	2.24E-04
Beggiatoa	45.89815	78.07432	0	1.78E-06	4.34E-04
Malassezia	46.35185	77.74324	0	8.71E-06	1.70E-03
Paracoccus	47.66667	76.78378	0	1.10E-05	1.79E-03
Leptotrichia	48.15741	76.42568	0	3.40E-05	4.74E-03
Filifactor	49.06481	75.76351	0	3.94E-05	4.80E-03
Crenothrix	48.76852	75.97973	0	5.14E-05	5.57E-03

Solobacterium	47.66667	76.78378	0	8.79E-05	8.56E-03
Sulfurovum	49.48148	75.45946	0	1.14E-04	9.64E-03
Eubacterium	80.07407	53.13514	1	1.19E-04	9.64E-03
Streptobacillus	52.35185	73.36486	0	1.44E-04	1.08E-02
Adhaeribacter	77.06481	55.33108	1	4.79E-04	3.33E-02
Moniliophthora	49.91667	75.14189	0	6.39E-04	4.15E-02

Supplementary Table S11. List of phyla significantly associating with CRC status in cohort C1.

Phylum	**Control rank mean**	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**
Fusobacteria	44.68519	78.95946	0	0.00000014	**0.000005**
Firmicutes	73.44444	57.97297	1	0.02924627	0.259876
Cloacimonetes	69.25926	61.02703	1	0.03419421	0.259876

Supplementary Table S12. IMG, mOTU and MLG species markers. IMG, mOTU and MLG species markers identified using random forest method among species associated with CRC (**Supplementary Table 8**). Marker species are listed by their importance reported by the method. MLG species identifiers starting

with “Con_” are enriched in control samples, and those starting with “CRC_” are enriched in CRC samples.

17 IMG species markers					
	Control rank mean	**Case rank mean**	**Enrichment (0:case/1:control)**	**P-value**	**q-value**
Peptostreptococcus stomatis	37.25926	84.37838	0	5.11E-12	1.32E-08
Parvimonas micra	38.43519	83.52027	0	4.21E-11	5.43E-08
Parvimonas sp. oral taxon 393	39.81481	82.51351	0	2.79E-10	2.40E-07
Parvimonas sp. oral taxon 110	43.52778	79.80405	0	6.17E-08	3.98E-05
Gemella morbillorum	43.87037	79.55405	0	1.53E-07	7.88E-05
Fusobacterium nucleatum	45.09259	78.66216	0	3.86E-07	1.56E-04
Leptotrichia buccalis	45.60185	78.29054	0	4.44E-07	1.56E-04
Fusobacterium sp. oral taxon 370	45.02778	78.70946	0	4.83E-07	1.56E-04
Burkholderia mallei	45.19444	78.58784	0	7.93E-07	2.27E-04
Prevotella intermedia	46.47222	77.65541	0	1.92E-06	4.95E-04
Streptococcus dysgalactiae	47.06481	77.22297	0	4.18E-06	8.99E-04
Beggiatoa sp. PS	46.53704	77.60811	0	5.03E-06	9.97E-04

Malassezia globosa	46.35185	77.74324	0	8.71E-06	1.60E-03
Paracoccus denitrificans	47.48148	76.91892	0	1.18E-05	2.02E-03
Eubacterium ventriosum	80.98148	52.47297	1	1.27E-05	2.05E-03
Filifactor alocis	49.06481	75.76351	0	3.94E-05	5.65E-03
Solobacterium moorei	47.66667	76.78378	0	8.79E-05	9.85E-03
7 mOTU species markers					
	Control rank mean	**Case rank mean**	**Enrichment(0:case/1:control)**	**P-value**	**q-value**
Gemella morbillorum	47.93518519	76.58783784	0	8.63E-07	1.18E-04
Parvimonas micra	46.2962963	77.78378378	0	2.31E-08	7.73E-06
Peptostreptococcus stomatis	46.25	77.81756757	0	2.81E-08	7.73E-06
motu_linkage_group_316	79.61111111	53.47297297	1	7.03E-05	3.86E-03
motu_linkage_group_407	81.13888889	52.35810811	1	8.51E-06	9.34E-04
motu_linkage_group_490	80.46296296	52.85135135	1	3.04E-05	2.78E-03
motu_linkage_group_624	51.01851852	74.33783784	0	1.33E-03	3.69E-02
27 MLG species markers					

	Control rank mean	Case rank mean	Enrichment(0:case/1:control)	P-value	q-value
Parvimonas micra	38.40741	83.54054	0	5.56E-12	4.84E-10
Fusobacterium nucleatum	40.32407	82.14189	0	1.72E-10	7.48E-09
Solobacterium moorei	42.2037	80.77027	0	4.01E-08	1.16E-06
Clostridium symbiosum	46.31481	77.77027	0	2.67E-06	5.80E-05
Con_10180	82.03704	51.7027	1	6.06E-06	1.05E-04
CRC_2881	51.25926	74.16216	0	7.57E-06	1.10E-04
Coprococcus sp. ART55/1	80.85185	52.56757	1	2.09E-05	2.05E-04
Clostridium hathewayi	46.77778	77.43243	0	2.12E-05	2.05E-04
Clostridiales bacterium 1_7_47FAA	48.16667	76.41892	0	2.49E-05	2.17E-04
CRC_4136	50.99074	74.35811	0	2.97E-05	2.32E-04
butyrate-producing bacterium SS3/4	80.57407	52.77027	1	3.19E-05	2.32E-04
Haemophilus parainfluenzae	80.49074	52.83108	1	4.18E-05	2.69E-04
Con_154	80.35185	52.93243	1	4.45E-05	2.69E-04
Bacteroides fragilis	49.09259	75.74324	0	5.56E-05	3.02E-04

Con_1979	79.94444	53.22973	1	6.03E-05	3.09E-04
Con_7958	75.27778	56.63514	1	7.40E-05	3.33E-04
Con_5770	79.39815	53.62838	1	7.66E-05	3.33E-04
CRC_6481	52.09259	73.55405	0	9.87E-05	3.90E-04
Con_1987	79.42593	53.60811	1	1.17E-04	4.23E-04
Con_4595	77.21296	55.22297	1	1.38E-04	4.81E-04
Eubacterium biforme	74.68519	57.06757	1	3.00E-04	8.70E-04
Desulfovibrio sp. 6_1_46AFAA	53.33333	72.64865	0	3.70E-04	9.87E-04
Clostridium citroniae	51.71296	73.83108	0	1.08E-03	2.19E-03
Fusobacterium varium	54.57407	71.74324	0	1.15E-03	2.28E-03
Roseburia intestinalis	76.99074	55.38514	1	2.20E-03	3.58E-03
Dorea formicigenerans	52.98148	72.90541	0	4.36E-03	5.84E-03
CRC_3579	54.05556	72.12162	0	6.09E-03	7.46E-03

Supplementary Table S13. 20 gene markers identified by the mRMR feature selection method in cohort C1. Detailed information regarding their enrichment, occurrence in CRC cases and controls, statistical test of association, taxonomy and identity percentage are listed.

Marker gene id	Enrich-ment	Wilcoxon rank-sum test		Occurrence				Identity	Taxonomy (Blastn to IMG v400)	Description (Blastp to KEGG v59)
				Control (n=54)		Case (n=74)				
		P-value	q-value	N	Rate(%)	N	Rate(%)			
2361423	Case	2.31E-13	4.88E-07	11	20.37037037	62	83.78378378	93.87	*Peptostreptococcus anaerobius*	transposase
3173495	Case	6.24E-13	6.58E-07	10	18.51851852	61	82.43243243	93.98	*Peptostreptococcus anaerobius*	transposase
2040133	Case	7.51E-10	4.06E-04	14	25.92592593	62	83.78378378	99.4	*Clostridium symbiosum*	cobalt/nickel transport system permease protein
1696299	Case	7.70E-10	4.06E-04	2	3.703703704	43	58.10810811	99.78	*Parvimonas micra*	DNA-directed RNA polymerase subunit beta
482585	Case	7.41E-09	1.05E-03	16	29.62962963	58	78.37837838	NA	NA	RNA-directed DNA polymerase
2211919	Control	4.98E-08	2.20E-03	49	90.74074074	47	63.51351351	80.99	*Coprobacillus* sp. 8_2_54BFAA	NA
4171064	Control	7.50E-08	2.61E-03	40	74.07407407	18	24.32432432	94.94	*Faecalibacterium prausnitzii*	cytidine deaminase
1704941	Case	7.53E-08	2.61E-03	2	3.703703704	39	52.7027027	99.13	*Fusobacterium nucleatum*	butyryl-CoA dehydrogenase

3319526	Control	1.08E-07	2.79E-03	32	59.25925926	10	13.51351351	90.01	*Faecalibacterium prausnitzii*	NA
3246804	Case	1.80E-07	3.24E-03	1	1.851851852	35	47.2972973	NA	NA	citrate-Mg2+:H+ or citrate-Ca2+:H+ symporter, CitMHS family
3976414	Control	4.42E-07	4.07E-03	30	55.55555556	9	12.16216216	87.12	*Faecalibacterium prausnitzii*	adenosylcobinamide-phosphate synthase CobD
4256106	Control	7.39E-07	4.53E-03	28	51.85185185	9	12.16216216	NA	NA	integrase/recombinase XerD
3531210	Control	1.44E-06	5.63E-03	13	24.07407407	0	0	NA	NA	GDP-L-fucose synthase
3611706	Control	1.68E-06	5.82E-03	15	27.77777778	0	0	NA	NA	anti-repressor protein
2206475	Control	1.81E-06	5.95E-03	28	51.85185185	9	12.16216216	98.59	*Eubacterium ventriosum*	beta-glucosidase
181682	Control	1.95E-06	6.09E-03	34	62.96296296	15	20.27027027	99.25	*Roseburia intestinalis*	NA
1804565	Control	2.03E-06	6.16E-03	22	40.74074074	4	5.405405405	NA	NA	branched-chain amino acid transport system ATP-binding protein
2736705	Case	5.71E-06	8.55E-03	2	3.703703704	32	43.24324324	99.68	*Clostridium hathewayi*	NA
1559769	Control	1.03E-05	1.04E-02	27	50	7	9.459459459	88.65	*Coprococcus catus*	polar amino acid transport system substrate-binding protein
370640	Control	2.64E-05	1.47E-02	14	25.92592593	0	0	99.4	*Bacteroides clarus*	NA

Supplementary Table S14. PERMANOVA analysis of variation in 20 CRC-associated gene markers in cohort C1. CRC status and stage explain the variation in these gene profiles, while fasting blood glucose (FBG) moderately explains the variation. See **Supplementary Table S4** for explanation of parameters in column 1.

Parameter	Df	SumsOfSqs	MeanSqs	F.Model	R^2	Pr(>F)	q-value
CRC Status	1	5.5793661	5.5793661	16.626711	0.116575	0.0001	0.00095
Stage of CRC	4	6.7812635	1.6953159	5.0761083	0.1416874	0.0001	0.00095
FBG	1	0.8119553	0.8119553	2.154786	0.0215146	0.0073	0.046233
Fecal sampling before or after colonoscopy	1	0.5473702	0.5473702	1.4588296	0.011536	0.0978	0.46455
Lesion location	1	0.500106	0.500106	1.4185104	0.0220202	0.1329	0.486163
Lesion specific location	7	2.7831853	0.3975979	1.1372468	0.1225468	0.1889	0.486163
HDL	1	0.4718905	0.4718905	1.2480119	0.0123263	0.203	0.486163
ALT/GPT	1	0.4650084	0.4650084	1.2366953	0.0115324	0.2047	0.486163
Duration between colonoscopy and fecal sample collection	1	0.4170429	0.4170429	1.1084063	0.0087893	0.3116	0.657822
Age	1	0.3976816	0.3976816	1.0557238	0.0083091	0.3669	0.676838
TCHO	1	0.3768657	0.3768657	0.9942006	0.0098441	0.4287	0.676838
DM	1	0.3653642	0.3653642	0.9692711	0.0076339	0.4617	0.676838
BMI	1	0.3660728	0.3660728	0.9708139	0.0077067	0.4631	0.676838

Cr	1	0.3412225	0.3412225	0.8963725	0.0084646	0.5617	0.719847
TNM	15	5.2686733	0.3512449	0.9797038	0.2021521	0.5683	0.719847
LDL	1	0.308397	0.308397	0.8136124	0.0080705	0.6624	0.741782
Gender	1	0.3092058	0.3092058	0.8193202	0.0064605	0.6637	0.741782
TG	1	0.291975	0.291975	0.7695216	0.0076365	0.7334	0.774144
eGFR	1	0.2043621	0.2043621	0.539403	0.0050159	0.9496	0.9496

Supplementary Table S15. CRC index estimated in cohort C1, a type 2 diabetes (T2D) cohort and an inflammatory bowel disease (IBD) cohort.

Cohort/group	**Median CRC index**	**Comparison with C1 patients**	
		***P*-value**	**q-value**
C1 patients	7.30636	NA	NA
C1 controls	-5.558923	3.91E-21	4.89E-21
T2D patients	0.2512602	1.71E-26	2.85E-26
T2D controls	-1.47849	2.00E-30	1.00E-29
IBD patients	-1.789305	6.00E-11	6.00E-11

IBD controls	-4.505388	1.27E-28	3.16E-28

Supplementary Table S16. Baseline characteristics of the Chinese cohort C2 consisting 47 CRC patients and 109 control individuals. For quantitative traits, the median, minimum and maximum are shown. FBG: fasting blood glucose; ALT/GPT: alanine transaminase/glutamate pyruvated transaminase; BMI: body mass index; DM: diabetes mellitus type 2; HDL: high density lipoprotein; TG: triglyceride; eGFR: epidermal growth factor receptor; TCHO: total cholesterol; Cr: creatinine; LDL; low density lipoprotein; TNM: tumor node metastasis staging system; Statistical tests used for identifying associations between metadata and CRC: † - Wilcoxon test, ‡ - Fisher's exact test.

Parameter	**Controls (n=109)**	**Cases (n=47)**	***P-value***	**q-value**
Age	58 (43,68)	69 (48,90)	3.146E-06†	1.363E-05
Gender (M:F)	40:69	25:22	0.07626‡	0.1824
BMI	23.02 (18.59,30.8)	20.94 (15.83,31.68)	0.7098†	0.7098
Stage of CRC (1:2:3:4)	n.a	4:24:15:4	n.a	n.a
Distribution of detailed TNM stages (T1N0:T3N0:T1N1:T3N1:T3N2:T4N1:T2N1M1:T3N1M1:T3N2M1:UT4:Mx)	n.a	4:23:1:9:4:1:1:1:1:1:1	n.a	n.a
Leison location (1:2:NA)	n.a	9:20:18	n.a	n.a
Leison specific location (2:3:4:6:7:8:9:NA)	n.a	3:3:3:2:7:4:7:18	n.a	n.a
Fecal sampling before or after colonoscopy	101:8 (93%:7%)	9:38 (19%:81%)	6.1669E-20‡	8.017E-19

(before:after)				
Duration between colonoscopy and fecal sample collection (days)	-63 (-202,92)	18 (-58,239)	4.064E-14†	2.642E-13
Duration of frozen storage of fecal samples (days)	374 (93,3526)	297 (30,3450)	0.2086†	0.3390
FBG	5 (4.5,6.3)	5.6 (4.5,7.9)	0.0842†	0.1824
TCHO	5.2 (3.8,5.9)	4.3 (3.6,5.3)	0.0769†	0.1824
LDL	2.9 (2,4.2)	2.5 (2.3,3.6)	0.6241†	0.6761
HDL	1.66 (1,2.03)	1.3 (0.9,2.6)	0.2822†	0.4076
TG	0.9 (0.7,2.08)	0.8 (0.5,1.9)	0.4680†	0.6084
Cr	74 (58,129)	70 (44,122)	0.5484†	0.6481
ALT/GPT	20 (14,68)	13 (10,36)	0.1043†	0.1937

Supplementary Table S17. Enrichment of two CRC-enriched and two control-enriched genes measured by qPCR in cohort C2.

Marker gene ID	**Gene description**	**Enrichment**	**Wilcoxon rank-sum test P-value**	**Wilcoxon rank-sum test stratified for colonoscopy**	**Mantel Haenszel Odds Ratio, adjusted for colonoscopy (95% CI)**	**Mantel Haenszel test P-value**

1704941	butyryl-CoA dehydrogenase	case	1.97E-09	1.52E-03	18.54 (2.62-131)	0.00509
482585	RNA-directed DNA polymerase	case	2.34E-03	4.55E-02	1.815 (0.653-5.05)	0.38
181682	gene with unknown function from Roseburia intestinalis	control	2.15E-01	3.13E-01	1.495 (0.456-4.9)	0.714
370640	gene with unknown function from Bacteroides clarus	control	3.11E-01	6.30E-01	1.647 (0.395-6.88)	0.778

Supplementary Table S18. Baseline characteristics of the Danish cohort (cohort D) consisting 16 CRC patients and 24 control individuals. For quantitative traits, the median, minimum and maximum are shown. BMI: body mass index; DM: diabetes mellitus type 2; TNM: tumor node metastasis staging system; Statistical tests used for identifying associations between metadata and CRC: † - Wilcoxon test, ‡ - Fisher's exact test.

Parameter	**Control (n=24)**	**Case (n=16)**	***P-value***	**q-value**
Age	65.5 (30, 87)	67.5 (47, 78)	0.4308219†	0.6376
Gender (M:F)	07:17	10:06	0.05309‡	0.15927
BMI	25.88 (18.94, 35.29)	25.89 (18.83, 33.20)	0.6328136†	0.6376
DM (YES:NO)	03:21	01:15	0.6376‡	0.6376
Stage of CRC (1:2:3:4)	n.a	1:9:5:1	n.a	n.a
Distribution of detailed TNM stages (T1N0M0V0:T3N0M0V0:T3N0M0V1: T3N1M0V0:T3N2M0V0:T4N0M0:	n.a	1:6:3:1:2:1:1:1	n.a	n.a

T4N2M0V1:T4NxMx)				
Cancer location (Distal:Proximal)	n.a	13:03	n.a	n.a
Cancer location (Adenocarcinom:Ascendens:Coecum:Rectum: Sigmoideum:Transversum)	n.a	1:1:1:9:3:1	n.a	n.a
Fecal sampling before or after colonoscopy (before:after)	24:0 (100%:0%)	12:4 (75%:25%)	0.0199‡	0.1194
Duration between colonoscopy and fecal sample collection (days)	7 (3, 89)	14 (-24, 252)	0.4466†	0.6376

Supplementary Table S19. Community structure differences between cohorts C1 and D. All comparisons were performed using Wilcoxon rank-sum test.

Gene count P-value					**Shannon index P-value**			
	D: Case	D: Control	C2: Case	C2: Control	D: Case	D: Control	C2: Case	C2: Control
D: Case		0.25991847	1.94E-05	0.000294527		0.772788361	5.84639E-05	4.02E-04
D: Control			7.86E-05	0.001729823			2.25586E-05	9.34E-04
C2: Case				0.212812929				0.178412749

Supplementary Table S20. Species annotation of the 1498 genes enriched in CRC patient microbiomes, both in cohort C1 and cohort D. A large fraction was

annotated to *Parvimonas micra*. Annotated species with more than 10 genes are listed here.

Species	**Gene numbers (Total=1452)**
Parvimonas micra	389
Solobacterium moorei	204
Clostridium symbiosum	177
Clostridium sp. 7_3_54FAA	108
Parvimonas sp. oral taxon 110	93
Parvimonas sp. oral taxon 393	93
Fusobacterium nucleatum	64
Peptostreptococcus stomatis	23
Clostridium hathewayi	17
Clostridium citroniae	14
Akkermansia muciniphila	11
[Clostridium] *difficile*	11
Peptostreptococcus anaerobius	10

Supplementary Table S21. List of CRC-associated species predicted from Chinese cohort C1 and validated in Danish cohort D with q<0.05

IMG species validated in cohort D					
	Control rank mean	Case rank mean	Enrichment(0:case/1: control)	P-value	q-value
Parvimonas sp. oral taxon 110	14.54166667	29.4375	0	9.06E-05	0.000808962
Parvimonas sp. oral taxon 393	14.66666667	29.25	0	0.000127394	0.000808962
Parvimonas micra	14.70833333	29.1875	0	0.00015168	0.000808962
Gemella morbillorum	15.70833333	27.6875	0	0.001465743	0.005862972
Peptostreptococcus stomatis	16.16666667	27	0	0.003409134	0.010909228
Fusobacterium sp. oral taxon 370	16.58333333	26.375	0	0.010235287	0.024739601
Fusobacterium nucleatum	16.70833333	26.1875	0	0.010823576	0.024739601
Malassezia globosa	17	25.75	0	0.023703729	0.047407459
mOTU species validated in cohort D					

	Control rank mean	Case rank mean	Enrichment(0:case/1:control)	P-value	q-value
Peptostreptococcus stomatis	16.5	26.5	0	0.000139835	0.000978842
Parvimonas micra	16.70833333	26.1875	0	0.000749378	0.002622823
Gemella morbillorum	18	24.25	0	0.004603221	0.010740848
MLG species validated in cohort D					
	Control rank mean	**Case rank mean**	**Enrichment (1:Control;0:Case)**	**P-value**	**q-value**
Parvimonas micra	15.20833333	28.4375	0	9.13E-05	0.002329351
Solobacterium moorei	16.22916667	26.90625	0	0.000172545	0.002329351

Supplementary Table S22. List of four gene markers predicted from cohort C1 that show significant associations in cohort D with q<0.05.

Gene	**Cohort C1**			**Cohort D**			**Blastn on IMG v400**	**Blastp on KEGG v59**	
Marker ID	**P-value**	**q-value**	**Enrich**	**P-value**	**q-value**	**Enrich**	**Species taxonomy**	**KEGG ID**	**Gene annotation**
2361423	2.31148E-13	4.87836E-07	case	1.16E-04	0.00116	case	*Peptostreptococcus anaerobius*	K07485	transposase

3173495	6.23501E-13	6.57946E-07	case	1.85E-04	0.00123	case	*Peptostreptococcus anaerobius*	K07485	transposase
1696299	7.69646E-10	0.000406082	case	7.87E-05	0.00116	case	*Parvimonas micra*	K03043	DNA-directed RNA polymerase subunit beta
1704941	7.53342E-08	0.002606428	case	2.08E-03	0.01040	case	*Fusobacterium nucleatum*	K00248	butyryl-CoA dehydrogenase

Supplementary Table S23. PERMANOVA analysis of variation in four gene markers validated in cohort D (No. of permutations = 9999). CRC status explains the variation in these gene profiles.

phenotype	**Df**	**Sums Of Sqs**	**Mean Sqs**	**F.Model**	**R^2**	**Pr (>F)**
CRC Status	1	8.11E-11	8.11E-11	4.8910108	0.1140335	0.0001
Stage of CRC	4	1.15E-10	2.86E-11	1.6816488	0.1612064	0.1375
Duration between colonoscopy and fecal sample collection	1	2.03E-11	2.03E-11	1.1199259	0.028628	0.2265
Cancer location (Distal:Proximal)	1	5.20E-11	5.20E-11	1.2648699	0.0828615	0.2383
Cancer location(Adenocarcinom:Ascendens:Coecum:Rectum:Sigmoideum:Transversum)	5	3.12E-10	6.24E-11	1.9756046	0.4969319	0.2998
Age	1	1.48E-11	1.48E-11	0.8097989	0.0208658	0.3989
DM	1	5.61E-12	5.61E-12	0.3020817	0.0078868	0.5654

Gender	1	6.48E-12	6.48E-12	0.3495622	0.0091152	0.571
BMI	1	7.51E-12	7.51E-12	0.4060178	0.0105717	0.5869
DNA purification date	1	3.66E-12	3.66E-12	0.1966498	0.0051484	0.6696
Fecal sampling before or after colonoscopy	1	6.95E-12	6.95E-12	0.3749813	0.0097715	0.6878
TNM	7	1.57E-10	2.25E-11	0.3823119	0.2506686	0.7061

Supplementary Table S24. Enrichment of four marker genes in published Austrian and French cohorts (A and F, respectively).

Marker Gene ID	**Cohort A**			**Cohort F**			**Blastn on IMG v400**	**Blastp on KEGG v59**	
	P-value	**q-value**	**Enrich**	**P-value**	**q-value**	**Enrich**	**Species taxonomy**	**KEGG ID**	**Gene annotation**
2361423	9.465681e-06	3.786272e-05	case	1.805948e-06	7.223791e-06	case	*Peptostreptococcus anaerobius*	K07485	transposase
3173495	1.021888e-04	3.065663e-04	case	1.311802e-05	3.935405e-05	case	*Peptostreptococcus anaerobius*	K07485	transposase
1696299	3.089198e-03	3.089198e-03	case	3.471676e-03	3.471676e-03	case	*Parvimonas micra*	K03043	DNA-directed RNA polymerase subunit beta
1704941	5.007540e-04	1.001508e-03	case	9.687230e-05	1.937446e-04	case	*Fusobacterium nucleatum*	K00248	butyryl-CoA dehydrogenase

Supplementary Table S25. Comparison of enrichment of 20 marker genes in Chinese (C1), Danish (D), Austrian (A) and French (F) cohorts. Cells marked in red: $P<0.05$. Enrichment in case or control is only reported when $P<0.2$. Only cohort C1 was used to discover gene biomarkers, and these 20 genes were among the 102,514 that associated with CRC. In cohorts D, A and F, association of only these 20 genes were verified.

Gene id	**Chinese cohort C1**		**Danish cohort D**		**Austrian cohort A**		**French cohort F**	
	Case (1) Vs. Controls (0)		Case (1) Vs. Controls (0)		Carcinoma (1) Vs Controls (0)		Case (1) Vs. Controls (0)	
	p.value	Enrichment	p.value	Enrichment	p.value	Enrichment	p.value	Enrichment
181682	1.95E-06	0	0.900619951	NA	0.678813728	NA	0.007181249	0
370640	2.64E-05	0	0.495680726	NA	0.862554181	NA	0.901689843	NA
482585	7.41E-09	1	0.467868103	NA	0.114070684	1	0.09202366	1
1559769	1.03E-05	0	0.627103852	NA	0.613815329	NA	0.318983729	NA
1696299	7.70E-10	1	7.87E-05	1	0.003089198	1	0.003471676	1
1704941	7.53E-08	1	0.002080194	1	0.000500754	1	9.68723E-05	1
1804565	2.03E-06	0	0.345063544	NA	0.719304711	NA	1	NA
2040133	7.51E-10	1	0.923193148	NA	0.037408072	1	0.3620777	NA
2206475	1.81E-06	0	0.559844892	NA	0.239405355	NA	0.086939707	0
2211919	4.98E-08	0	0.343905238	NA	0.8730299	NA	0.403859093	NA
2361423	2.31E-13	1	0.000116036	1	9.46568E-06	1	1.80595E-06	1
2736705	5.71E-06	1	0.653175645	NA	0.085244448	1	0.321243655	NA
3173495	6.24E-13	1	0.00018455	1	0.000102189	1	1.3118E-05	1

3246804	1.80E-07	1	0.586270986	NA	0.834009147	NA	0.893668207	NA
3319526	1.08E-07	0	0.646619859	NA	0.847882874	NA	0.085059441	0
3531210	1.44E-06	0	0.23124459	NA	0.014329165	1	0.142060944	0
3611706	1.68E-06	0	1	NA	0.889823764	NA	0.346149329	NA
3976414	4.42E-07	0	0.539082044	NA	0.748143815	NA	0.458758072	NA
4171064	7.50E-08	0	0.705131044	NA	0.171937649	1	0.081938362	0
4256106	7.39E-07	0	0.702861448	NA	0.05048434	1	0.880361689	NA

Supplementary Table S26. Classification accuracy of the two marker genes measured by qPCR in cohort C2, stratified into early (I-II) and late (III-IV) stage cancer.

Group	Marker ID	Enrichment	Wilcox rank-sum test, P-value	Wilcoxon rank-sum test stratified for colonoscopy, P-value	Mantel Haenszel Odds Ratio adjusted for colonoscopy (95% CI)	Mantel-Haenszel test P-value
Stages I and II	1696299	case	6.51E-14	3.35E-06	21.5 (3.18-146)	1.38E-05
	1704941	case	4.15E-07	0.008654411	27.77 (1.64-469)	0.0322
	1696299 or 1704941		N.A.	N.A.	33.37 (4.49-248)	1.68E-06
Stages III and IV	1696299	case	1.51E-11	0.00027574	15.44(3.06-77.9)	0.00109
	1704941	case	4.40E-09	0.002700628	25.34(2.91-221)	0.00842
	1696299 or 1704941		N.A.	N.A.	15.77(3.52-70.6)	0.000653

Supplementary Table S27. Primer and probe sequences for qPCR measurement of five gene markers and controls.

Gene	Sequence type	Nucleotide sequence
1696299	Forward	AAGAATGGAGAGAGTTGTTAGAGAAAGAA
	Reverse	TTGTGATAATTGTGAAGAACCGAAGA
	Probe	AACTCAAGATCCAGACCTTGCTACGCCTCA
1704941	Forward	TTGTAAGTGCTGGTAAAGGGATTG
	Reverse	CATTCCTACATAACGGTCAAGAGGTA
	Probe	AGCTTCTATTGGTTCTTCTCGTCCAGTGGC
181682	Forward	CGGATTTGCAGTGGCAAGTT
	Reverse	TGATTGCAGACGCCAATGTC
	Probe	CGTGAAAAATCCGCGCATCTGGC
370640	Forward	TCCATCCGCAAGCCTTTACT
	Reverse	GCTTCCGGTGCCATTGACTA
	Probe	TTCATCATCACAGCCGACAACGCA

482585	Forward	AATGGGAATGGAGCGGATTC
	Reverse	CCTGCACCAGCTTATCGTCAA
	Probe	AAGCCTGCGGAACCACAGTTACCAGC
control	Forward	CGTCAGCTCGTGTCGTGAG
	Reverse	CGTCGTCCCCACCTTCC
	Probe	TTAAGTCCCACAACGAGCGCAACCC