Research Paper

Estabishing high-accuracy biomarkers for colorectal cancer by comparing fecal microbiomes in patients with healthy families

Jian Yang,a Dongfang Li,b Zhenyu Yang,c,d Wenkui Dai,e,f,Xin Fengg Yanhong Li*i Yiqi Jiang,g Pingang Li*c Yinhui Li*, Bo Tang*, Qian Zhou*, Chuangzhao Qiu*, Chao Zhang*, Ximing Xu,c,d Su Feng,c,d Daxi Wang*, Heping Wang,f,h Wenjian Wang,f,h Yuejie Zheng,f,h Lin Zhang,1 Wenjie Wang,b,d Yanhong Liu,j Daxi Wang*, and Peiwu Yu2

a Department of General Surgery and Center of Minimal Invasive Gastrointestinal Surgery, The First Hospital Affiliated to Army Medical University, Chongqing, China; b Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China; c School of statistics and data science, NanKai University, Tianjin, China; d Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, Nankai University, Tianjin, China; e Department of Microbial Research, WeHealthGene Institute, Shenzhen, China; f Joint Laboratory of Microecology and Children’s Health, Shenzhen Children’s Hospital & Shenzhen WeHealthGene Co. Ltd, Shenzhen, China; g Department of Computer Science, College of Science and Engineering, City University of Hong Kong, Hong Kong, China; h Department of Respiratory Diseases, Shenzhen Children’s Hospital, Shenzhen, China; i Department of Pediatrics, The Third Hospital of Hebei Medical University, Shijiazhuang, China; j Department of Pathology, Children’s Hospital of Shanhui, Shanhui, China; k Department of Pathology, Women Health Center of Shanhui, Shanhui, China.

ABSTRACT

Colorectal cancer (CRC) causes high morbidity and mortality worldwide, and noninvasive gut microbiome (GM) biomarkers are promising for early CRC diagnosis. However, the GM varies significantly based on ethnicity, diet and living environment, suggesting varied GM biomarker performance in different regions. We performed a metagenomic association analysis on stools from 52 patients and 55 healthy family members who lived together to identify GM biomarkers for CRC in Chongqing, China. The GM of patients differed significantly from that of healthy controls. A total of 22 microbial genes were included as screening biomarkers with high accuracy in additional 46 cases and 40 randomly selected healthy adults in Chongqing (area under the receiver-operation curve (AUC) = 0.905, 95% CI 0.832–0.977). The classifier based on the 22 biomarkers also performed well in the cohort from Hong Kong (AUC = 0.811, 95% CI 0.715–0.907) and French (AUC = 0.859, 95% CI 0.773–0.944) populations. Quantitative PCR was applied for measuring three selected biomarkers in the classification of CRC patients in independent Chongqing population containing 30 cases and 30 controls, and the best biomarker from Coprobacillus performed well with high AUC (0.930, 95% CI 0.904–0.955). This study revealed increased sensitivity and applicability of our GM biomarkers compared with previous biomarkers significantly promoting the early diagnosis of CRC.

CONTACT Ke Zhou, k.zhou@hust.edu.cn; Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China; Shuaicheng Li, shuaicli@cityu.edu.hk; Department of Computer Science, College of Science and Engineering, City University of Hong Kong, Hong Kong, China; Peiwu Yu, yupeiwu01@sina.com; Department of General Surgery, Southwest Hospital, Third Military Medical University, Chongqing, China.

ARTICLE HISTORY

Received 20 August 2019
Revised 22 November 2019
Accepted 23 December 2019

KEYWORDS

Colorectal cancer; Gut microbiome; Biomarkers; Family cohort; qPCR validation

Introduction

Colorectal cancer (CRC) is the third most frequently diagnosed cancer in men and the second in women.1,2 In China, the number of new CRC patients was estimated to be 376,300, and the CRC death toll was 191,000 in 2015.3 The American Cancer Society recommends a colonoscopy every 10 years after age 45 for the early diagnosis of colorectal cancers, especially CRC.4 However, colonoscopy is not widely acceptable in common health investigations given its invasive nature. Noninvasive tests, such as liquid biopsy and fecal immunochemical test, are promising for CRC screening but have limited detection precision and prediction accuracy.5

Emerging reports have demonstrated dysbiosis of gut microbiome (GM) in CRC patients and the potential of GM biomarkers in CRC screening.6,7 The imbalanced GM included Fusobacterium nucleatum, Peptostreptococcus stomatis, Parvimonas micra, and Solobacterium moorei, which have been validated as biomarkers for CRC screening in several studies.8–11

© 2020 Taylor & Francis Group, LLC
Microbial genes from the GM have also been increasingly accepted as early diagnostic CRC biomarkers.\textsuperscript{5,12} However, dietary habits and exterior environments can impose long-term effects on GM configuration.\textsuperscript{13} Thus, we hypothesized that GM biomarkers in CRC screening differ based on region given the decreasing accuracy observed when biomarkers from Hong Kong (HK) population were applied to French and Austrian populations.\textsuperscript{8} Additionally, we hypothesized that high-accuracy CRC screening could be achieved using GM biomarkers that were identified by comparing the fecal microbiome of CRC patients with that of healthy family members.

To identify and investigate GM biomarkers for CRC screening in Chongqing (CQ), China, and test our hypothesis, we collected 107 fecal samples from 52 CRC patients and 55 corresponding healthy adult family members who lived together for at least 1 year. Metagenomic analysis was conducted to explore GM disequilibrium in CRC patients and associated biomarkers. We also assessed the performance of our GM biomarkers in additional 86 fecal samples in CQ and 269 microbial samples in different populations. Additionally, three biomarkers were validated with quantitative PCR (qPCR) in an independent Chongqing population with randomly selected 30 patients and 30 healthy controls.

### Results

#### Sample information

Fecal samples from 193 Chinese subjects, including 98 CRC patients (68 males and 30 females) and 95 healthy controls (49 males and 46 females), were collected at the Southwest Hospital in CQ (Table 1). Of the recruited subjects, 52 cases and their 55 healthy family members were included to identify GM biomarkers (collectively regarded as Family cohort, Table S1). The remaining 86 fecal samples were selected randomly as the validation cohort (Table S1). Additionally, independent 30 CRC patients and 30 healthy controls were recruited from the same hospital for qPCR assessment (Table S2). We also collected published metagenomics data from 75 CRC patients and 53 healthy controls in HK (PRJEB10878) and 53 CRC patients and 88 healthy individuals in France (ERP005534).\textsuperscript{8,14}

#### GM alterations in CRC patients compared with healthy family members

Metagenomic sequencing was performed for 52 CRC patients and 55 corresponding healthy family members who were recruited based on various inclusion criteria (Table 1, Table S1). We

| Table 1. Sample information. |
|-------------------------------|
| **Family cohort** | **Validation cohort** |  |
| **CRC (n = 52)** | **Controls (n = 55)** | **Adjusted P** | **CRC (n = 46)** | **Controls (n = 40)** | **Adjusted P** | **Overall (n = 193)** |
| **Gender (male:female)** | 35:17 | 26:29 | 0.08 | 33:13 | 23:17 | 0.23 | 117:76 |
| **Ages (years)** | 53 (33, 74) | 42 (20, 72) | < 0.001 | 59 (28, 76) | 42.5 (22, 65) | < 0.001 | 48 (20, 76) |
| **BMI (kg/m\textsuperscript{2})** | 23.49 (17.51, 28.73) | 23.21 (19.87, 27.6) | 0.66 | 22.5 (17.92, 30.43) | 22.82 (19, 28.27) | 0.53 | 23.06 (17.51, 30.43) |
| **GGT (IU/L)** | 21 (8, 71) | 20 (11, 38) | 0.55 | 25 (11, 166) | 29.5 (11, 55) | 0.56 | 22 (8, 166) |
| **AST (IU/L)** | 20.6 (7.9, 47.9) | 15.8 (9.2, 22.4) | < 0.001 | 20.8 (10, 105.8) | 17.4 (9.3, 23) | 0.03 | 17.95 (7.9, 105.8) |
| **ALT (IU/L)** | 17.1 (3.7, 69.4) | 15.5 (9, 22.9) | 0.06 | 15.2 (1.1, 98.5) | 15.5 (9, 23) | 0.51 | 15.6 (1.1, 98.5) |
| **ALB (g/L)** | 40.4 (27.8, 52.4) | 41.9 (37.1, 49.9) | 0.16 | 41.85 (31.1, 51.5) | 42.4 (37, 49.8) | 0.23 | 41.85 (27.8, 52.4) |
| **Fasting glucose (mmol/L)** | 5.49 (4.16, 8.12) | 4.65 (3.93, 6.08) | < 0.001 | 5.51 (3.9, 13.32) | 5.31 (4.02, 6.06) | 0.06 | 5.29 (3.9, 13.32) |
| **HGB (g/L)** | 127 (65, 177) | 135 (108, 160) | 0.09 | 132 (75, 159) | 142 (108, 160) | 0.03 | 133 (65, 177) |
| **Cr (mmol/L)** | 71 (40.5, 115) | 79.7 (59.1, 102.5) | 0.04 | 71.2 (33, 102.1) | 80.9 (59.1, 102.8) | 0.005 | 74.65 (33, 115) |
| **CEA (ng/mL)** | 3.28 (0.32, 178.36) | - | - | 2.84 (0.68, 397.6) | - | - | 2.96 (0.32, 397.9) |
| **AJCC stage (I:II:III:IV)** | 13:18:16:5 | - | - | 10:18:15:3 | - | - | 23:36:31:8 |
| **FOBT (positive:negative)** | 46:6 | 2:53 | < 0.001 | 31:15 | 02:38 | < 0.001 | 81:112 |
| **Localization (rectum:colon)** | 29:23 | - | - | 29:17 | - | - | 58:40 |

*"* represents no detection result; BMI, body mass index; GGT, γ-glutamyl transpeptidase; AST, aspartate aminotransferase; ALT, serum alanine aminotransferase; ALB, albumin; HGB, hemoglobin; Cr, creatinine; CEA, carcinoembryonic antigen; AJCC, American Joint Committee on Cancer staging system; FOBT, fecal occult blood test.
produced a total of 7.25 billion high-quality sequencing reads (37.56 million reads per individual on average, Table S3) using the Illumina HiSeq platform (Illumina, San Diego, California, USA). The rarefaction curve exhibited a plateau in all samples, suggesting a sufficient sequencing depth for the following analysis (Figure S1).

Age, aspartate aminotransferase, creatinine, fasting glucose and fecal occult blood test had significant differences between CRC patients and controls (Table 1). Permutational multivariate analysis of variance (PERMANOVA) indicated that diagnosis status ($P = .015$, Table S4) and age ($P = .04$, Table S4) contributed significantly to the differences of GM between patients and controls. Gender took a slight effect on the discrepancy of GM ($P = .072$, Table S4), while other clinical indicators, including body mass index, aspartate aminotransferase, fasting glucose, creatinine, etc, associated insignificantly with GM ($P > .05$, Table S4). We adjusted the confounding effect from age and gender with a generalized additive model. Microbial diversity slightly increased in the GM of CRC patients compared with that of healthy individuals ($P > .05$, Figure 1a). The phylum Bacteroides and Firmicutes dominated the GM of both healthy and diseased individuals, but the difference of them was not observed between healthy and patients (adjusted $P > .05$, Figure 1b). In addition, Proteobacteria was slightly enriched without significant difference in the GM of CRC patients (adjusted $P > .05$, Figure 1b). In addition, the difference in gene number was not obvious between two groups ($P > .05$, Figure 1c). Compared with that of healthy family members, the genera in CRC patients distinctly enriched by Coprobacillus, Burkholderia, Porphyromonas, Paracoccus, Peptoniphilus, Synechococcus and Cyanothece (adjusted $P < .05$, Figure 1d, Table S5). We also identified several accumulated microbial species in the GM of patients including Roseburia inulinivorans, Clostridium ramosum, Porphyromonas gingivalis, F. nucleatum, Gemella morbillorum, etc. (adjusted $P < .05$, Figure 1d, Table S5). We also observed that the correlations of the microbial species were obviously different in CRC patients and healthy people (Figure S2).

On average, 624,404 microbial genes were observed in the GM of healthy controls, whereas only 585,092 genes were found in CRC cases (Figure 1c). In addition, a decreased proportion of functional category named amino acid metabolism was observed in the GM of inpatients (Figure 1e).

**Discovery of CRC biomarkers in CQ family cohort**

By comparing the GM of CRC patients with that of healthy individuals, we identified 22 microbial genes that were strongly associated with CRC (Table S6). Of the 22 genes, 20 enriched in the GM of CRC cases and 2 enriched in the GM of healthy family members (Figure 2a). Almost half of the biomarker genes ($N = 8$) were assigned as from phylum Firmicutes. Four biomarker genes were classified into Fusobacteria and 10 biomarker genes remained unknown. In eight genes from Firmicutes, five genes were classified as belonging to two microbial species, including *Clostridium symbiosum* ($N = 4$) and *S. moorei* ($N = 1$). Of the remaining three genes, one was assigned to the genus *Coprobacillus* and the other two were unknown. In four genes from phylum Fusobacteria, one gene was classified into *F. nucleatum* and the rest three genes were identified as genus *Fusobacterium* (Table S6). Of the 12 candidate genes with known functional annotation, four CRC-associated genes were annotated to quorum sensing ($N = 2$), butanoate metabolism ($N = 1$) and microbial metabolism in diverse environments ($N = 1$). The other eight candidate genes were annotated to acting on a sulfur group of donors, ribosome, genetic information processing, type II secretion system, porphyrin and chlorophyll metabolism, replication and repair, acting on peptide bonds (peptidases) and arginine biosynthesis (Table S6).

The receiver operating characteristic (ROC) curve indicated that the 22 candidate genes significantly differentiate CRC cases and healthy families (area under the ROC (AUC) = 0.998, 95% CI 0.993–1.000). When applied to 46 additional CRC cases and 40 randomly selected healthy controls (validation cohort, Table S1), the 22 candidate genes also exhibited high sensitivity for CRC screening (Figure 2b, AUC = 0.905, 95% CI 0.832–0.977). In parallel, these 22 biomarker genes were similarly distributed in additional 86 CQ individuals as in the family cohort (Figure 2a).
Sensitivity of GM biomarkers for screening CRC revealed significant regional tendency

To test the performance of 22 biomarker genes in screening CRC in other regions, we collected published metagenomics data from HK (74 CRC cases and 54 healthy adults) and French (53 cases and 88 healthy adults) populations to serve as an external validation cohort (Table S7). Similar to CQ populations, 20 of 22 biomarker genes tended to accumulate in CRC patients in the HK and French cohorts while the remaining two genes enriched in healthy controls (Figure 2a). The ROC curve showed a high sensitivity of biomarker genes for classifying CRC patients in HK (AUC = 0.811, 95% CI 0.715–0.907) and French (AUC = 0.859, 95% CI 0.773–0.944) populations (Figure 2b), while these biomarker genes from...
CQ cohort seemed to exhibit reduced accuracy when applied to CRC screening in population with different living regions and diet habits (Figure 2b).

**Living region was more significant for differentiating microbial samples compared to CRC risk**

To understand the reduced accuracy of CQ GM biomarkers to detect CRC in other region populations, we analyzed the distribution of all microbial samples in three cohorts. Non-metric multidimensional scaling analysis indicated that microbial samples in the same region were clustered together (Figure 3a). In addition, microbial samples in CQ were more similar to those in HK than in French populations (Figure 3a).

Further analyses indicated the lowest GM diversity in the CQ cohort and the highest in the French cohort (Figure S3a, adjusted $P < .05$). *Bacteroides* dominated GM in three cohorts, whereas *Prevotella* represented the second highest abundance in the GM of the CQ cohort (Figure S4, Table S8). In addition, compared with the other two cohorts, these two dominant genera enriched higher in CQ cohort. In contrast, *Escherichia* and *Ruminococcus*

![Figure 2](image2.png)

**Figure 2.** Enrichment and performance of 22 biomarker genes among three cohorts. (a) Candidate genes are shown in the left vertical column, and corresponding phylogenetic and functional assignments are shown in the right column. The relative abundance is represented by different colors ($\log_{10}$ (relative abundance); white, not detected; red, the most abundant). The red and green gene names indicate CRC enriched and health enriched genes, respectively. (b) The red curve indicates ROC in the validation cohort; the green and aquamarine curves indicate ROC in the HK and French cohorts, respectively.

![Figure 3](image3.png)

**Figure 3.** Cluster of microbial samples from differed populations and distribution of biomarker bacteria in three cohorts. (a) Dot and triangle indicate microbiomes from CRC cases and healthy controls, respectively. Blue, green and red colors represent microbial samples in CQ, HK and French cohorts, respectively. (b) Relative abundance of the four biomarker microbes in three cohorts. The significant difference of abundance within the individual cohort is labeled with * (*, $P < .05$; **, $P < .01$; ***, $P < .001$).
represented the second most abundant microbial genera in the GM of the HK and French cohorts, respectively (Figure S4, Table S8). However, *Escherichia* had the lowest relative abundance in CQ cohort and was different among the three cohorts. The relative abundance of *Ruminococcus* was highest in French cohort and was different among the three cohorts. At the species level, *Bacteroides dorei*, *vulgatus*, *Escherichia coli* and *Bacteroides uniformis* were dominant in the GM of the CQ, HK and French populations, respectively (Figure S4, Table S8). *Prevotella copri* differed among three cohorts. Compared to the discrepancy in GM among different regions (Table S8), structural variations in GM between diseased and healthy individuals of the same region reduced slightly (Figure S4). For instance, the top five microbial genera in the GM of both inpatients and healthy controls in CQ included *Bacteroides*, *Prevotella*, *Faecalibacterium*, *Eubacterium* and *Escherichia* (Figure S4a, Table S8). Regarding the HK and French cohorts, the top 10 GM genera between cases and healthy individuals were similar (Figure S4b, c, Table S8).

Additionally, abundance of biomarker gene-related microbes were different in three cohorts. Relative abundance of *Coprobacillus* differed between diseased and healthy people only in CQ cohort, and that of *F. nucleatum* differed between CRC patients and healthy controls in CQ, HK and French cohorts. The distribution of *S. moorei* showed a discrepancy between diseased and healthy subjects in French cohort. The relative abundance of *C. symbiosum* revealed a distinct difference between patients and healthy people both in HK and in French cohorts (Figure 3).

**Three biomarker genes were verified by qPCR in independent Chongqing population**

According to the degree of gene’s contribution to the accuracy of classification CRC patients, we selected three genes from 22 biomarkers for further verification (gene 8122329, unknown function from *Coprobacillus*; gene 3742340, nitrilase from *C. symbiosum*; gene 5053929, peptide methionine sulf-oxide reductase msrA/msrB from *Fusobacterium*). Their relative abundances were measured by qPCR in an independent Chongqing population (30 cases and 30 controls, age and gender were not significant different between patients and controls groups, Table S2). The results showed that the relative abundance of two biomarker genes (from *Coprobacillus* and *C. symbiosum*) differed significantly between CRC patients and healthy controls (adjusted *P* < .05, Figure 4a). But the biomarker gene from *Fusobacterium* did not show a significant difference (adjusted *P* > .05, Figure 4a). Density curve illustrated discrepant distribution of two biomarker genes’ (from *Coprobacillus* and *C. symbiosum*) relative abundance between CRC cases and healthy controls (Figure 4b). However, the biomarker gene from *Fusobacterium* was not regarded as an appropriate independent CRC indicator and was excluded in the following analyses. Based on random forest classification algorithm, ROC curve indicated that the selected biomarker gene from *Coprobacillus* significantly classified CRC patients from healthy controls in 60 independent Chongqing population (Figure 4c, AUC = 0.930, 95% CI 0.904–0.955). In addition, the AUC value increased slightly after adding genes from *C. symbiosum* into classification model (Figure 4d, AUC = 0.935, 95% CI 0.883–0.987).

**Discussion**

CRC is one of the most malignant tumors worldwide and early screening can help reduce associated mortality. Currently available methods for early CRC diagnosis include fecal occult blood test, colonoscopy examination, fecal immunochemical test and carcino-embryonic antigen test. Such tests have many limitations, including low accuracy and invasive techniques.

Emerging studies have demonstrated imbalanced GM in CRC patients and the promise of GM biomarkers in CRC screening. Prior studies documented that the low-abundance microbes in GM of CRC patients contributed to CRC development, such as *F. nucleatum, P. stomatis, P. micra* and *S. moorei*, most of which were recognized as potential GM biomarkers. Previous reports suggested the potential role of our identified low-abundance microbial species and genera, including *F. nucleatum, P. gingivalis, G. morbillorum, P. nigrescens*, *Porphyromonas* and *Coprobacillus*, in CRC development. Recent
study found that *F. nucleatum* was able to promote intestinal tumorigenesis via adhering to cancer cells, modulating immune cells and modifying the tumor microenvironment.\(^{24-26}\) In addition, *F. nucleatum* also stimulated Toll-like receptor 4 to modulate CRC patient’s chemotherapeutic response, which possibly influenced the treatment outcome.\(^{27}\) The other low-abundant microbes, such as *P. gingivalis*, *P. nigrescens*, *G. morbillorum*, and *porphyromonas*, were also considered to be strongly associated with CRC development.\(^{22,23,28}\) Additionally, further study should be conducted to analyze *Coprobacillus*, being rarely reported but identified as high-accuracy biomarker in CQ cohort.

Emerging studies indicated that GM-related metabolism, such as synthesis, deamination and decarboxylation, affected cancerous conditions in the gut.\(^{29,30}\) Previous researches identified increased abundance of microbial genes related to amino acid metabolism in GM of CRC patients,\(^{31}\) which was consistent with our results. The association between microbial metabolism of amino acids and CRC development may be established by producing toxic metabolites. For example, GM was able to degrade phenylalanine, being an aromatic amino acid, into CRC-associated toxic phenylacetate through catabolic pathways.\(^{32,33}\) Additionally, GM could metabolize sulfur-containing amino acid, including cysteine and methionine, via sulfide-producing pathway and generate toxic H\(_2\)S which contributed to CRC incidence.\(^{34}\)

Given that GM could be shaped by various external factors, including living environments, diet, lifestyle and antibiotics exposure,\(^{35-37}\) which should be considered when applying GM biomarkers in screening CRC. Our study selected CRC patients and their healthy families to decrease the confounding effects of living environment and lifestyle on GM differences. After adjusting the confounding effect of the age in the generalized additive model, identified GM biomarkers showed higher accuracy in screening CRC compared to published biomarkers.\(^{8,14}\) Two biomarker genes from *C. symbiosum* and *Coprobacillus* were verified with high accuracy in Chongqing population via qPCR. *C. symbiosum* was confirmed to be related to CRC in previous studies and reported as a GM
biomarker for detecting early and advanced CRC patients in Shanghai population in China.\textsuperscript{14,38,39} In addition, another biomarker gene from genus \textit{Coprobacillus} was enriched in GM of ethanol-related CRC patients,\textsuperscript{40} which coincided with the fact that alcoholic beverages were widely accepted in Chongqing.\textsuperscript{41} This may partly explain that \textit{Coprobacillus} performed well when screening CRC in Chongqing population. Recent studies also showed the improved accuracy of GM biomarkers when diminishing the confounding effect of environmental factors from various population cohorts.\textsuperscript{22,23} However, these cross-cohort studies might neglect population-specific microbes or DNA fragments, such as \textit{Coprobacillus} which was identified in our study and validated by qPCR in an independent cohort with high accuracy (AUC = 0.930).

Several limitations should be noted in our study. The family paired sampling design in our study is unable to totally eliminate effects from external factors, attributing to the difference of dietary, activity area and mental state in each family member. Therefore, the external factor contributing to confounding effect is supposed to be considered and adjusted in the following study, such as the age in our study. In addition, the accuracy of biomarkers identified from family-paired cohort is also limited by external factors in different populations. The recent cross-cohort studies give us a new angle to face this limitation: cross-cohort and family paired sampling method could be integrated for identifying higher accuracy and wilder applicable range biomarkers under appropriate inclusion criteria, sufficient clinical information and minimum sample size. Moreover, updated integrated gene catalog (IGC) database was employed for taxonomic annotation, which is the largest free accessible non-redundant gene catalog with 11,446,577 genes of human GM.\textsuperscript{42} But the limitations of this database are lacked verification from cultivation experiment and are not timely updated taxonomic and functional annotation.\textsuperscript{43} In 2019, the culturable genome reference (CGR) was established by 1,520 cultivated and assembled bacteria genomes.\textsuperscript{44} Although some strictly anaerobic bacteria might be inevitably omitted through cultivating, CGR database fills the limitation of IGC through cultivation experiments. We hope to revise our taxonomic annotation when CGR is available. CGR and IGC databases have a slight different scope of application: the CGR database might be fit to the study focused on specific species in GM, and the IGC database is more appropriate to the study of whole GM, such as identifying biomarkers from a dataset.

In general, GM from different population displayed respective characteristic because of a greatly confounding effect from regional environmental factors. However, regional specific biomarkers could be identified based on the GM and displayed high accuracy in a particular population. Therefore, regional specific GM biomarkers would be an efficient clinical tool for CRC diagnosis. Further analysis is required to promote the clinical application of GM biomarkers, including validating our findings in a large cohort study and exploring the diagnosing index integrating GM biomarkers and clinical indices. It is foreseen that, with the GM research developments, the regional specific GM biomarker will promisingly become conventional diagnosis method for CRC screening.

\section*{Materials and methods}

\subsection*{Subject inclusion and sample collection}

CRC patients were diagnosed histopathologically by colonoscopy at the hospital.\textsuperscript{45} The exclusion criteria for patients were as follows: ≤18 years old or ≥76 years old, colorectal benign lesion, antibiotics exposure, radiotherapy therapy or continuous treatment by systemic corticosteroids 1 month prior to sampling and serious mental disorder. Healthy individuals were selected based on the following exclusion criteria: dysentery, chronic enteritis, inflammatory bowel disease, irritable bowel syndrome, Crohn’s disease, metabolic diseases (BMI ≥ 32, diabetes or malnutrition), long-term probiotics uptake, continuous treatment by systemic corticosteroids 1 month prior to sampling and serious mental disease. Selected cases’ healthy family members lived together with the patients for at least 1 year. The ethics committee of The First Hospital Affiliated to Army Medical University approved this study (Registration Number: KY201737). Informed consent were
obtained from recruited patients and healthy individuals.

Fresh stool, blood and urine of each patient were collected in the early morning after admission to the hospital. All collected samples were frozen on dry ice within 30 min and then stored at $-80^\circ C$ until further analysis. Fresh stools from healthy subjects were collected during physical examination in the hospital.

**Laboratory assessment**

A blood auto-analyzer (Beckman Coulter AU5800, Brea, CA, USA) was used to analyze serum creatinine, fasting plasma glucose, serum alanine aminotransferase, aspartate aminotransferase, $\gamma$-glutamyl transpeptidase and albumin. Tumor markers, such as carcinoembryonic antigen, were tested in a detection system (Lumix200 xMAP, Austin, TX, USA). Fresh stools were processed on an automatic processing platform for standard stool examination, including fecal occult blood test, stool color and shape (JinHua JHSPSY-I, Nanchang, JiangXi, China).

**DNA extraction, library construction and sequencing**

DNA extraction was performed using the QIAamp DNA Stool Mini Kit (QIAGEN, 19593). DNA concentration was measured using a NanoDrop spectrophotometer (Thermo Scientific, Scoresby, Victoria, Australia) and Qubit Fluorometer (Life Technologies, Grand Island, NY, USA). Molecular size was evaluated by agarose gel electrophoresis. All DNA samples were stored at $-20^\circ C$ before further process. DNA libraries were constructed following the manufacturer’s instruction (Illumina, San Diego, CA, USA) with an insert size of 350 bp. Paired-end sequencing was performed on the Illumina HiSeq platform (San Diego, CA, USA). Sequencing data were accessible in the GenBank database under accession number: SRP128485.

**Phylogenetic and functional profiling**

Paired-end reads were filtered using the following criteria: one of the paired reads contained 10% ambiguous Ns or contained 50% bases with a quality score <20. The filtered paired-end reads were mapped to the human genome (hg19) to eliminate genomic contamination with SOAPaligner 2.21 (-m 244 -x 455 -v 5 -r 1 -l 35 -M 4). The remaining high-quality reads were aligned with an updated IGC using bowtie2 (v 2.3.0) (very-sensitive-local – score-min L0,1,6 -k 10 -p 8 -I 200 -X 500 – reorder – omit-sec-seq -N 1 – ignore-quals). Three types of alignment results could be observed when paired-end reads aligned to genes in the database: (1) paired-end reads aligned to one gene, which was counted as two reads mapping; (2) paired-end reads aligned to more than one gene, the primary aligned gene was counted as two-read mapping, and the remaining genes were counted zero; or (3) one of paired-end reads aligned to one gene, which was counted as one-read mapping.

Then, the number of mapped paired-end reads to each gene in IGC was counted. For a given gene g, its relative abundance is $\text{Ab}(g)_{\text{relative}}$, which was calculated using the following formula:

$$\text{Ab}(g)_{\text{relative}} = \text{Ab}(g) \times 100/\text{Ab}(G)$$

$\text{Ab}(g)$ is the abundance of gene g, $\text{Ab}(G)$ is the sum of all genes abundance, U is the number of mapped reads in gene g and L is the length of gene g.$^49$

The relative abundance of taxonomy (Phylum, genus and species) and KEGG orthology was calculated by summing $\text{Ab}(g)$ of corresponding genes based on the annotated results in the IGC database.

**Statistical analysis**

Based on taxonomy profiling, alpha diversity was assessed by Shannon index, and the rarefaction curve was generated with vegan package (version: 2.4.4; R: 3.3.1). The confounding effect was tested by PERMANOVA. Generalized additive model was applied for removing the confounding effect from age and gender.$^51$ ROC curves were generated with pROC in R (pROC: 1.10.0, R: 3.3.1). Density curve was drawn by R software (version: 3.3.1). Wilcoxon rank-sum test was applied to explore inter-group differences and the significance ($P$-value) was adjusted using the Benjamini and Hochberg method.$^52$
Biomarkers identification

Genes absent in more than 80% of the samples were removed. Generalized additive mode was employed for identifying different microbial genes between diseased and healthy subjects ($P < .01$). A two-step schema was then applied for biomarker identification. Firstly, we performed feature selection on these retained genes which were filtered using the mRMR algorithm and the top 100 ones were enrolled as candidates for further analysis. In the second stage, random forest model was used to build the classifier based on the top k-th genes by varying k in the set 1 to 100. The biomarkers were identified based on the performance of the classifiers which was assessed by 5-fold cross-validation and ROC.  

Quantitative PCR

Primers for three microbial biomarker genes and bacterial universal 16S rDNA gene were synthesized and purified by Sangon Biotech (Shanghai, China). The primer sequences were presented as follows: 8122329 gene For-CTCTGTAAGGAGAAGTTC AATTCTTG; 8122329 gene Rev-CTCTCCTTTCTTT GGAITCTGCAAGTG; 3742340 gene For-TTATAG TGGGGACGGATAATGCG; 3742340 gene Rev-GG TATTTTATGTATTGGCCGCC; 5053929 gene For-TGGAAAGATGGAAAAACCAACTTGTGTT; 5053929 gene Rev- CACGACATTTAATATCCTT GATAATTCACCTT. Internal reference primer for total bacterial DNA was determined by 16s rRNA using the following primers: 16s rRNA For-GTTGTGCAGCTCGTGTGC; 16s rRNA Rev- GCAGTCTCGCTAGTGCC. qPCR amplifications were performed in a 20-μl reaction system containing 10 μl of SYBR Green premix Dimer Eraser (Takara Bio, RR091A), 2 μl of extracted fecal DNA and 1.2 μl of primers. Amplification and detection of DNA were performed with the C1000 Thermal cycler (Bio-Rad, Hercules, CA, USA) via the following reaction conditions: 95°C for 3 min, followed by 40 cycles of 95°C for 5 s and 60°C for 30 s. Each sample was assayed in triplicate and the mean of the three cycle threshold (Ct) values was used for subsequent analysis. The Ct value is defined by the number of cycles for the fluorescent signal reaching the threshold in qPCR. The abundance of the biomarkers in each sample was calculated by $2^{-\Delta \Delta Ct}$ methods ($\Delta Ct = Ct_{target} - Ct_{control}$, $\Delta \Delta Ct = \Delta Ct - \Delta Ct_{maximum}$).

Acknowledgments

We thank all the nurses who helped with physical examination, clinical recording and fecal collection at The First Hospital Affiliated to Army Medical University. We also thank the authors who made their data publicly available. We also thank Mr. Xiaofeng Lin from EasyPub for improved the language.

Disclosure of potential conflict of interests

The authors declare that they have no competing interests.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65:87–108. PMID: 25651787. doi: 10.3322/caac.21262.
2. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A, Jemal A. Colorectal cancer statistics, 2017. CA Cancer J Clin. 2017;67:177–193. PMID: 28248415. doi: 10.3322/caac.21395.
3. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ, He J. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66:115–132. PMID: 26808342. doi: 10.3322/caac.21338.
4. Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, Etzioni R, McKenna MT, Oeffinger KC, Shih YT, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American cancer society. CA Cancer J Clin. 2018;29846947. doi:10.3322/caac.21457.
5. Pantel K, Alix-Panabieres C. Liquid biopsy in 2016: circulating tumour cells and cell-free DNA in gastrointestinal cancer. Nat Rev Gastroenterol Hepatol. 2017;14:73–74. PMID: 28096542. doi: 10.1038/nrgastro.2016.198.
6. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol. 2014;12:661–672. PMID: 25198138. doi: 10.1038/nrmicro3344.
7. Sears CL, Garrett WS. Microbes, microbiota, and colon cancer. Cell Host Microbe. 2014;15:317–328. PMID: 24629338. doi: 10.1016/j.chom.2014.02.007.
8. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, Tang L, Zhao H, Stenvang J, Li Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. 2015;66:70–78. PMID: 26408641. doi:10.1136/gutjn-2015-309800.
9. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma
Fusobacterium nucleatum promotes colorectal carcinoma by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. Cell Host Microbiol. 2013;14:195–206. PMID: 23954158. doi:10.1016/j.chom.2013.07.012.

Pedersen RM, Holt HM, Justesen US. Solobacterium moorei bacteremia: identification, antimicrobial susceptibility, and clinical characteristics. J Clin Microbiol. 2011;49:2766–2768. PMID: 21525228. doi:10.1128/JCM.02525-10.

Nistal E, Fernandez-Fernandez N, Vivas S, Olcoz JL. Factors determining colorectal cancer: the role of the intestinal microbiota. Front Oncol. 2015;5:220. doi:10.3389/fonc.2015.00220.

Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Bohm J, Brunetti F, Habermann N, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10:766. doi:10.1038/msb.20145645.

Schoen RE, Pinsky PF, Weissfeld JL, Yokochi LA, Church T, Laiyemo AO, Bresalier R, Andriole GL, Churchman CD, Moore N, et al. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome. 2018;6:70. PMID: 29642940. doi:10.1186/s40168-018-0451-2.

Wong SH, Kwong TNY, Wu CY, Yu J. Clinical applications of gut microbiota in cancer biology. Semin Cancer Biol. 2018. PMID: 29782923. doi:10.1016/j.semcancer.2018.05.003.

Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, et al. Meta-analysis of fecal metabolomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25:679–689. PMID: 30936547. doi:10.1038/s41591-019-0406-6.

Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zollo M, Beghini F, Manara S, Karcher N, Pozzi C, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019;25:667–678. PMID: 30936548. doi:10.1038/s41591-019-0405-7.

Abed J, Emdarg JE, Zamir G, Faroja M, Almogy G, Grenov A, Sol A, Naor R, Pikarsky E, Atlan KA, et al. Fap2 mediates fusobacterium nucleatum colorectal adenocarcinoma enrichment by binding to tumor-expressed Gal-GalNAc. Cell Host Microbe. 2016;20:215–225. PMID: 27512904. doi:10.1016/j.chom.2016.07.006.

Gur C, Ibrahim Y, Isaacson B, Yamin R, Abed J, Gamliel M, Enk J, Bar-On Y, Stanietskyy-Kaynan N, Coppenhagen-Glazer S, et al. Binding of the Fap2 protein of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. Immunity. 2015;42:344–355. PMID: 25680274. doi:10.1016/j.immuni.2015.01.010.

Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, Clancy TE, Chung DC, Lochhead P, Hold GL, et al. Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe. 2014;14:207–215. PMID: 23954159. doi:10.1016/j.chom.2013.07.007.

Yu T, Guo F, Yu Y, Sun T, Ma D, Han J, Qian Y, Kryczek I, Sun D, Nagarsheth N, et al. Fusobacterium nucleatum promotes chemoresistance to colorectal cancer by modulating autophagy. Cell. 2017;170:548–63e16. PMID: 28753429. doi:10.1016/j.cell.2017.07.008.

Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, Goedert JJ, Shi J, Bork P, et al. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. PLoS One. 2016;11:e0155362. PMID: 27171425. doi:10.1371/journal. pone.0155362.

Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol. 2014;12:661–672. PMID: 25198138. doi:10.1038/nrmicro3344.

Neis EP, Dejong CH, Rensen SS. The role of microbial amino acid metabolism in host metabolism. Nutrients.
31. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25:968–976. PMID: 31171880. doi:10.1038/s41591-019-0458-7.

32. Russell WR, Duncan SC, Scobie L, Duncan G, Cantlay L, Calder AG, Anderson SE, Flint HJ. Major phenylpropanoid-derived metabolites in the human gut can arise from microbial fermentation of protein. Mol Nutr Food Res. 2013;57:523–535. PMID: 23349065. doi:10.1002/mnfr.201200594.

33. Windey K, De Preter V, Verbeke K. Relevance of protein fermentation to gut health. Mol Nutr Food Res. 2012;56:184–196. PMID: 22121108. doi:10.1002/mnfr.201100542.

34. Ma N, Tian Y, Wu Y, Ma X. Contributions of the interaction between dietary protein and gut microbiota to intestinal health. Curr Protein Pept Sci. 2017;18:795–808. PMID: 28215168. doi:10.2174/1389203718666170216153505.

35. Makk K, Deegan EC, Walter J, Backhed F. The impact of dietary fiber on gut microbiota in host health and disease. Cell Host Microbe. 2018;23:705–715. PMID: 29902436. doi:10.1016/j.chom.2018.05.012.

36. Conlon MA, Bird AR. The impact of diet and lifestyle on gut microbiota and human health. Nutrients. 2014;7:17–44. PMID: 25545101. doi:10.3390/nu7010017.

37. Marchesi JR, Adams DH, Fava F, Hermes GD, Hirschfeld GM, Hold G, Quraishi MN, Kinross J, Smidt H, Tuohy KM, et al. The gut microbiota and host health: a new clinical frontier. Gut. 2016;65:330–339. PMID: 26338727. doi:10.1136/gutjnl-2015-309990.

38. Xie YH, Gao QY, Cai GX, Sun XM, Zou TH, Chen HM, Yu SY, Qiu YW, Gu WQ, Chen XY, et al. Fecal Clostridium symbiosum for Noninvasive Detection of Early and Advanced Colorectal Cancer: test and Validation Studies. EBioMedicine. 2017;25:32–40. PMID: 29033369. doi:10.1016/j.ebiom.2017.10.005.

39. Ai L, Tian H, Chen Z, Chen H, Xu J, Fang JY. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. Oncotarget. 2017;8:9546–9556. PMID: 28061434. doi:10.18632/oncotarget.14488.

40. Tsuuryu A, Kuwahara A, Saito Y, Yamaguchi H, Tsubo T, Suga S, Imai M, Aoki Y, Takahashi S, Tsutsui M, et al. Ecophysiologial consequences of alcoholism on human gut microbiota: implications for ethanol-related pathogenesis of colon cancer. Sci Rep. 2016;6:27923. PMID: 27295340. doi:10.1038/srep27923.

41. Healthy drinking habits survey in 25 provinces over China. Beijing (China): China Health Care Association; 2007. p. 2008.

42. Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, Ward KJ, Jackson MA, Xia Y, Chen X, et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. Cell Syst. 2016;3:572–584e3. PMID: 27818083. doi:10.1016/j.cels.2016.10.004.

43. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–841. PMID: 24997786. doi:10.1038/nbt.2942.

44. Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol. 2019;37:179–185. PMID: 30718686. doi:10.1038/s41587-018-0008-8.

45. Bureau of Medical Administration, National Health and Family Planning Commission of the PRC, and Oncology Branch of Chinese Medical Association. Standardization of diagnosis and treatment for colorectal cancer in China (2015 edition). Chongqing (China): Chinese Journal of Digestive Surgery; 2015.

46. Wagner Mackenzie B, Waite DW, Taylor MW. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Front Microbiol. 2015;6:130. PMID: 25741335. doi:10.3389/fmicb.2015.00130.

47. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25:1966–1967. PMID: 19497933. doi:10.1093/bioinformatics/btp336.

48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–359. PMID: 22388286. doi:10.1038/nmeth.1923.

49. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Luo G, Deng Z, Qi Y, Guo R, et al. Metagenome and microbiome analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2017;23:1459–1461. doi:10.1038/nm.4309.