Accurate Five-category Classification for Colorectal Cancer Using Gut Microbiome 16S rRNA Sequencing

Liying Zhang
New Horizon Health  https://orcid.org/0000-0003-3276-8009

Jiaqi Zhu
New Horizon Health

Qiutao Ding
New Horizon Health  https://orcid.org/0000-0002-2184-884X

Yanqi Huang
"Cancer Institute (Key Laboratory of Cancer Prevention and Intervention, China National Ministry of Education, Key Laboratory of Molecular Biology in Medical Sciences, Zhejiang Province, China), Th

Hongbo Zhang
New Horizon Health

Shiqing Ma
New Horizon Health Limited

Mingya Han
New Horizon Health Limited

Shu Zheng
Zhejiang University  https://orcid.org/0000-0002-2521-190X

Weixian Zheng
New Horizon Health Limited

Jun Liu  (jun.liu@nhbio.com.cn)
New Horizon Health Limited

Article

Keywords: 16S rRNA Sequencing, Colorectal Cancer (CRC), Model Classification, CRC Early Detection, Random Forest

DOI: https://doi.org/10.21203/rs.3.rs-763700/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Accurate Five-category Classification for Colorectal Cancer

Using Gut Microbiome 16S rRNA Sequencing

Liying Zhang¹, #, Jiaqi Zhu¹, #, Qutao Ding¹, #, Yanqin Huang², Hongbo Zhang¹, Shiqing Ma¹, Mingya Han¹, Shu Zheng², Weixian Zheng¹, Jun Liu¹ *

¹New Horizon Health Co. Ltd, 13F/T1, 400 Jiang’er Rd., Herui Technology Park, Binjiang District, Hangzhou, Zhejiang 310052, China; ²Cancer Institute (Key Laboratory of Cancer Prevention and Intervention, China National Ministry of Education, Key Laboratory of Molecular Biology in Medical Sciences, Zhejiang Province, China), The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310009, China

# These authors contributed equally to this work.

* To whom correspondence should be addressed.

Email: jun.liu@nhbio.com.cn

Tel: 86-571-81956716

Running title: Classification for Colorectal Cancer Using Gut Microbiome 16S rRNA Sequencing

Keywords: 16S rRNA Sequencing, Colorectal Cancer (CRC), Model Classification, CRC Early Detection, Random Forest
The association between the gut microbiome and the five stages of colorectal cancer (CRC) (healthy, polyposis, nonadvanced adenoma, advanced adenoma, and cancer) remains unclear. We performed 16S rRNA sequencing of the V3-V4 amplicon from 999 samples from subjects at various stages of CRC development and constructed an accurate predictive random forest model for CRC development. In the testing set, our five-category CRC prediction classifier had accuracies of 0.84 and 0.74 using the relative operational taxonomic unit (OTU) abundances and relative genus abundances, respectively. Specifically, the OTU-based classifier had a sensitivity of 0.97 and specificity of 0.97 for CRC samples, and the genus-based classifier had a sensitivity of 0.97 and specificity of 0.95 for CRC samples. Meanwhile, the gut microbiota was found to differ at all stages of CRC development. The differential abundances of closely related bacteria were used to accurately classify the five stages of CRC development. Additionally, both unannotated and annotated OTUs played important roles in classifier modelling. Our work not only provides valuable 16S rRNA sequencing data from patients and healthy individuals on a large scale but also identifies reproducible gut microbiome biomarkers for CRC staging and highlights their potential applications as noninvasive microbiome biomarkers for diagnosis and as predictive CRC screening tests.
INTRODUCTION

Colorectal cancer (CRC) is the 3rd most common cancer in the world and the 2nd most common cancer in China, and it accounts for more than 0.57 million deaths worldwide. Most sporadic CRCs develop through polypoid adenomas to an intramucosal carcinoma (high-grade dysplastic adenoma), which then progresses to malignant forms over 5-10 years. This is known as the adenoma-carcinoma process, which is induced by a multistep mechanism that involves intestinal flora, metabolism, inflammation, gene mutations, etc. Because the cancerization process can last for decades, the detection of precancerous-stage adenoma or early-stage CRC could bring the 5-year relative survival rate to up to 90%, alleviate the mortality rate, and reduce the economic burden.

Previous investigations have shown that the intestinal flora is different between healthy individuals and patients with diseases, including CRC. Recent studies have shown that from nodular polyps to early cancer to metastatic disease, the intestinal flora changes at all stages of CRC development. With in-depth research advances, increasing evidence shows that the intestinal flora and its related metabolites are closely related to the occurrence of malignant tumours through the development of chronic inflammation and immune surveillance disorders. Changes in the type and number of intestinal flora constituents significantly affect the occurrence and development of CRC. For example, certain strains of Escherichia coli and Bacteroides fragilis have a significant promotion effect on CRC. Nevertheless, the relative abundance of any one kind of microorganism may not be able to distinguish between healthy individuals and CRC patients. Alternatively, metagenomic profiling of faecal microbiomes was used to discover microbial biomarkers between CRC patients and healthy individuals. The
validation result achieved an area under the curve (AUC) of 0.84 in the Chinese cohort. A combined faecal immunochemical test and 16S rRNA gene sequencing method was used to diagnose CRC and adenoma, with AUCs reaching 0.83 and 0.95, respectively. These results demonstrate that detecting biomarkers of faecal microbiomes is a promising non-invasive strategy for diagnosing colorectal diseases, but challenges remain. However, most previous studies using microbial genetic biomarkers are limited by small sample sizes and may not be able to identify predictive signals from hundreds to thousands of features. Therefore, prospective studies are urgently needed for large-scale stool specimen collection with long-term follow-up to validate the influence of microorganisms contributing to CRC development and determine features of the intestinal microbiome that can be used for CRC screening and modified for clinical prevention or treatment.

In this work, we analysed the microbiota of 999 stool specimens from patients with CRC and adenoma and normal controls and solved the problem of insufficient sample size. Similar to existing studies²⁴⁻²⁸, 16S rRNA gene (V3-V4 region) sequencing was used to investigate the microbiota content, although it may not be possible to achieve species-level differentiation. We formalized the task of CRC or adenoma prediction as a classification problem and focused on operational classification units (OTUs). By using OTUs instead of taxonomic annotations as features, the problem that the same OTU may be inconsistently annotated when using different reference databases was avoided²⁹,³⁰. Unlike the multibacteria model, all OTUs that passed the quality control standards were used as features instead of selecting the most distinguishing OTUs. Using the method of standardizing the sample to reduce batch impact, the subjects could be further divided
into CRC, advanced adenoma (AA), nonadvanced adenoma (NAA), polyposis (PL) and healthy individual (HI) groups using the random forest model, providing increased diagnostic value for potential clinical applications. The proposed strategy is a promising method for disease stratification.
RESULTS

Alterations in gut microbial composition in serial CRC stages

To identify reproducible relationships between the gut microbiome and developmental stages of CRC, we performed 16S rRNA gene amplicon sequencing of the stool microbiome of 999 patients with CRC or healthy individuals recruited. In total, 200 samples were collected from CRC patients, 195 from subjects with AA, 197 from patients with NAA, and 200 patients with PL and 207 HIs.

In this study, the association of microbiome abundance with CRC progression was characterized. The abundance of genera in samples from various stages of CRC showed a stark difference (Extended Data Fig. 1a). The two most abundant genera, *Bacteroides* and *Prevotella*, had a complementary relationship, i.e., the higher the abundance of *Bacteroides* in the sample was, the lower the abundance of *Prevotella*. At the same time, in the development of CRC, the abundance of *Bacteroides* increased, and the abundance of *Prevotella* decreased. Intriguingly, the abundances of *Faecalibacterium*, *Roseburia* and *Lachnospiraceae incertae sedis* in the samples were also high, but significant differences were not observed among the HI, PL, AA, or CRC samples. According to the distribution of each bacterial genus in the five groups of samples (Fig. 1a), the abundance of *Bacteroides* and *Prevotella* in the CRC samples was much higher and lower than that in the other groups, respectively. Meanwhile, *Faecalibacterium* had a lower proportion in the NAA stage, and *Roseburia* had a lower proportion in the CRC samples.

To identify significantly different components, dimensionality reduction analysis on 561-dimensional data features was performed using principal component analysis (PCA)\(^{31}\). A two-sided Wilcoxon test was performed to test the differences between the first and
second principal components (Fig. 1b). The CRC samples were significantly lower than in other groups of samples in the first principal component (upper boxplots). According to the second principal component test (left boxplots), the most significant difference was found between the NAA and AA groups and between the NAA and HI samples (P < 0.001). The CRC and AA, NAA, and HI samples were significantly different, with a P value < 0.05. Meanwhile, significant differences between the NAA and PL samples and between the HI and PL samples were also found. These results indicated that the gut microbiotas of each stage of CRC progression share similar bacterial components but could be significantly separated from those of other stages.

To identify stage-specific gut microbial signatures with CRC progression, the sequences of 561 OTUs were annotated and analysed at the phylum (Extended Data Fig. 1b), family, and genus levels. The OTUs were divided into 17 families, and only 16 genera were found to have significant differences among serial CRC stages using a two-sided Mann-Whitney U test (Fig. 2). Among them, *Intestinibacter*, *Romboutsia*, *Blautia*, *Faecalibacterium*, and *Clostridium sensu stricto* showed a state of reduced abundance during CRC development. In contrast, the abundances of *Clostridium XIVb*, *Parabacteroides*, *Odouribacter*, *Bacteroides*, and *Fusobacterium* during the development of CRC continued to increase. Altogether, the abundance of significantly altered genera in the gut microbiome could be used as a signature of CRC staging and potential microbial biomarkers for early detection of CRC.

Assessment of classification models using taxonomic abundances

To test whether the stool microbiome could be used as a reproducible CRC prescreening tool, we performed prediction and validation for various stages of CRC using diverse
classification algorithms. Classification models using gut microbiome 16S rRNA sequencing data were constructed according to the relative OTU abundance and the relative genus abundance. Before data classification tests, a total of 561 OTUs passed feature screening, and 987 samples satisfied the cut-off of variance after applying stringent inclusion and exclusion criteria (see Methods). The samples were randomly shuffled and divided into a training set and a testing set at a ratio of 7:3. Sample composition in the training set included HI (n=136), PL (n=146), NAA (n=149), AA (n=118), and CRC (n=141). In the testing set, there were 59 HI, 54 PL, 48 NAA, 77 AA, and 59 CRC samples.

To test the impact of algorithms on the classification accuracy, the training samples (690) were used for training the classifiers to classify samples into five stages of CRC progression using 11 classification algorithms, including 5 weak classifiers (K-Nearest Neighbours, Support Vector Machines, Logistic Regression, Naïve Bayes, and Decision Tree) and 6 ensemble classifiers (Majority Vote Classifier, Adaboost (Adaptive Boosting) Classifier, Bagging Classifier, Stacking Ensemble Classifier, XGBoost (Extreme gradient-boosted trees), and Random Forest Classifier). As a result, the accuracy of classification in both the training set and testing set was apparently affected by using different algorithms, and the ensemble classifiers generally outperformed the weak classifiers in accuracy using either relative OTU or genus abundances (Table 1).

For weak classifiers using the relative OTU abundances, the classifier using the decision tree algorithm had the highest accuracy (0.616) in the testing set, followed by using naïve Bayes algorithms (0.512). Using the relative genus abundances, the accuracy rank was consistent with the ranking by the relative OTU abundance modelling but with a smaller
accuracy value. Compared with that of the weak classifiers, the classification accuracy of
the ensemble classifiers was considerably improved. The ensemble classifier constructed
using the relative OTU abundances and relative genus abundances with the highest
classification accuracy was the random forest classifier, with testing accuracies of 0.845
and 0.743, respectively. Altogether, the relative abundances using OTUs have a better
classification performance than that of the genera, and the random forest model has the
best classification effect. These findings highlighted that the relative taxonomic
abundances from 16S rRNA gene sequencing data can be accurately associated with
different stages of CRC using the random forest classifier.

Assessment of CRC staging classification using the microbiome random forest
model

In addition to the benchmarking of classification accuracy using various classifiers, we
assessed the prediction ability of our random forest classifier for the 297 samples in the
testing set. The sensitivity (true-positive rate) and specificity (1–false-positive rate) of
the classifier were evaluated using the relative OTU abundances and the relative genus
abundances (Table 2). Using the relative OTU abundances, the sensitivity and specificity
of the classification model for CRC were both 0.97. As demonstrated in the matrix of
confusion (Fig. 3a), the OTU-dependent random forest classifier divides the 59 samples
from CRC patients into 57 CRCs, 1 AA, and 1 NAA. Meanwhile, the 59 healthy control
samples were divided into 55 HIs, 2 PLs, 1 NAA and 1 AA. Together, our trained
random forest classifier performed better in the classification of CRC and healthy stages
with a higher sensitivity and specificity than other stages of CRC development. For
classification using the relative genus abundances (Fig. 3b), the performance of the
classifier was generally worse than that of using the relative OTU abundances (Table 2). The genus-dependent classifier had a sensitivity value of 0.97 with a specificity of 0.95 for CRC samples, but the sensitivity for healthy samples was only 0.71. These results suggested that the trained random forest classifier using the relative genera abundances will have a high false positive rate among healthy samples and thus is impractical for clinical application, especially in CRC early detection. Therefore, these results suggest that a faecal microbiome predictive test using 16S rRNA gene sequencing with relative OTU abundances with our trained random forest classifier could serve as a promising clinical tool for large-scale CRC screening.

Correlations among critical taxonomic biomarkers

Based on the outstanding ability of our random forest classifier to discriminate each stage of CRC progression from the training set to the testing test, we tried to explain the correlation among biomarkers in the OTU- and genus-based classification model. The relative abundances of the top 5 important OTUs and OTU-annotated genera were extracted, and calculations were performed (Fig. 3c, see Extended Data Table 1 for OTUs not in the database). As a result, the top five OTUs were all highly correlated with absolute values greater than 0.2 (Fig. 4f). For instance, OTU511 (Dialister) and OTU325 (Bacteroides) have a positive correlation with a correlation coefficient of 0.41, and OTU153 (Bacteroides) and OTU202 (Prevotella) have a negative correlation with a correlation coefficient of -0.37. These results highlighted the importance of each OTU even from the same genus. However, the top five genera generated from the relative OTU abundances were found to be less correlated (Fig. 4l). For example, Clostridium sensu stricto was found to have a negative correlation between Bacteroides with a correlation
coefficient of -0.23 and a positive correlation between *Prevotella* (0.21). Using the relative OTU abundances, the correlations among *Clostridium sensu stricto*, *Diaslister*, and *Fusobacterium* were all less than 0.1, which demonstrates the weak correlation of genera annotated from OTU data in the random forest classifier. Meanwhile, the top 5 important features from the model using relative genus abundances were also highly correlated ([Fig. 3d, Extended Data Fig. 2](#)). Considering the false positive rate of the genera model, the importance of each genus was perturbed by the neighbouring groups because of the inappropriate taxonomic level.

In addition to the correlation analysis among features in the classification models, the distribution of critical OTUs and genera showed that the relative abundances of important OTUs and genera had diverse distribution patterns between stages of CRC development ([Fig. 4a-e, g-k](#)). For example, any one of the 5 OTUs had a significant difference between CRC and other stages (P < 0.0001, Wilcoxon test), and both OTU325 ([Fig. 4a](#)) and OTU153 ([Fig. 4c](#)) could effectively distinguish AA from other stages (P < 0.001, Wilcoxon test). In the distribution of genus abundances, the abundances of *Bacteroides* ([Fig. 4k](#)), *Prevotella* ([Fig. 4h](#)), *Clostridium sensu stricto* ([Fig. 4i](#)), and *Dialister* ([Fig. 4g](#)) in the CRC samples were significantly different from those in the other samples, while the abundances of *Fusobacterium* were significantly different only between CRC and NAA or PL ([Fig. 4j](#)). According to the distribution divergence of OTUs and genera, these results supported that the random forest classifier using the significantly different distributed relative OTU abundances could have a better classification accuracy.
Microbiome functional changes during CRC progression

To evaluate the correlation of signalling pathways between normal samples and abnormal samples, the functional potentials of the microbiome were inferred using PICRUSt2 \(^{32}\). A total of 29 signalling pathways were found to be highly correlated with different statuses of colorectal diseases (Extended Data Fig. 3). The correlation coefficient of OTUs and signalling pathways exhibited the major differences among colorectal diseases (Spearman correlation coefficient). For instance, compared with the normal samples, the signalling pathways of the ethylmalonyl-CoA pathway (MetaCyc: PWY-5741) and super pathway of (R,R)-butanediol biosynthesis (MetaCyc: P125-PWY) in OTU192 were found to be highly correlated with the PL group samples. Meanwhile, the pathway of chorismate biosynthesis II (MetaCyc: PWY-6165) in Clostridium cluster XIVa was found to be tightly linked to the PL and CRC groups but was less correlated in the NAA and AA stages. Moreover, the isopropanol biosynthesis pathway (MetaCyc: PWY-6876) in Clostridium cluster I was highly correlated with the PL, NAA, and NA groups but less linked to the CRC group. These results showed that the predicted functional abundances of the microbiome using 16S rRNA gene sequencing data from faecal samples were highly consistent with the KEGG orthologue for microbial metabolism from subjects at different stages of colorectal diseases (Extended Data Table 2).
DISCUSSION

Using 999 samples collected with detailed colonoscopy and histopathological data, this study assessed alterations in the gut microbiome during the development of CRC and the capability of microbial markers to detect CRC at various stages before cancer. The best performing model achieved a high accuracy (CRC, sensitivity=0.97, specificity=0.97) with 54 important OTUs to distinguish CRC from a nontumour control. Meanwhile, a strong correlation between modelling OTUs was found, and 29 signalling pathways were tightly correlated with CRC stages. The differentially abundant OTUs and signalling pathways were constantly altered at different stages, and these signatures have potential as biomarkers for determining the CRC development stage without colonoscopy and early detection of CRC among populations.

It has long been reported that faecal bacteria, such as *F. nucleatum, Escherichia coli*, and *Bacteroides fragilis*, could serve as biomarkers for the noninvasive diagnosis of CRC\(^{19,23,33,34}\). Our research also shows that *Bacteroides* and *Prevotella* are dominant in the human gut microbiota. With the occurrence of CRC, the relative abundance of *Bacteroides* is increased with decreased *Prevotella*, indicating that the depletion of *Prevotella* may induce the development of CRC. Meanwhile, increased *Bacteroides* abundances in samples suggest the occurrence of CRC and development. In addition, at the genus level, 16 genera had significant differences in relative abundances in the CRC, AA, NAA, PL and HI groups.

In a previous study, adenoma-related classifiers showed lower accuracies in distinguishing adenomas from healthy controls (AUC =0.54) or CRC (AUC =0.69)\(^{35}\). One explanation is that the adenoma-associated gut microbiome closely resembles that
associated with a healthy status\textsuperscript{35-37}. In addition, it is probably also influenced by the limited coverage of taxonomy and the high dependence on reference genomes in whole-metagenome shotgun sequencing (WMS) taxonomic profiling\textsuperscript{38,39}. WMS data are well recognized to possess the advantage of species- and strain-level resolution. However, the current strategies for characterizing microbial community compositions with WMS strongly rely on the known reference genome database\textsuperscript{40-42}, which is likely missing species without well-annotated genomes. Consequently, WMS data analysis has biases in the relative abundance normalization resulting from unknown bacteria. To overcome the absent species annotation, we performed 16S rRNA sequencing to profile the OTU level and genus-level abundance of gut bacteria. Moreover, considering inconsistent abundance changes among OTUs assigned as the same species\textsuperscript{29,30}, we constructed classifiers using the relative OTUs and genera to capture the most informative taxonomies that could effectively distinguish CRC patients from healthy people.

In previous studies, researchers mainly constructed a two-class model that distinguishes CRC/adenoma and healthy controls\textsuperscript{35-37}. This study included all the stages of CRC development ranging from HI, PI, NAA, AA to CRC and constructed a five-category classification model using 11 algorithms. The classifiers using ensemble algorithms are superior to the weak classifiers. The relative OTU abundances achieved the best classification performance on the random forest model with accuracies above 0.8. Furthermore, using OTUs as a feature gave better results than using genera as a feature for classification. This probably resulted from the interference of the addition of OTUs without classification value into relative genus abundances (Extended Data Table 3). Together, it is recommended to use the relative OTU abundances as features when using
16S rRNA sequencing data to construct a classification model. Since more valuable classification factors among categories can be retained using the relative OTU abundances, the classification model can be more accurate than using a genera model. Overall, the testing results of the 297 samples exhibit the robustness of our trained five-category random forest classifier.

In addition to evaluating the performance of our OTU-based classification model, we tried to explain the correlation of critical features that contribute to the model. Using correlation analysis, we found that only 54 OTUs played an important role in the classification model, with 8 important OTUs not annotated in the database, which is worth further exploration of their biological functions in future research. Meanwhile, a total of 29 signalling pathways were found to be highly correlated with different stages of colorectal diseases, and 10 signalling pathways were consistent with those reported in the literature. Most of the other signalling pathways were closely related to the occurrence of enteritis, and the pathogenic mechanism of these signalling pathways is worth further exploring to refine their contribution to CRC progression. Nevertheless, this study identifies reproducible gut microbiome biomarkers for CRC staging and highlights their potential applications as noninvasive microbiome biomarkers for diagnosis and as predictive CRC screening tests.
MATERIALS AND METHODS

Faecal sample collection and storage

Faecal samples were collected from the Second Affiliated Hospital, Zhejiang University School of Medicine and the Cancer Prevention and Treatment Institute of Jiashan County in Zhejiang Province using a faecal collection and pretreatment kit (New Horizon Health Technology Co., Ltd., China). The inclusion criteria were as follows: age 40 to 75; available colonoscopy and pathological results; no history of any other cancers before the collection of faecal samples; and no preoperative treatment before the collection of faecal samples. For each sample, approximately 5 g of stool specimen was collected and stored in the collection tube with storage buffer. For long-term storage, faecal samples were stored at -80°C before DNA extraction. All patients signed the informed consent form of the study. The use of human samples was approved by The Second Affiliated Hospital, Zhejiang University School of Medicine.

Sample grouping

Samples were divided into five groups, HI, PL, NAA, AA, and CRC, according to the following histopathological criteria: HI was defined as normal without hyperplasia or neoplastic findings under colonoscopy; PL was defined as non-adenomatous hyperplasia; NAA was defined as having adenoma with a diameter <10 mm; AA was defined as adenoma with high-grade dysplasia, adenoma with a diameter ≥10 mm, villous differentiation in ≥25% of the visible area under a microscope, or a serrated lesion with a diameter ≥10 mm; and CRC was defined as all stages of CRC (I-IV).
Library preparation and sequencing

Total genomic DNA from faecal samples was extracted and purified using nucleic acid extraction and purification kits (New Horizon Health Technology Co., Ltd., China). The DNA concentration was measured using a NanoDrop2000 (Thermo), and the quality was assessed on a 1% agarose gel (1%, w/v). Each DNA sample was diluted to 1 ng/μL with sterile water before PCR amplification. The V3 and V4 hypervariable regions of the 16S rRNA gene were amplified using the primer pair 341F (5’-CCTAYGGGRBGCASCAG-3’) and 806R (5’-GGACTACNNGGGTATCTAAT-3’) as described. PCR was performed using Phusion High Fidelity PCR Master Mix (NEB, cat. M0531 L) with thermal cycling conditions consisting of 98°C for 1 min, followed by 30 cycles of 98°C for 10 sec, 50°C for 30 sec, and 72°C for 30 sec, and finally 72°C for 5 min. PCR products were separated by electrophoresis in agarose gels (2%, w/v), and samples with bright main bands between 400-500 bp were selected and purified using a GeneJET Gel Extraction Kit (Thermo, cat. K0692). Sequencing libraries were prepared using the TruSeq DNA PCR-Free Sample Preparation Kit (Illumina, Cat. FC-121-3003) following the manufacturer’s recommendations. Library quality was assessed using an Agilent Bioanalyzer 2100 system. The pooled libraries were sequenced on an Illumina HiSeq2500 at 250 bp paired-end (PE) read lengths in three batches by Novogene Bioinformatics Technology Co., Ltd. (Beijing, China).

Analysis pipeline

The analysis pipeline consisted of a combination of public programs and customized scripts. PE reads from each sample were first merged using FLASH (v1.2.10) and quality filtered using USEARCH (v10.0.240) with a minimal 10 bp overlap and Q20
cut-offs, respectively. Then, the reads were input into the UPARSE program for chimaeric read filtering and cluster (OTU) generation with 99% similarity. Next, the abundances of the OTUs in each sample were stored in a matrix, with each row denoting an OTU. For each sample, the total abundances of the OTUs were normalized to 50,000. Moreover, the OTUs were classified and annotated by SINTAX using the taxonomic information in the RDP and SILVA databases. Finally, the relative OTU abundances were compared among groups using linear discriminant analysis effect size (LEfSe).

**Data preprocessing**

Data preprocessing was performed on OTU features and samples. In feature screening, OTUs with normalized abundances less than 10 in more than 90% of the samples were excluded from further analysis. Then, the sum of the remaining OTU abundances in each sample was normalized to 100 for standardization. As a result, the OTU abundances of each sample were standardized, and the sequencing data were converted into characteristic ratio data. Samples with a standard deviation of the relative OTU abundances greater than 3 were treated as abnormal samples and excluded from the following analysis. To solve the problem of interbatch differences between samples in the sequencing process, we normalized the samples. Specifically, under each OTU sequence, the relative abundance ratio of each sample corresponding to the OTU was calculated and used as the characteristic expression value of the model. That is, in the OTU relative abundance matrix, the samples are normalized to eliminate batch-to-batch problems during the sequencing process.
Classifier development

All classifier developments were performed using Python (v.3.8.5), and some statistical analyses were performed using R (v.4.0.2). K-Nearest Neighbours, Decision Tree, Logistic Regression, Support Vector Machines, and Naïve Bayes were generated using the scikit-learn package. For ensemble classifiers, the majority vote classifier was integrated using logistic regression, decision tree, and K-nearest neighbours with default parameters, except for the voting parameter. The voting parameter in this class is set to soft, which predicts the class label based on the argmax of the sums of the predicted probabilities. The bagging classifier uses a decision tree as the basic classifier. Entropy is used in the base classifier to evaluate the information content of each feature. In the bagging classifier, the number of base estimators is set to 10, and the training for each base estimator draws 50% of the sample each time. The AdaBoost classifier uses a decision tree as the basic classifier. The maximum number of estimators to terminate boosting is set to 2. The learning rate was set to 0.01. The stacking ensemble classifier integrated K-nearest neighbours, random forest, and naïve Bayes as the first-level classifiers and logistic regression as the second-level classifier. The XGBoost classifier was set using a step size of 0.1 when the weight was updated in each iteration. The total number of iterations was 10, the depth of the tree was 5, and the minimum loss function reduction value required for node splitting was set to 0. When training each tree, the ratio of the used data to the total training set was 0.8, and the ratio of the used features to the total features was 0.8. The objective function was set to logistic, and the weight of the positive sample was 1. While training the random forest model, the class of sklearn.model_selection.GridSearchCV was applied to exhaustively search over specified
parameter values for the estimator. The number of trees in the forest was set to 40, and
the number of features to consider when looking for the best split was set to 7.

Data availability

All sequence data were submitted to the European Nucleotide Archive (ENA) under
project PRJEB29239.

Computer code

The codes are available at https://github.com/liyingya-123/NHBio_16S_Analysis.git. The
customized code was written in Python 3.8.5 and R 4.0.2.
REFERENCES

1. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249, doi:10.3322/caac.21660 (2021).

2. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767, doi:10.1016/0092-8674(90)90186-i (1990).

3. Tajik, N. et al. Targeting zonulin and intestinal epithelial barrier function to prevent onset of arthritis. *Nat Commun* **11**, 1995, doi:10.1038/s41467-020-15831-7 (2020).

4. Ge, Y. et al. Gut microbiota influence tumor development and Alter interactions with the human immune system. *J Exp Clin Cancer Res* **40**, 42, doi:10.1186/s13046-021-01845-6 (2021).

5. Schott, E. M. et al. Targeting the gut microbiome to treat the osteoarthritis of obesity. *JCI Insight* **3**, doi:10.1172/jci.insight.95997 (2018).

6. Chen, X. et al. Gut dysbiosis induces the development of pre-eclampsia through bacterial translocation. *Gut* **69**, 513-522, doi:10.1136/gutjnl-2019-319101 (2020).

7. Chattopadhayay, I. et al. Exploring the Role of Gut Microbiome in Colon Cancer. *Appl Biochem Biotechnol*, doi:10.1007/s12010-021-03498-9 (2021).

8. Ashktorab, H., Kupfer, S. S., Brim, H. & Carethers, J. M. Racial Disparity in Gastrointestinal Cancer Risk. *Gastroenterology* **153**, 910-923, doi:10.1053/j.gastro.2017.08.018 (2017).

9. Shen, T. D. Diet and Gut Microbiota in Health and Disease. *Nestle Nutr Inst Workshop Ser* **88**, 117-126, doi:10.1159/000455220 (2017).

10. Cheng, Y., Ling, Z. & Li, L. The Intestinal Microbiota and Colorectal Cancer. *Front Immunol* **11**, 615056, doi:10.3389/fimmu.2020.615056 (2020).

11. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J Clin* **70**, 7-30, doi:10.3322/caac.21590 (2020).

12. Yachida, S. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med* **25**, 968-976, doi:10.1038/s41591-019-0458-7 (2019).

13. Qi, X. et al. Gut microbiota mediated molecular events and therapy in liver diseases. *World J Gastroenterol* **26**, 7603-7618, doi:10.3748/wjg.v26.i48.7603 (2020).

14. Sheng, Q. et al. Characteristics of fecal gut microbiota in patients with colorectal cancer at different stages and different sites. *Oncol Lett* **18**, 4834-4844, doi:10.3892/ol.2019.10841 (2019).

15. Ahmed, I. & Umar, S. Microbiome and Colorectal Cancer. *Curr Colorectal Cancer Rep* **14**, 217-225, doi:10.1007/s11888-018-0416-7 (2018).

16. Chen, C. et al. Bacteroides, butyric acid and t10,c12-CLA changes in colorectal adenomatous polyp patients. *Gut Pathog* **13**, 1, doi:10.1186/s13099-020-00395-0 (2021).

17. Cuevas-Ramos, G. et al. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc Natl Acad Sci USA* **107**, 11537-11542, doi:10.1073/pnas.1001261107 (2010).
Goodwin, A. C. et al. Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis. Proc Natl Acad Sci U S A 108, 15354-15359, doi:10.1073/pnas.1010203108 (2011).

Wu, S. et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. Nat Med 15, 1016-1022, doi:10.1038/nm.2015 (2009).

Arthur, J. C. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science 338, 120-123, doi:10.1126/science.1224820 (2012).

Huycke, M. M., Abrams, V. & Moore, D. R. Enterococcus faecalis produces extracellular superoxide and hydrogen peroxide that damages colonic epithelial cell DNA. Carcinogenesis 23, 529-536, doi:10.1093/carcin/23.3.529 (2002).

Eklof, V. et al. Cancer-associated fecal microbial markers in colorectal cancer detection. Int J Cancer 141, 2528-2536, doi:10.1002/ijc.31011 (2017).

Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut 66, 70-78, doi:10.1136/gutjnl-2015-309800 (2017).

Cao, Q. et al. Effects of Rare Microbiome Taxa Filtering on Statistical Analysis. Front Microbiol 11, 607325, doi:10.3389/fmicb.2020.607325 (2020).

Jin, Y. et al. Prediction of Postoperative Ileus in Patients With Colorectal Cancer by Preoperative Gut Microbiota. Front Oncol 10, 526009, doi:10.3389/fonc.2020.526009 (2020).

Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457-463, doi:10.1038/nature24621 (2017).

Baxter, N. T., Ruffin, M. T. t., Rogers, M. A. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med 8, 37, doi:10.1186/s13073-016-0290-3 (2016).

Hirano, A. et al. Comparison of the microbial community structure between inflamed and non-inflamed sites in patients with ulcerative colitis. J Gastroenterol Hepatol, doi:10.1111/jgh.14129 (2018).

Balvociute, M. & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics 18, 114, doi:10.1186/s12864-017-3501-4 (2017).

Gwak, H. J. & Rho, M. Data-Driven Modeling for Species-Level Taxonomic Assignment From 16S rRNA: Application to Human Microbiomes. Front Microbiol 11, 570825, doi:10.3389/fmicb.2020.570825 (2020).

Dien, J. Applying principal components analysis to event-related potentials: a tutorial. Dev Neuropsychol 37, 497-517, doi:10.1080/87565641.2012.697503 (2012).

Douglas, G. M. et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 38, 685-688, doi:10.1038/s41587-020-0548-6 (2020).

Kostic, A. D. et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res 22, 292-298, doi:10.1101/gr.126573.111 (2012).

Gao, R. et al. Dysbiosis signature of mycobiota in colon polyp and colorectal cancer. Eur J Clin Microbiol Infect Dis 36, 2457-2468, doi:10.1007/s10096-017-3085-6 (2017).
Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 25, 667-678, doi:10.1038/s41591-019-0405-7 (2019).

Feng, Q. et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 6, 6528, doi:10.1038/ncomms7528 (2015).

Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 10, 766, doi:10.15252/msb.20145645 (2014).

Rausch, P. et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome 7, 133, doi:10.1186/s40168-019-0743-1 (2019).

Feng, Q. et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun 6, 667-678, doi:10.1038/ncomms7528 (2015).

Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 10, 766, doi:10.15252/msb.20145645 (2014).

Rausch, P. et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome 7, 133, doi:10.1186/s40168-019-0743-1 (2019).

Laudadio, I., Fulci, V., Stronati, L. & Carissimi, C. Next-Generation Metagenomics: Methodological Challenges and Opportunities. OMICS 23, 327-333, doi:10.1089/omi.2019.0073 (2019).

Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9, 811-814, doi:10.1038/nmeth.2066 (2012).

Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 10, 1014, doi:10.1038/s41467-019-08844-5 (2019).

Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15, R46, doi:10.1186/gb-2014-15-3-r46 (2014).

Yu, Y., Lee, C., Kim, J. & Hwang, S. Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. Biotechnol Bioeng 89, 670-679, doi:10.1002/bit.20347 (2005).

Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).

Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10, 996-998, doi:10.1038/nmeth.2604 (2013).

Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41, D590-596, doi:10.1093/nar/gks1219 (2013).

Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73, 5261-5267, doi:10.1128/AEM.00062-07 (2007).

Yilmaz, P. et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res 42, D643-648, doi:10.1093/nar/gkt1209 (2014).

Segata, N. et al. Metagenomic biomarker discovery and explanation. Genome Biol 12, R60, doi:10.1186/gb-2011-12-6-r60 (2011).

Pedregosa., F., Varoquaux., G., Gramfort., A., Michel., V. & Thirion., B. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830 (2011).
Acknowledgements

We thank Ms. Weixian Zheng and Ms. Jiying Wang for sample collection. The Second Affiliated Hospital and National Key R&D Program of China (No. 2016YFC1302803). Zhejiang University School of Medicine and New Horizon Health Co., Ltd. supported this work.

Author contributions

J. L., L. Z., S. Z., and Y. H. conceived the study. J. L and Y. H. supervised the study. W. Z., J. W., S. Z., and M. H. organized the clinical study, recruited patients, and collected samples. J. Z. and L. Z. and S. M. and Y. H. collected and analysed the 16S rRNA sequencing data. L. Z. and H. Z. and J. Z. and Y. H. performed machine learning and statistical analyses. L. Z. and Q. D. designed and generated the figures. L. Z. and Q. D. wrote the manuscript. All authors discussed and approved the manuscript.

Competing interests

L. Z., J. Z., H. Z., Q. D., S. M., M. H., and J. L. are employees of New Horizon Health Technology Co., Ltd.

Ethics approval

2017-075 of The Second Affiliated Hospital, Zhejiang University School of Medicine, China.
### Table 1. Accuracy evaluation of the classifier using different algorithms

| Algorithms                  | Training accuracy (Average ± SD) | Testing accuracy |
|-----------------------------|----------------------------------|------------------|
|                             | OTU     | Genus   | OTU     | Genus   |
| **Weak classifier**         |         |         |         |         |
| K-Nearest Neighbours        | 0.39 ± 0.04 | 0.33 ± 0.04 | 0.367 | 0.326 |
| Support Vector Machine      | 0.43 ± 0.04 | 0.37 ± 0.02 | 0.364 | 0.354 |
| Logistic Regression         | 0.47 ± 0.02 | 0.38 ± 0.03 | 0.421 | 0.392 |
| Naïve Bayes                 | 0.53 ± 0.04 | 0.48 ± 0.06 | 0.512 | 0.5   |
| Decision Tree               | 0.59 ± 0.01 | 0.54 ± 0.03 | 0.616 | 0.514 |
| **Ensemble classifier**     |         |         |         |         |
| Majority Vote Classifier    | 0.62 ± 0.02 | 0.53 ± 0.02 | 0.606 | 0.552 |
| Adaboost Classifier         | 0.63 ± 0.06 | 0.52 ± 0.03 | 0.646 | 0.559 |
| Bagging Classifier          | 0.67 ± 0.03 | 0.60 ± 0.04 | 0.754 | 0.601 |
| Stacking Ensemble Classifier| 0.76 ± 0.01 | 0.56 ± 0.04 | 0.747 | 0.615 |
| XGBoost Classifier          | 0.84 ± 0.05 | 0.67 ± 0.08 | 0.818 | 0.681 |
| Random Forest               | 0.85 ± 0.04 | 0.74 ± 0.08 | 0.845 | 0.743 |

a Standard deviation.
| Group     | Sensitivity | Specificity |
|-----------|-------------|-------------|
| **CRC**   | 0.97        | 0.97        |
| **AA**    | 0.64        | 0.97        |
| **NAA**   | 0.92        | 0.92        |
| **PL**    | 0.80        | 0.98        |
| **HI**    | 0.93        | 0.99        |
| **Relative OTU abundances** |          |             |
| **CRC**   | 0.97        | 0.95        |
| **AA**    | 0.62        | 0.96        |
| **NAA**   | 0.75        | 0.88        |
| **PL**    | 0.67        | 0.94        |
| **HI**    | 0.71        | 0.95        |
| **Relative genus abundances** |          |             |
Figure 1. Alterations in gut microbial composition in serial CRC stages. **a** Relative proportion of bacterial genus in the HI, PL, NAA, AA and CRC samples. **b** Principal component analysis of samples (HI, n=195; PL, n=200; NAA, n=197; AA, n=195; CRC, n=200) from five groups, which shows that the faecal microbiota composition was divergent in serial CRC stages. The groups (HI, PL, NAA, AA and CRC) are indicated by different shapes and colors. The upper-right and the bottom-left boxplots illustrate the samples projected onto the two principal coordinates broken down by disease status, respectively. P values of the first and second principal components were calculated with a two-sided Wilcoxon test for study and group. “****” indicates P ≤ 0.001, “***” indicates P ≤ 0.01, and “**” indicates P ≤ 0.05. Boxplots represent the 25th–75th percentile of the distribution; the median is shown in a thick line at the middle of the box; the whiskers extend up to values within 1.5 times of IQR, and outliers are represented as circles.
Figure 2. Distinct stage-specific gut microbial signatures with CRC progression.

The phylogenetic tree of 561 OTUs was divided into 17 families. The genus nodes of different gates are marked with different colors, and the outmost purple squares indicate nodes has a relative abundance greater than 0.005. A two-sided Mann–Whitney U test revealed 16 genera with a significance level of less than 0.05 in serial CRC stages, and the box plot shows the distribution of these genera. The y axis for each box plot represents the relative abundance. Each plot shows boxes for the five groups (HI, PL, NAA, AA and CRC) in order from left to right.
Figure 3. Classification results of random forest classifiers and corresponding importance of biomarkers. **a, b** The confusion matrix of the random forest classifiers using the relative OTU abundances (a) and relative genus abundances (b), respectively. The horizontal axis represents the result of the model prediction, and the vertical axis represents the true value of the sample. The proportion in the matrix is the prediction performance in each group. **c, d** The importance of features in the random forest classification model using the OTU (c) and genus (d), respectively. The horizontal axis represents the contribution ratio, and the vertical axis represents the features.
Figure 4. Comparison of distinguishing ability between OTU markers and corresponding genera. a-e Distribution of relative OTU abundances at different stages
of colorectal cancer development. Spearman correlation coefficient between the top 5 OTUs (f) and genera based on the classified OTUs (l). The numbers indicate the Spearman correlation coefficients of the corresponding OTUs and genera, respectively.

distribution of the relative abundances of genus, based on the classified OTUs, at different stages of colorectal cancer development. The groups are color-coded and indicated by different shapes. P values were calculated with a two-sided Wilcoxon test for study and group. “***” indicates P≤0.001, “**” indicates P≤0.01, and “*” indicates P≤0.05. All boxplots represent the 25th–75th percentile of the distribution; the median is shown in a thick line at the middle of the box; the whiskers extend up to values within 1.5 times of IQR, and outliers are represented as dots.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf