




## RESEARCH ARTICLE

# Development and evaluation of a colorectal cancer screening method using machine learning-based gut microbiota analysis

Yusuke Konishi<sup>1</sup> | Shintaro Okumura<sup>1,2,3</sup> | Tomonori Matsumoto<sup>1</sup> |  
Yoshiro Itatani<sup>3</sup> | Tsuyoshi Nishiyama<sup>4</sup> | Yuki Okazaki<sup>4</sup> | Masatsune Shibutani<sup>4</sup>  |  
Naoko Ohtani<sup>4</sup>  | Hisashi Nagahara<sup>4</sup> | Kazutaka Obama<sup>3</sup> | Masaichi Ohira<sup>4</sup> |  
Yoshiharu Sakai<sup>3</sup> | Satoshi Nagayama<sup>5,6</sup> | Eiji Hara<sup>1,2,7,8</sup> 

<sup>1</sup>Research Institute for Microbial Diseases (RIMD), Osaka University, Suita, Japan

<sup>2</sup>The Cancer Institute, Japanese Foundation for Cancer Research (JFCR), Tokyo, Japan

<sup>3</sup>Graduate School of Medicine, Kyoto University, Kyoto, Japan

<sup>4</sup>Osaka City University Graduate School of Medicine, Osaka, Japan

<sup>5</sup>The Cancer Institute Hospital, JFCR, Tokyo, Japan

<sup>6</sup>Uji-Tokushukai Medical Center, Uji, Japan

<sup>7</sup>Immunology Frontier Research Centre (IFReC), Osaka University, Suita, Japan

<sup>8</sup>Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita, Japan

## Correspondence

Satoshi Nagayama, Uji-Tokushukai Medical Center, Uji 611-0041, Japan.  
Email: [satoshi.nagayama@jfc.or.jp](mailto:satoshi.nagayama@jfc.or.jp)

Eiji Hara, RIMD, Osaka University, Suita 565-0871, Japan.  
Email: [ehara@biken.osaka-u.ac.jp](mailto:ehara@biken.osaka-u.ac.jp)

## Funding information

Japan Agency for Medical Research and Development (AMED), Grant/Award Number: 21cm0106401h0006 and JP21gm1010009; Japan Science and Technology Agency (JST), Grant/Award Number: JPMJMS2022

## Abstract

Accumulating evidence indicates that alterations of gut microbiota are associated with colorectal cancer (CRC). Therefore, the use of gut microbiota for the diagnosis of CRC has received attention. Recently, several studies have been conducted to detect the differences in the gut microbiota between healthy individuals and CRC patients using machine learning-based gut bacterial DNA meta-sequencing analysis, and to use this information for the development of CRC diagnostic model. However, to date, most studies had small sample sizes and/or only cross-validated using the training dataset that was used to create the diagnostic model, rather than validated using an independent test dataset. Since machine learning-based diagnostic models cause overfitting if the sample size is small and/or an independent test dataset is not used for validation, the reliability of these diagnostic models needs to be interpreted with caution. To circumvent these problems, here we have established a new machine learning-based CRC diagnostic model using the gut microbiota as an indicator. Validation using independent test datasets showed that the true positive rate of our CRC diagnostic model increased substantially as CRC progressed from Stage I to more than 60%

Yusuke Konishi and Shintaro Okumura are contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

for CRC patients more advanced than Stage II when the false positive rate was set around 8%. Moreover, there was no statistically significant difference in the true positive rate between samples collected in different cities or in any part of the colorectum. These results reveal the possibility of the practical application of gut microbiota-based CRC screening tests.

#### KEYWORDS

biomarkers, colorectal cancer, next generation sequencing, screening

## 1 | INTRODUCTION

In developed countries, the morbidity and mortality of colorectal cancer (CRC) are increasing year by year, and its countermeasures are becoming an urgent issue.<sup>1</sup> Since effective therapeutic drugs for CRC are still under development, the only effective measure at the moment is early detection and surgical removal of CRC. Thus, there is a need for a simple and accurate CRC screening test. Currently, the most widespread screening tests for CRC are colonoscopy and the fecal occult blood test (FOBT).<sup>1</sup> Although the colonoscopic examination achieves an accurate and sensitive diagnostic test for CRC detection, it is invasive, expensive, *labor-intensive*, and time consuming. Thus, it is difficult to use it for a population-wide CRC screening.<sup>1</sup> On the other hand, FOBT is a non-invasive, simple, and inexpensive screening test for CRC detection, but often gives false positive results especially when applied to those with bleeding by hemorrhoids or menstruation. Furthermore, the sensitivity of FOBT for early-stage CRC and proximal colon cancer is not adequately high because of the difficulty to detect the low amount of blood in the stool.<sup>2</sup> Therefore, there is a need to develop a new non-invasive, simple, and effective CRC screening test that can compensate for the problems of the FOBT.

The gut microbiota is an ecosystem created by a wide variety of bacteria that reside in the intestinal lumen. It is known that the gut microbiota helps to maintain the homeostasis of the organism through the construction of the host's immune system and assistance in food digestion.<sup>3-5</sup> However, when the composition of the gut microbiota is disrupted by overeating, an imbalanced diet, or antibiotic medication, the number of useful bacteria decreases, and pathogenic bacteria proliferate instead, causing intestinal and systemic inflammation, and metabolic disorders.<sup>3-5</sup> Accumulating evidence indicates that gut bacteria are also involved in tumorigenesis in the liver<sup>6,7</sup> and colon.<sup>8-10</sup> Notably, *Fusobacterium nucleatum*,<sup>11-13</sup> enterotoxigenic *Bacteroides fragilis*,<sup>14</sup> *pks*<sup>+</sup> *Escherichia coli*,<sup>15-18</sup> and *Peptostreptococcus anaerobius*<sup>19</sup> are reportedly involved in the development of CRC. Furthermore, we

have recently reported that *Porphyromonas gingivalis* and *Porphyromonas asaccharolytica* may promote the development of CRC through the production of butyrate.<sup>20</sup> Thus, examination of gut microbiota to detect those pathogenic bacteria would be a promising method to screen CRCs. However, since these pathogenic bacteria are not detected in all patients with CRCs and some of these bacteria are also detected in healthy individuals, screening for CRC cannot be satisfactorily performed if the presence of these pathogenic bacteria alone is used as markers.<sup>13,21,22</sup>

In recent years, several studies have been conducted to detect the differences in the gut microbiota between healthy individuals and CRC patients using machine learning-based meta-sequencing analysis of bacterial DNA, and to use this information for the diagnosis of CRC.<sup>23-27</sup> However, most of the studies have small sample sizes, and some of them do not have independent test datasets, but only cross-validate the diagnostic model using the training dataset used to create the model.<sup>23-27</sup> Since machine learning-based diagnostic models can cause overfitting if the sample size is small.<sup>25</sup> Furthermore, cross-validation is not an appropriate validation method because the results can reflect the characteristics of the training data and cause an overfitting problem. Therefore, it is essential to develop a diagnostic model using a training dataset with sufficient sample size, and at the same time, to validate it using an independent test dataset. Furthermore, some studies have used sequence data published in databases to increase the sample size. However, since differences in experimental conditions and sample population characteristics may affect the prediction results of machine learning models, care should be taken in interpreting the results obtained. It should also be noted that although shotgun metagenomic sequencing analysis has higher bacterial classification accuracy than meta-16S rRNA gene sequencing analysis, it is currently difficult to use for primary screening of large populations because of the high cost of analysis and the need for high-performance computers to analyze large amounts of data.

To solve the above problems, in this study, we attempted to develop a new diagnostic model for CRC screening

test using meta-sequencing analysis of gut bacterial 16S rRNA genes with a carefully designed machine learning approach. Notably, our model has been created using a sufficient number of training datasets and validated using independent test datasets collected from three hospitals in different regions of Japan. With these improvements, we succeeded in developing a more reliable CRC screening method using the gut microbiota, and we report it here and discuss the potential and limitations of the method using the gut microbiota.

## 2 | MATERIALS AND METHODS

### 2.1 | Human fecal sample collection

Feces were collected from study participants who visited the JFCR (Tokyo) from December 2013 to March 2015 (cohort-1), January to September 2017 (cohort-2), and May 2019 to August 2021 (cohort-3), and those who visited Kyoto University Hospital (Kyoto) or Osaka City University Hospital (Osaka) from May 2019 to August 2021 (cohort-3) using a fecal sampling tool (TechnoSuruga Laboratory). All study participants underwent colonoscopy at the time of stool collection. Colorectal cancer patients (CRC patients) were defined as patients with primary malignant epithelial colorectal tumors according to the Third English Edition of the Japanese Classification of Colorectal, Appendiceal, and Anal Carcinoma.<sup>2</sup> Advanced adenoma patients were defined as patients with colorectal adenomas larger than 10 mm in diameter. Healthy individuals (HI) were defined as individuals without colorectal cancers nor colorectal advanced adenomas. HI were classified into two groups: clean HI who had no colorectal adenomas and HI with colorectal polyps smaller than 10 mm in size. We excluded those with a history of inflammatory bowel disease, prior gastrointestinal reconstructive surgery, severe liver dysfunction, anticancer and/or antibiotic treatment within 1 month, stool collection within 3 days of colonoscopy, and those without access to detailed clinical information. Patients who received chemotherapy, radiation therapy, or colonic stent placement before fecal sample collection, who had fecal samples collected after endoscopic resection of tumors, or patients whose tumors were not primary colorectal tumors (e.g., squamous cell carcinoma of the anal canal cancer, metastasis or direct invasion of other cancers to the large intestine) were also excluded. HI with a history of colorectal cancers, or with malignant tumors other than colorectal cancer or abnormal endoscopic findings, such as enteritis and hamartomas at the time of stool collection were excluded. Written informed consent was obtained from all participants for the use of anonymized samples

and the publication of the patients' clinical information under the protocol approved by the ethics committee of the JFCR hospital, Kyoto University Hospital, and Osaka City University Hospital. The tumor profiles of CRC patients were classified based on the Third English Edition of the Japanese Classification of Colorectal, Appendiceal, and Anal Carcinoma.<sup>2</sup>

### 2.2 | 16S rRNA gene sequencing analysis and microbiome analysis

Bacterial DNA extraction from fecal samples was performed using a QIAamp Fast DNA Stool Mini Kit (QIAGEN) (samples of cohort-1) or a Magstration System 12GC (Precision System Science) in TechnoSuruga Laboratory (samples of cohort-2) or a GENE STAR PI-480 automated DNA isolation system (Kurabo Industries, Ltd., Osaka, Japan) (samples of cohort-3). The polymerase chain reaction (PCR) amplification of the V1-V2 region of the bacterial 16S rRNA gene was performed using KAPA HiFi Hot Start Ready Mix (Roche) with universal 16S rRNA primers followed by the secondary amplification adding the Illumina flow cell adapters and indices. The PCR primers used are shown in Table S1. Meta-16S rRNA gene sequencings were carried out per 192 samples on the Illumina MiSeq platform (Illumina Inc.) using MiSeq Reagent Kit v2 (Illumina Inc.) (paired-end, 250 cycles × 2). These processes were performed at Biken Biomics, Inc. Sequencing reads were processed according to the QIIME2 (version 2020.8) pipeline.<sup>28</sup> Fastq files were de-noised with the DADA2 plugin<sup>29</sup> and amplicon sequence variants (ASVs) were counted. These processes were performed separately for samples from the training data and the test data. Subsequently, de novo clustering was performed on ASVs using the VSEARCH plugin<sup>30</sup> to obtain operational taxonomic units (OTUs) with a similarity of more than 99%. Open-reference clustering based on OTUs detected from the training data were performed on the test data. Finally, the OTU counts were converted to relative abundance per sample. A phylogenetic tree was generated from the ASVs, and beta diversity analyses (principal coordinate analyses of weighted UniFrac distance) were performed with a sampling depth of 10,000 reads. Phylogenetic classification of the detected OTUs was performed by a Naive Bayes classifier trained on the SILVA 16S rRNA sequence database (version 138)<sup>31</sup> in the QIIME2 pipeline. Identification of the specific bacterial species corresponding to each OTU was performed by using the 16S rRNA database provided by the National Center for Biotechnology Information (NCBI) (last modified on 12 June 2021) and a similarity search with BLAST+ (version 2.9.0).<sup>32</sup>

## 2.3 | Statistical modeling

For the construction of the colorectal cancer screening model, h2o.automl function with 10-fold cross-validations in the h2o package of R (version 3.32.0.1) (<https://www.h2o.ai/>) was performed. After repeating this process 10 times, the StackedEnsemble\_BestOfFamily model with the highest AUC for the cross-validation predictions was selected. The h2o package displays the feature (variable) importance scaled between 0 and 1, except for stacked ensemble learning. In this study, the scaled feature importance in a stacked ensemble model was defined as the rescaled sum of the product of the scaled importance of its constituent models within a meta-learner and the scaled importance of each feature within each model.

## 2.4 | Quantitative real-time PCR analysis

Quantitative real-time PCR was performed on Thermal Cycler Dice® Real-Time System III (Takara Bio Inc.) using TB Green® Premix Ex Taq™ II (Takara Bio Inc.). Universal 16S rDNA was used as internal control, and the abundances of the bacteria were expressed as relative levels to 16S rDNA. The PCR primer sequences used are shown in Table S1.

## 2.5 | Statistical analysis

Statistical analysis was performed using R (version 4.0.5). The differences in the characteristics of the samples were analyzed by the Wilcoxon rank sum test and the chi-squared test. For the performance testing of the colorectal cancer screening model, a bootstrapping method with 10,000 resamples by the pROC package of R (version 1.16.2)<sup>33</sup> was used. To compare the variables of the multiple sample groups, the pairwise method with the adjustment of *p* values by the Benjamini–Hochberg false-discovery rate correction at 0.05 was performed. The Cochran–Armitage test for trends in proportions was used to evaluate the statistical significance of trends in positive rates across colorectal cancer progression. Statistical tests were two-tailed and *p* < 0.05 was considered significant.

# 3 | RESULTS

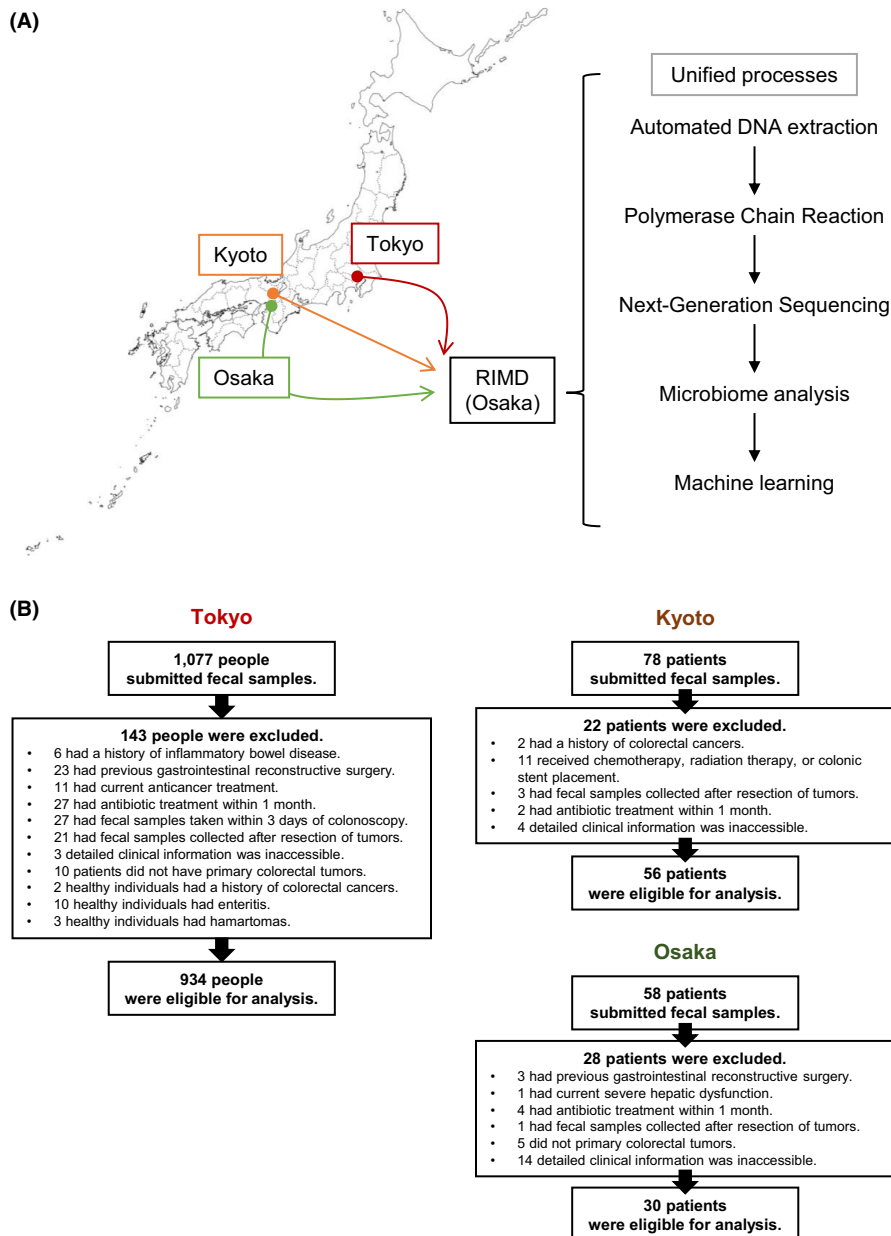
## 3.1 | Strategies for optimizing the diagnostic model of CRC

In our previous study, stool samples from healthy individuals and patients with CRC were collected twice

at the JFCR hospital in Tokyo (cohort-1 in 2013–2015 and cohort-2 in 2017–2018) and their gut microbiota profiles were analyzed by bacterial 16S rRNA gene meta-sequencing.<sup>20</sup> As a result, we found 12 common CRC-related bacterial species that were significantly increased in CRC patients but almost undetectable in healthy individuals in both cohorts.<sup>20</sup> Unexpectedly, however, the CRC diagnostic model created by machine learning based on the dataset of cohort-1 failed to correctly diagnose the sample of cohort-2. To investigate the cause of this, we compared the  $\beta$ -diversity of the gut microbiota between the two cohorts and found that they differed substantially (Figure S1). These two cohorts used different methods to extract DNA from stool, which may have caused differences in the gut microbiota profile and reduced diagnostic efficiency. Therefore, considering this result and the problems that have been pointed out regarding the generation of a CRC diagnostic model using the gut microbiota and machine learning,<sup>34–36</sup> we attempted to develop a new CRC diagnostic model by paying attention to the following three points<sup>1</sup>: Create a diagnostic model for CRC screening using a sufficient number of training datasets and validate it appropriately using independent samples as test datasets.<sup>2</sup> Stool collection, DNA extraction, and sequencing are performed based on a consistent pipeline to minimize data variability due to differences in operations.<sup>3</sup> Test samples were collected at three hospitals located in different regions of Japan (Tokyo) to evaluate the diagnostic robustness regardless of the region of donor residency (Figure 1A).

## 3.2 | Determining the sample size required to create a CRC diagnostic model

The small sample size can lead to overfitting in generating diagnostic models based on machine learning. In addition, cross-validation, which is performed using part of the training dataset rather than being evaluated using an independent test dataset, also increases the risk of overfitting.<sup>25</sup> Therefore, we attempted to estimate the number of training data samples required to create a diagnostic model for the CRC screening test using the training dataset and independent test dataset of cohort-1. Note that because there is a slight batch effect for each sequencing run, there may be some errors when using amplicon sequence variants (ASVs), which are classified as identical only if they are 100% identical, in the analysis. To circumvent this problem, we utilized an operational taxonomic unit (OTU) clustering ASVs with 99% identity rather than ASV itself, and then create and validate a CRC diagnostic model using h2o AutoML (<https://www.h2o.ai/>)



**FIGURE 1** The strategy of this study. (A) flow diagram of cohort-3. Stool samples from three hospitals in Tokyo (JFCR hospital), Kyoto (Kyoto University Hospital), and Osaka (Osaka City University Hospital) were collected in one place and analyzed using a uniform method. (B) Workflow chart for enrolling healthy individuals (HI) and CRC patients for microbiome analysis in cohort-3

h2o.ai/), an open-source machine learning platform (Figure S2). The h2o AutoML can run seven different machine learning algorithms: Generalized Linear Model, Distributed Random Forest, Extremely Randomized Trees, Gradient Boosting Machine, XGBoost, Deep learning, and Stacked Ensemble. However, it is generally known that Stacked Ensemble has the best performance, and Stacked Ensemble is further divided into Stacked Ensemble\_AllModels and Stacked Ensemble\_BestOfFamily (Figure S2). Although the performance of Stacked Ensemble\_AllModels and Stacked Ensemble\_BestOfFamily is almost the same, the model created by Stacked Ensemble\_AllModels has a very large data size (Figure S2). Therefore, in this study, we decided to use Stacked Ensemble\_BestOfFamily which has a manageable data size to create a CRC diagnostic model. We

created a diagnostic model by varying the number of clean healthy individuals and CRC patients in the training dataset of cohort-1, and evaluated the diagnostic efficiency of the model by calculating the AUC (area under the curve) of the ROC (receiver operating characteristic) curve using independent test dataset<sup>37</sup> (Table S2 and Figure S3A). As the number of clean healthy individuals and CRC patients used to create the model was increased step by step, the value of AUC increased until the number of subjects reached 120 each (Figure S3B). However, when the sample size was increased to more than 120 subjects each, the AUC values did not increase anymore (Figure S3B), indicating that the training data of 120 clean healthy individuals and 120 CRC patients were necessary and sufficient for creating the CRC diagnostic model.

### 3.3 | Development of a diagnostic model for CRC screening

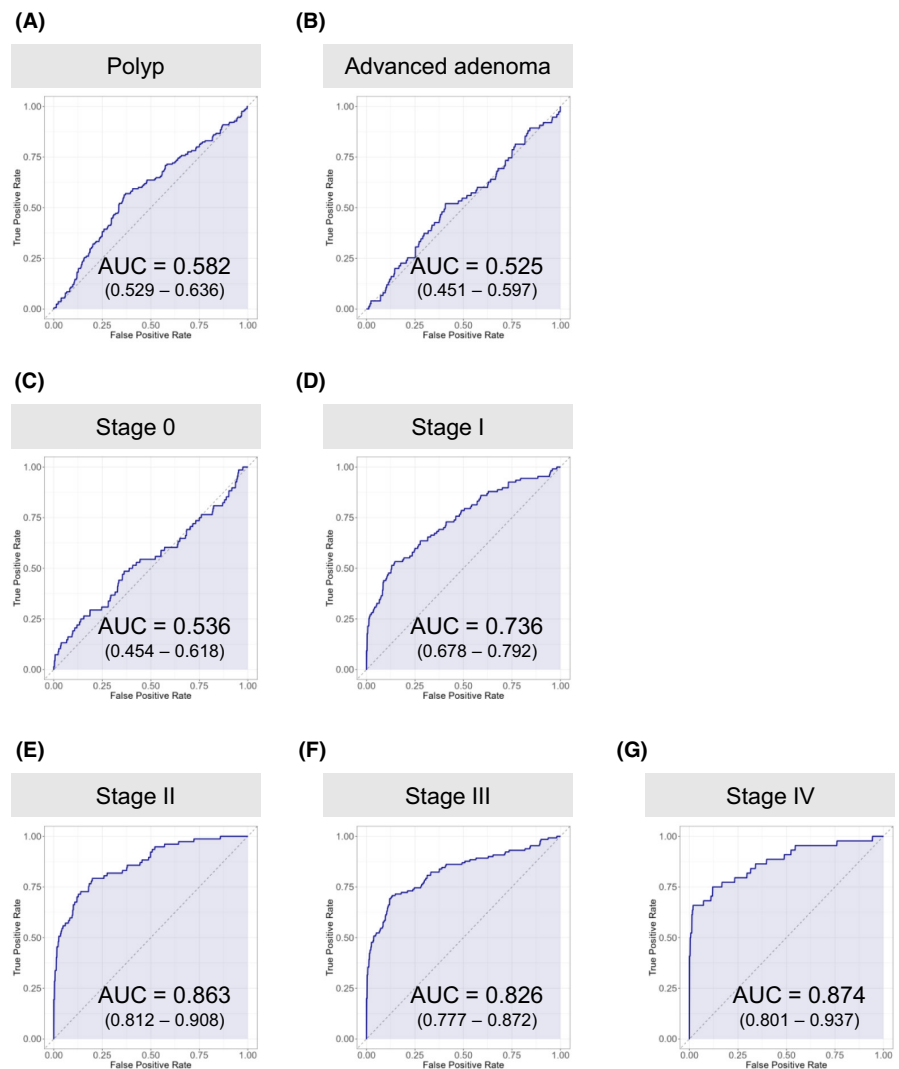
Given the estimated sample size required to create a diagnostic model, samples required for the development of the CRC diagnostic model (more than 120 clean healthy

individuals and CRC patients each) were newly collected in Tokyo for training data as cohort-3. Independent samples for test data were also collected in Kyoto and Osaka in addition to Tokyo. All collected stool samples were subjected to DNA extraction, sequencing of bacterial 16S rRNA gene, and clustering of sequence data by a consistent

**TABLE 1** Summary of the study participants in cohort-3

Study population	Clean HI	HI with polyps	Advanced adenoma patients	CRC patients stage				
				0	I	II	III	IV
Training data								
Tokyo	120			22	43	31	17	7
Test data								
Tokyo	234	165	75	30	46	36	76	32
Kyoto					21	16	19	
Osaka				1	8	12	5	4
Cohort-2 (Tokyo)	83			37	32	13	30	8

Abbreviation: HI: Healthy individuals.



**FIGURE 2** Evaluation of CRC diagnostic model by ROC curve. (A)–(G) ROC curves were drawn separately for the different sample categories and for each stage of CRC progression. Healthy individuals with colorectal polyps (A), advanced adenoma patients (B), Stage 0 (C), Stage I (D), Stage II (E), Stage III (F), Stage IV (G). AUC (area under the curve). Ranges in parentheses are 95% confidence intervals with 10,000 bootstrap replicates

pipeline (Figure 1A). Among a total of 1213 samples collected, 193 samples inappropriate for analysis were excluded, and 1020 samples were processed to create a CRC diagnostic model (Figure 1B). From cohort-3, 120 samples each of clean healthy individuals and CRC patients from the Tokyo sample were selected as training data (Table 1), and a new CRC diagnostic model was generated by the Stacked Ensemble\_BestOfFamily machine learning algorithm. Subsequently, the model generated from training data were evaluated by test dataset. It should be noted that we were unable to collect samples from healthy individuals in Kyoto and Osaka and that the distribution of cancer progression stage among CRC patients was uneven (Table 1). Therefore, in order to increase the number of test data, we also used the dataset of the cohort-2, where DNA extraction from stool was performed in the same way as in the cohort-3, and its  $\beta$ -diversity was similar to that of the cohort-3 as judged by principal coordinate analysis (PCoA) of weighted UniFrac distance<sup>38</sup> (Figure S4). Thus, the test dataset was composed of all samples that were not used for training data in cohort-3 and cohort-2, whose specimens were processed in a consistent way (Table 1). Note that the value of AUC is highly dependent on the sample composition of the dataset. Therefore, as a measure to evaluate the diagnostic model, we decided to also compare the true positive rate (sensitivity) for each stage of cancer when the threshold of the algorithm was set to a value that is expected to result in a false positive rate (i.e., 1 - specificity) of about 8%, as in Zeller et al.<sup>26</sup>

Judging from the independent test dataset, the AUC and true positive rate of CRC patients increased substantially as CRC progressed from stage I, with AUCs above 0.80 and true positive rates above 60% in patients with stage II or higher CRC (Figure 2 and Table 2). The true positive rate also significantly increased in correlation with the depth of tumor invasion (T classification),<sup>2</sup> indicating that our model is suitable to detect advanced CRCs (Table 2). On the other hand, the true positive rate of patients with early-stage cancers belonging to stage 0 or T0 was as low as 19.1% and the AUC = 0.536, almost indistinguishable from clean healthy individuals (Table 2 and Figure 2). Healthy individuals with colorectal polyps and patients with advanced adenoma were also indistinguishable from clean healthy individuals (Table 2 and Figure 2A and B). Consistent with these results, the mean relative abundance of the top 50 most important bacteria (OTUs) for diagnosis in our CRC diagnostic model changed after stage I of CRC (Figure 3). It is worth emphasizing that there was no significant difference in the true positive rate between the data from the three hospitals (Tokyo, Kyoto, and Osaka) (Table 3) and no statistically significant difference in the true positive rate in any part of the colorectum (Table 4). It

should also be noted that although there was a statistically significant difference in age and BMI (body mass index) between the healthy individual group and the CRC patient group in the test dataset, the CRC diagnostic model obtained by training data combining gender, age, BMI, and gut microbiota composition did not significantly improve performance from the model trained on gut microbiota composition data alone (Figure 4). This result suggests that changes in gut microbiota are more strongly correlated with colorectal cancer than age or BMI.

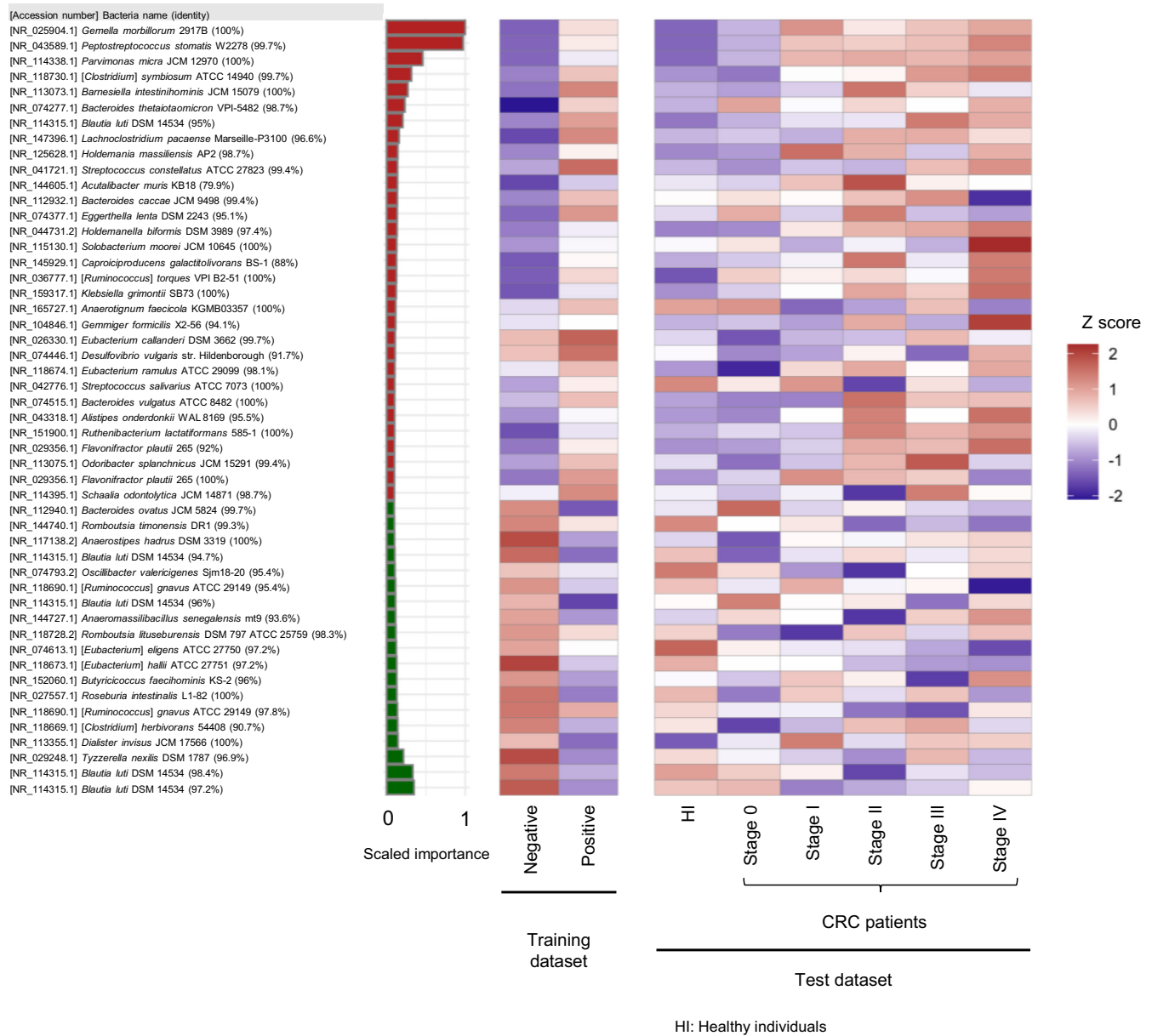
### 3.4 | Comparison with qPCR methods for detection of specific gut bacteria

Recently, Guo et al.<sup>39</sup> reported that a highly accurate CRC diagnostic model was developed by combining the results of qPCR quantification of the abundance of three gut bacteria, *Fusobacterium nucleatum*, *Faecalibacterium prausnitzii*, and *Bifidobacterium* spp. Therefore, in order to compare the efficiency of our model and Guo's model,<sup>39</sup> we randomly selected samples of 67 clean healthy individuals

TABLE 2 Positive rate of our CRC diagnostic model for clean healthy individuals, healthy individuals with colorectal polyps, advanced adenoma patients, and patients with CRC in different stages of CRC progression (stage and T classification)

Group	n	Positive rate (%)
Clean healthy individuals	317	10.7 (7.3–14.2)
Healthy individuals with polyps	165	13.3 (8.5–18.8)
Advanced adenoma patients	75	12.0 (5.3–20.0)
CRC patients		
Stage		
0	68	19.1 (10.3–29.4)
I	107	45.8 (36.4–55.1)
II	77	66.2 (55.8–76.6)
III	130	63.1 (54.6–71.5)
IV	44	68.2 (54.5–81.8)
		$p = 7.99 \times 10^{-10}$
T classification		
Tis	68	19.1 (10.3–29.4)
T1	77	48.1 (36.4–59.7)
T2	60	46.7 (33.3–60.0)
T3	166	66.3 (59.0–73.5)
T4	55	67.3 (54.5–80.0)
		$p = 6.88 \times 10^{-11}$

Ranges in parentheses are 95% confidence intervals with 10,000 bootstrap replicates. Statistical significance was determined with the Cochran–Armitage test for trends in proportions.  $p$  values < 0.05 were considered significant.



**FIGURE 3** The top 50 most important bacteria (OTUs) for diagnosis in our CRC diagnostic model developed in this study. The heat map shows the mean relative abundance of the top 50 most important bacteria (OTUs) for diagnosis in our CRC diagnostic model in the training and test datasets. The bar chart shows the scaled feature (variable) importance of each OTU. The color of bar chart is red if the average relative presence ratio of positive samples is greater than the average relative presence ratio of negative samples in the training dataset, and green otherwise. The relative amounts were multiplied by 10,000 and logarithmically transformed (pseudo-count = 1), then the mean of each sample group was calculated and finally normalized by the z-score

and 59 CRC patients from our cohort-3, performed a qPCR analysis with the same primers as Guo et al., and applied it to Guo et al.'s diagnostic model. Although the CRC diagnostic model by Guo et al. achieved very high diagnostic accuracy with  $AUC = 0.964$  in their dataset, their model only showed  $AUC = 0.654$  in our cohort-3 dataset, which was lower than our CRC diagnostic model (Figure 5). It is unclear why this difference occurred, but it may reflect differences in DNA extraction methods or regional differences between Japan and China. However, our data

suggest that screening for CRC using the diagnostic model of Guo et al. would be difficult, at least in a Japanese sample.

## 4 | DISCUSSION

The colon contains the highest density of metabolically active microbiota, and it is becoming clear that changes in the composition of the gut microbiota are associated



**TABLE 3** Positive rate of our CRC diagnostic model for CRC patients in different hospitals by stage

Group	<i>n</i>	Positive rate (%)
Clean healthy individuals	317	10.7 (7.3–14.2)
CRC patients		
Stage 0/I/II		
Tokyo	112	47.3 (37.5–56.3)
Kyoto	37	59.5 (43.2–75.7)
Osaka	21	57.1 (38.1–76.2)
<i>p</i> = 0.371		
Stage III/IV		
Tokyo	108	67.6 (59.3–75.9)
Kyoto	19	68.4 (47.4–89.5)
Osaka	9	66.7 (33.3–100)
<i>p</i> = 0.995		

Ranges in parentheses are 95% confidence intervals with 10,000 bootstrap replicates. Statistical significance was determined with the chi-squared test. *p* values < 0.05 were considered significant.

with colorectal cancer.<sup>8–10</sup> Although only a limited number of gut bacteria have been reported to be involved in the development of CRC,<sup>11–20</sup> many studies have reported increased or decreased abundance of certain gut bacteria in patients with CRC.<sup>26,27,40,41</sup> Therefore, changes in the gut microbiota may be useful for screening for colorectal cancer. In this study, we established a new machine learning-based CRC diagnostic model using gut microbiota as an indicator, overcoming the problems that have been pointed out in the past, such as sample size and independent test datasets. However, since several attempts have already been made to use the gut microbiota as an indicator for CRC diagnosis by machine learning,<sup>23–27</sup> it is important to compare our CRC diagnostic model with Zeller's model, which seems to be the most reliable so far because it uses independent test dataset for validation.<sup>26</sup> In Zeller's model, however, the training data were small (88 healthy individuals and 53 CRC patients), and the test data included only 38 CRC patients.<sup>26</sup> Furthermore, most of the healthy individual data in the test dataset were cited from other studies.<sup>26</sup> It is, therefore, possible that the characteristics of the other study's data, rather than being healthy individuals or not, are responsible for the differences between CRC patients and healthy individuals and increase the accuracy of the diagnosis in a pretense. In addition, the diagnostic model obtained in Zeller's study showed AUC = 0.85 in the test data, which at first glance appears to be a high performance,<sup>26</sup> but the reason for this value may be that only two patients with CRC were in Stage 0, and the others were in more advanced stages. In fact, in Zeller's CRC diagnostic model, when the false

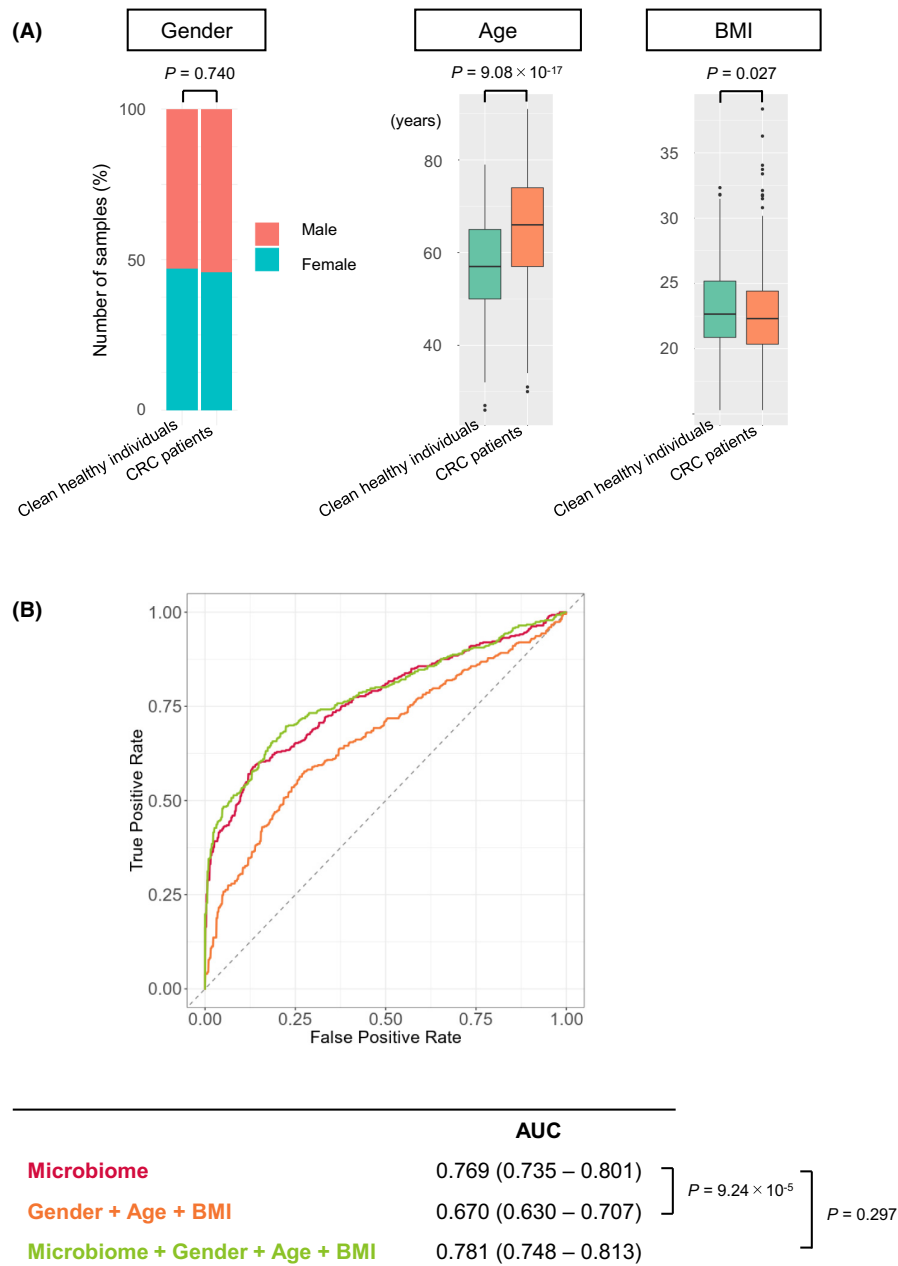
**TABLE 4** Positive rate by stage of our CRC diagnostic model for CRC with a different location

Group	<i>n</i>	Positive rate (%)
Clean healthy individuals	317	10.7 (7.3–14.2)
CRC patients		
Stage 0		
Proximal colon	32	9.4 (0.0–21.9)
Distal colon	13	30.8 (7.7–53.8)
Rectum	23	26.1 (8.7–43.5)
<i>p</i> = 0.148		
Stage I		
Proximal colon	36	38.9 (22.2–55.6)
Distal colon	24	45.8 (25.0–66.7)
Rectum	47	51.1 (36.2–66.0)
<i>p</i> = 0.544		
Stage II		
Proximal colon	32	62.5 (46.9–78.1)
Distal colon	22	59.1 (36.4–77.3)
Rectum	23	78.3 (60.9–95.7)
<i>p</i> = 0.335		
Stage III/IV		
Proximal colon	45	57.8 (42.2–71.1)
Distal colon	43	58.1 (44.2–72.1)
Rectum	86	70.9 (61.6–80.2)
<i>p</i> = 0.203		

Ranges in parentheses are 95% confidence intervals with 10,000 bootstrap replicates. Statistical significance was determined with the chi-squared test. *p* values < 0.05 were considered significant.

positive rate is 7.7%, the true positive rate is less than 60% even for Stage III/IV.<sup>26</sup> On the other hand, our CRC diagnostic model has a true positive rate of more than 60% for CRC that had progressed to stage II or higher (Table 2).

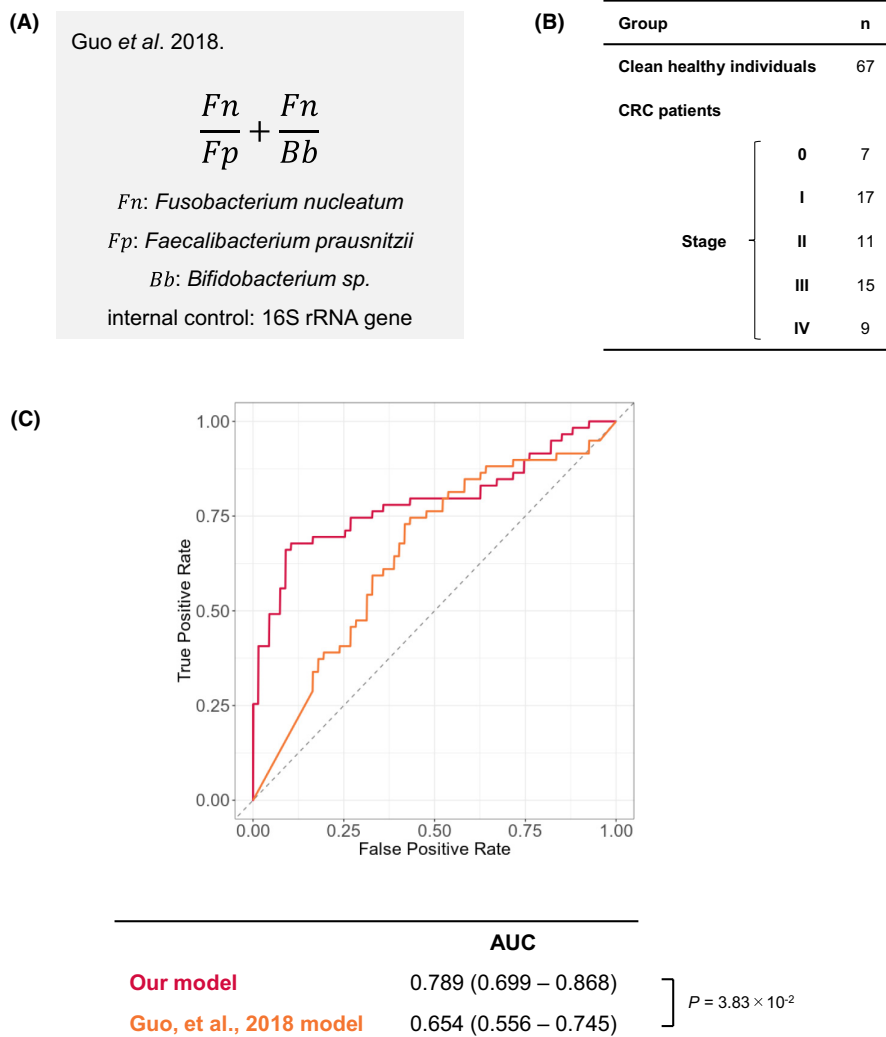
Nevertheless, *Gemella morbillorum*, *Peptostreptococcus stomatis*, and *Parvimonas micra*, which ranked among the most important bacterial for diagnosis in our CRC diagnostic model (Figure 3), have also been identified as CRC-associated bacteria in other reports<sup>20,27,40,41</sup> including Zeller's study,<sup>26</sup> suggesting that these bacterial species are likely to be increased commonly in CRC patients, regardless of geographical localization, technical protocols, and ethnicity. However, since, there have been no reports to date that these bacteria are involved in the development of CRC, it is possible that some bacteria have increased or decreased as a result of developing CRC (Figure 3). In this regard, it is important to note that several known tumor-promoting bacteria, such as *Fusobacterium nucleatum*,<sup>11–13</sup> *Peptostreptococcus anaerobius*,<sup>19</sup> *Porphyromonas gingivalis*,<sup>20</sup> and *Porphyromonas asaccharolytica*,<sup>20</sup> are not listed



**FIGURE 4** Gender, age, and body mass index (BMI) of the cohort-3 samples. (A) Distribution of gender, age, and BMI in the test dataset (317 clean healthy individuals and 426 CRC patients). Missing data have been skipped. The boxes in the graph of age and BMI represent 25th–75th percentiles, black lines indicate the median, whiskers extend to the maximum and minimum values within 1.5× the interquartile range and dots indicate outliers. Statistical significance was determined with the chi-squared test (gender) or the two-tailed Wilcoxon rank sum test (age and BMI).  $p$  values  $< 0.05$  were considered significant. (B) ROC curves for the test datasets of the “microbiome” model, the “Gender + Age + BMI” model, and the “Gender + Age + BMI + microbiome” model. The “microbiome” model was trained to distinguish between CRC patients and clean healthy individuals based solely on gut microbiota. The “Gender + Age + BMI” model was trained based on gender, age, and BMI. The “Gender + Age + BMI + microbiome” model was trained based on gender, age, BMI, and gut microbiome. Despite the significant differences in Age and BMI between clean healthy individuals and CRC patients (A), the AUC of the “Gender + Age + BMI” model was low. The AUC of the “Gender + Age + BMI + microbiome” model did not differ from that of the “microbiome” model. Ranges in parentheses are 95% confidence intervals with 10,000 bootstrap replicates. Statistical significance was determined with a bootstrapping method with 10,000 resamples.  $p$  values  $< 0.05$  were considered significant

in Figure 3, as they overlap in appearance pattern with the above-mentioned bacteria and do not further improve the diagnostic efficiency of CRC. Therefore, the bacteria shown

in Figure 3, which are good diagnostic markers for CRC, may not necessarily be the bacteria involved in the development of CRC. In other words, we should be cautious about



**FIGURE 5** Comparison with quantitative PCR method (Guo *et al.*, 2018). (A) Formula for Guo's model, using the relative presence of the three bacteria. (B) The sample configuration used to validate Guo's model. (C) Comparison of the ROC curves from our CRC diagnostic model and Guo's model for the samples listed in Panel B. Ranges in parentheses are 95% confidence intervals with 10,000 bootstrap replicates. Statistical significance was determined with a bootstrapping method with 10,000 resamples. *p* values < 0.05 were considered significant

discussing the causal relationship between gut bacteria and CRC based only on quantitative changes in gut bacteria, and biological function analysis of gut bacteria is also necessary.

Although we did not perform FOBT in this cohort study, the test data of cohort-3 included 21 clean healthy individuals and 13 healthy individuals with colorectal polyps who had false positive results on FOBT. Notably, 76.2% (95% confidence interval, 57.1%–90.5%) of these 21 clean healthy individuals and 76.9% (53.8%–100.0%) of these 13 healthy individuals with colorectal polyps tested negative in our CRC diagnostic model, suggesting that combining FOBT with our gut microbiota-based diagnostic model may reduce false positives by more than 75%. Moreover, as mentioned earlier, the FOBT is known to have a poor positive rate for proximal colon cancer.<sup>42</sup> On the other hand, our model showed no statistically significant difference in the positive rate in any part of the colorectum (Table 4). Proximal colon cancer is known to have a low survival rate, which may be partly due to the delay in detection caused by the low true positive rate by FOBT. Therefore, our model may also be useful for improving the survival

rate of proximal colon cancer. However, contrary to our expectations, the CRC diagnostic model we developed could not sufficiently achieve early detection, which is an important issue in CRC screening tests. This indicates that in the early stages of CRC, the gut microbiota is not altered enough to be detected by stool examination, implying that there may be limitations in analyzing the gut microbiota in feces for the diagnosis of early stage CRC. Further studies are therefore needed to overcome this limitation.

#### ACKNOWLEDGMENTS

The authors thank Dr. Kenya Honda (Keio University) for valuable advice for gut bacterial culture and Dr. Shota Nakamura (Osaka University) for performing a 16S rRNA gene meta-sequencing analysis. The authors also thank Ms. Satoko Otomi, Ms. Shizuka Murata, Mr. Masanori Tanaka, Dr. Tomohiro Tsuchida, and Dr. Shoichi Saito (JFCR) for collecting human specimens throughout this study. The authors are grateful to members of Hara's laboratory for helpful discussion during the preparation of this manuscript. This work was supported

in part by grants from the Japan Agency for Medical Research and Development (AMED) under grant number 21cm0106401h0006 and JP21gm1010009, and from the Japan Science and Technology Agency (JST) under grant number JPMJMS2022.

## CONFLICT OF INTEREST

None.

## AUTHOR CONTRIBUTIONS

Eiji Hara and Satoshi Nagayama designed the experiments. Yusuke Konishi did most of the analyses. Shintaro Okumura collected and organized the clinical information of the healthy subjects and colorectal cancer patients who participated in this study. All the authors discussed the results and commented on the manuscript.

## ETHICS STATEMENT

This study was approved by the Institutional Review Board of Osaka University, JFCR hospital, Kyoto University Hospital, and Osaka City University Hospital.

## DATA AVAILABILITY STATEMENT

Microbiome analysis (bacterial 16S rRNA gene meta-sequence) data generated in this study have been deposited in the DNA Data Bank of Japan (DDBJ) with the accession codes DRA011735 (cohort-1) or DRA011736 (cohort-2) or DRA012966 or DRA013152 (cohort-3) (<https://www.ddbj.nig.ac.jp>). The deposited data are available in NCBI under accession numbers DRA011735, DRA011736, DRA012966, and DRA013152. The databases referred in microbiome analysis are as follows: SILVA (<https://www.arb-silva.de/>) and National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>).

## ORCID

Masatsune Shibutani  <https://orcid.org/0000-0002-2164-7394>

Naoko Ohtani  <https://orcid.org/0000-0001-8934-0797>

Eiji Hara  <https://orcid.org/0000-0001-7821-3960>

## REFERENCES

- Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209-249. doi:10.3322/caac.21660
- Japanese Society for Cancer of the Colon and Rectum. Japanese classification of colorectal, appendiceal, and anal carcinoma: the 3d English edition. *J Anus Rectum Colon*. 2019;3:175-195. doi:10.23922/jarc.2019-018
- Schroeder BO, Backhed F. Signals from the gut microbiota to distant organs in physiology and disease. *Nat Med*. 2021;22:1079-1089. doi:10.1038/nm.4185
- Janney A, Powrie F, Mann EH. Host-microbiota maladaptation in colorectal cancer. *Nature*. 2020;585:509-517. doi:10.1038/s41586-020-2729-3
- Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiology*. 2021;19:55-71. doi:10.1038/s41579-020-0433-9
- Yoshimoto S, Loo TM, Atarashi K, et al. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature*. 2013;499:97-101. doi:10.1038/nature12347
- Schwabe RF, Jobin C. The microbiome and cancer. *Nat Rev Cancer*. 2013;13:800-812. doi:10.1038/nrc3610
- Wong SH, Yu J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat Rev Gastroenterology & Hepatology*. 2019;16(11):690-704. doi:10.1038/s41575-019-0209-8
- Sears CL, Garrett WS. Microbes, microbiota, and colon cancer. *Cell Host Microbe*. 2014;15:317-328. doi:10.1016/j.chom.2014.02.007
- Irrazabal T, Belcheva A, Girardin SE, Martin A, Philpott DJ. The multifaceted role of the intestinal microbiota in colon cancer. *Mol Cell*. 2014;54:309-320. doi:10.1016/j.molcel.2014.03.039
- Castellarin M, Warren RL, Freeman JD, et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res*. 2012;22:299-306. doi:10.1101/gr.126516.111
- Kostic AD, Gevers D, Pedamallu CS, et al. Genomic analysis identifies association of fusobacterium with colorectal carcinoma. *Genome Res*. 2012;22:292-298. doi:10.1101/gr.126573.111
- Kostic AD, Chun E, Robertson L, et al. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe*. 2013;14:207-215. doi:10.1016/j.chom.2013.07.007
- 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;15:1016-1022. doi:10.1038/nm.2015
- Nougayrède JP, Homburg S, Taieb F, et al. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science*. 2006;313:848-851. doi:10.1126/science.1127059
- Arthur JC, Perez-Chanona E, Mühlbauer M, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science*. 2012;338:120-123. doi:10.1126/science.1224820
- Wilson MR, Jiang Y, Villalta PW, et al. The human gut bacterial genotoxin colibactin alkylates DNA. *Science*. 2019;363:eaar7785. doi:10.1126/science.aar7785
- Pleguezuelos-Manzano C, Puschhof J, Huber AR, et al. Mutational signature in colorectal cancer caused by genotoxic pks(+) *E. coli*. *Nature*. 2020;580:269-273. doi:10.1038/s41586-020-2080-8
- Long X, Wong CC, Tong L, et al. *Peptostreptococcus anaerobius* promotes colorectal carcinogenesis and modulates tumour immunity. *Nat Microbiology*. 2019;4:2319-2330. doi:10.1038/s41564-019-0541-3
- Okumura S, Konishi Y, Narukawa M, et al. Gut bacteria identified in colorectal cancer patients promote tumorigenesis via butyrate secretion. *Nat Commun*. 2021;12:5674. doi:10.1038/s41467-021-25965-x
- Bolej A, Hechenbleikner EM, Goodwin AC, et al. The *Bacteroides fragilis* toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin Infectious Diseases*. 2015;60(2):208-215. doi:10.1093/cid/ciu787

22. Shimpoh T, Hirata Y, Ihara S, et al. Prevalence of pks-positive *Escherichia coli* in Japanese patients with or without colorectal cancer. *Gut Pathogens*. 2017;9(1):1-8. doi:10.1186/s13099-017-0185-x
23. Zackular JP, Rogers MA, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev Res (Phila)*. 2014;7:1112-1121. doi:10.1158/1940-6207.Capr-14-0129
24. Baxter NT, Ruffin MT, Rogers MA, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med*. 2016;8:37. doi:10.1186/s13073-016-0290-3
25. Ai L, Tian H, Chen Z, et al. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget*. 2017;8(6):9546. doi:10.18632/oncotarget.14488
26. 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. doi:10.15252/msb.20145645
27. Yachida S, Mizutani S, Shiroma H, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med*. 2019;25:968-976. doi:10.1038/s41591-019-0458-7
28. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnology*. 2019;37(8):852-857. doi:10.1038/s41587-019-0209-9
29. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13(7):581-583. doi:10.1038/nmeth.3869
30. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 2016;4:e2584. doi:10.7717/peerj.2584
31. Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nuc Acids Res*. 2007;35(21):7188-7196. doi:10.1093/nar/gkm864
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403-410. doi:10.1016/s0022-2836(05)80360-2
33. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. doi:10.1186/1471-2105-12-77
34. He Y, Wu W, Zheng HM, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med*. 2018;24(10):1532-1535. doi:10.1038/s41591-018-0164-x
35. Voigt AY, Costea PI, Kultima JR, et al. Temporal and technical variability of human gut metagenomes. *Genome Biol*. 2015;16(1):1-12. doi:10.1186/s13059-015-0639-8
36. Sinha R, Abu-Ali G, Vogtmann E, et al. Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (MBQC) project consortium. *Nat Biotechnology*. 2017;35(11):1077-1086. doi:10.1038/nbt.3981
37. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006;27(8):861-874. doi:10.1016/j.patrec.2005.10.010
38. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. *ISME J*. 2011;5(2):169-172. doi:10.1038/ismej.2010.133
39. Guo S, Li L, Xu B, et al. A simple and novel fecal biomarker for colorectal cancer: ratio of fusobacterium nucleatum to probiotics populations, based on their antagonistic effect. *Clin Chemistry*. 2018;64(9):1327-1337. doi:10.1373/clinchem.2018.289728
40. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med*. 2019;25:679-689. doi:10.1038/s41591-019-0406-6
41. Thomas AM, Manghi P, Asnicar F, 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. doi:10.1038/s41591-019-0405-7
42. Morikawa T, Kato J, Yamaji Y, Wada R, Mitsushima T, Shiratori Y. A comparison of the immunochemical fecal occult blood test and total colonoscopy in the asymptomatic population. *Gastroenterology*. 2005;129(2):422-428. doi:10.1016/j.gastro.2005.05.056

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Konishi Y, Okumura S, Matsumoto T, et al. Development and evaluation of a colorectal cancer screening method using machine learning-based gut microbiota analysis. *Cancer Med*. 2022;11:3194–3206. doi: [10.1002/cam4.4671](https://doi.org/10.1002/cam4.4671)