frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 25 July 2023
DOI 10.3389/fmicb.2023.1178744

OPEN ACCESS

EDITED BY
Benoit St-Pierre,
South Dakota State University, United States

REVIEWED BY
Liqiang Li,
National Clinical Research Center for Infectious Diseases, China
Xianzhi Lin,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Jiayin Wang
✉ wangjiayin@xjtu.edu.cn

†These authors have contributed equally to this work

RECEIVED 03 March 2023
ACCEPTED 14 June 2023
PUBLISHED 25 July 2023

CITATION
Liu G, Li T, Zhu X, Zhang X and Wang J (2023) An independent evaluation in a CRC patient cohort of microbiome 16S rRNA sequence analysis methods: OTU clustering, DADA2, and Deblur. *Front. Microbiol.* 14:1178744. doi: 10.3389/fmicb.2023.1178744

An independent evaluation in a CRC patient cohort of microbiome 16S rRNA sequence analysis methods: OTU clustering, DADA2, and Deblur

Guang Liu[1,2†], Tong Li[3†], Xiaoyan Zhu[1], Xuanping Zhang[1] and Jiayin Wang[1*]

[1]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, [2]Guangdong Hongyuan Pukong Medical Technology Co., Ltd., Guangzhou, China, [3]School of Bioscience and Bioengineering, South China University of Technology, Guangzhou, China

16S rRNA is the universal gene of microbes, and it is often used as a target gene to obtain profiles of microbial communities via next-generation sequencing (NGS) technology. Traditionally, sequences are clustered into operational taxonomic units (OTUs) at a 97% threshold based on the taxonomic standard using 16S rRNA, and methods for the reduction of sequencing errors are bypassed, which may lead to false classification units. Several denoising algorithms have been published to solve this problem, such as DADA2 and Deblur, which can correct sequencing errors at single-nucleotide resolution by generating amplicon sequence variants (ASVs). As high-resolution ASVs are becoming more popular than OTUs and only one analysis method is usually selected in a particular study, there is a need for a thorough comparison of OTU clustering and denoising pipelines. In this study, three of the most widely used 16S rRNA methods (two denoising algorithms, DADA2 and Deblur, along with *de novo* OTU clustering) were thoroughly compared using 16S rRNA amplification sequencing data generated from 358 clinical stool samples from the Colorectal Cancer (CRC) Screening Cohort. Our findings indicated that all approaches led to similar taxonomic profiles (with $P > 0.05$ in PERMNAOVA and $P < 0.001$ in the Mantel test), although the number of ASVs/OTUs and the alpha-diversity indices varied considerably. Despite considerable differences in disease-related markers identified, disease-related analysis showed that all methods could result in similar conclusions. *Fusobacterium*, *Streptococcus*, *Peptostreptococcus*, *Parvimonas*, *Gemella*, and *Haemophilus* were identified by all three methods as enriched in the CRC group, while *Roseburia*, *Faecalibacterium, Butyricicoccus*, and *Blautia* were identified by all three methods as enriched in the healthy group. In addition, disease-diagnostic models generated using machine learning algorithms based on the data from these different methods all achieved good diagnostic efficiency (AUC: 0.87–0.89), with the model based on DADA2 producing the highest AUC (0.8944 and 0.8907 in the training set and test set, respectively). However, there was no significant difference in performance between the models ($P > 0.05$). In conclusion, this study demonstrates that DADA2, Deblur, and *de novo* OTU clustering display similar power levels in taxa assignment and can produce similar conclusions in the case of the CRC cohort.

KEYWORDS

CRC, gut microbiome, denoising algorithms, comparison, DADA2, Deblur, OTU clustering

Introduction

Colorectal cancer (CRC) is the third most common cancer globally, causing more than one million deaths annually (Brenner et al., 2014; Stoffel and Murphy, 2020). The occurrence and development of colorectal cancer often involve an adenoma-carcinoma process, which often occurs over many years and involves a variety of mechanisms and gene mutations (Fearon and Vogelstein, 1990; Jones et al., 2008). The early symptoms of CRC are not obvious: they consist of body discomfort, dyspepsia, occult blood in the stool, and other symptoms (Brenner et al., 2014). As the disease progresses, clearer symptoms gradually appear, including changes in defecation habits, blood in the stool, diarrhea, alternating diarrhea and constipation, and local abdominal pain, among others (Brenner et al., 2014). Not only does CRC inflict mental and physical distress on patients, but it also imposes a heavy economic burden and places pressure on patients and their families. Developed countries have now made significant advancements in CRC screening, resulting in a gradual decline in incidence (Brenner et al., 2014). Therefore, early cancer screening and timely intervention are important for CRC patients (Díaz-Tasende, 2018; Shaukat et al., 2021; Xi and Xu, 2021).

An increasing number of studies have shown a strong correlation between health and the intestinal microbiota; the microbiota can influence human health by regulating host immunity, inflammation, and cognitive function (Ashktorab et al., 2017; Zhang et al., 2017; Dalal et al., 2021). Studies have confirmed that the intestinal microbiota is a key environmental factor in the occurrence and development of CRC, as the composition of the intestinal microbiota is significantly different in CRC compared to that of healthy people (Zhang et al., 2017; Park et al., 2021). For example, an increase in *Fusobacterium nucleatum* has been confirmed to be closely related to CRC (Castellarin et al., 2012; Bullman et al., 2017). Therefore, fecal metagenomics may play a beneficial role in early screening and clinical diagnosis of colorectal cancer.

16S rRNA has a unique structure that contains both conserved and variable regions, and it is present in all known bacteria and archaea, so it is commonly used as a marker gene for bacterial community research using next-generation sequencing (NGS) technology (Coenye and Vandamme, 2003; Sanschagrin and Yergeau, 2014; Yarza et al., 2014; Muthappa et al., 2022). Not only can the 16S rRNA sequencing approach decrease the high cost of metagenomic sequencing, but it can also mitigate the problem of host contamination (Boers et al., 2019). However, sequencing errors can also introduce some non-real nucleotide differences (Kunin et al., 2010; Aird et al., 2011; Schloss et al., 2011). Traditionally, sequences are clustered into operational taxonomic units (OTUs) with a particular identity threshold (usually 97%) to reduce the interference of sequencing errors using the OTU clustering method (Edgar, 2013; Patin et al., 2013), and *de novo* clustering methods have been regarded as the optimal method of assigning the 16S rRNA gene to OTUs (Westcott and Schloss, 2015). In recent years, several denoising algorithms have been created to solve this problem, such as DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017), which can correct sequencing errors by generating amplicon sequence variants (ASVs). DADA2 has been reported to be more accurate than the OTU clustering method in mock communities, as it can accurately resolve sequence variants differing by a single nucleotide and present in as few as two reads, identify more real variants, and output fewer spurious sequences; this provides alternative methods to explore strain-level variation (Callahan et al., 2016). In another study, it has been reported that both Deblur and DADA2 have high consistency, and both of them achieve outputs close to the ground truth in simulated communities (Amir et al., 2017). Deblur also exhibited higher performance stability than DADA2, with a higher frequency cutoff, when samples from the American Gut Project (http://americangut.org) underwent two sequencing runs, as a larger fraction of ASVs could be recalled in the second sequencing run (Amir et al., 2017).

Along with the widespread use of 16S rRNA sequencing, high-resolution ASVs have become more popular than OTUs. As only a single analysis method is usually selected in a particular study, there is a need for a thorough comparison of OTU clustering and denoising pipelines, as different methods may lead to different conclusions in some cases. However, there are not many studies on this topic. OTU clustering approaches have been compared using samples from chicken cecum (Allali et al., 2017) and preterm infants (Plummer and Twin, 2015), and a comparison of denoising algorithms using environmental samples was reported in 2018 (Nearing et al., 2018). However, a thorough comparison between denoising algorithms and OTU clustering approaches has yet to be conducted using clinical samples, and no comparison has been conducted specifically in colorectal cancer patients. In this study, based on 358 16S rRNA sequencing samples, including 184 CRC samples and 174 healthy human samples, a comparison was conducted between two selected denoising methods (DADA2 and Deblur), as well as *de novo* OTU clustering; additionally, disease-related markers were identified and the potential efficiency of a disease-diagnostic model based on machine learning algorithms was evaluated. The aim of this study was to assess whether similar biological conclusions regarding microbiome composition could be obtained using different methods.

Methods

Data acquisition and study design

A total of 358 samples from a Chinese cohort were selected for inclusion in this study; these consisted of 184 CRC samples and 174 healthy control samples. For each of these samples, the V3-V4 region of the 16S rRNA gene was amplified using 319F/806R primer, and an Illumina MiSeq was used to generate 2×300bp reads. Reads were downloaded from the SRA database with the accession number PRJNA763023 (Yang et al., 2021). We selected samples from the older adult population (late-onset CRC patients and age-matched healthy controls) within the Fudan cohort, and selected samples with > 20,000 sequences in order to reduce the effect of lower numbers of sequences. The taxonomic composition of bacteria as established using multiple approaches (OTU clustering, DADA2, and Deblur) was first compared, and disease-related markers obtained based on the aforementioned methods were subsequently also compared (Figure 1). Next, the samples were randomly divided into a training set (70% of the data) for construction of a CRC classifier and a random forest model test

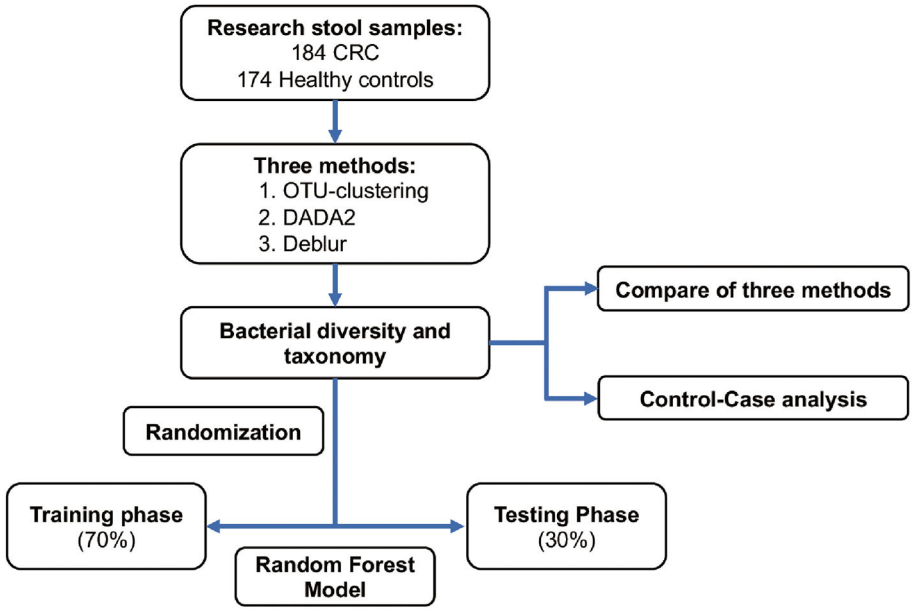

FIGURE 1
Study design and flow chart. A total of 358 stool samples from China were included in this study (184 samples from CRC patients and 174 from healthy controls). The fecal microbiota was assessed using 16S rRNA gene sequencing. Samples were randomly divided into a training set (70%) and a test set (30%). CRC, colorectal cancer; OTU clustering, *de novo* OTU-clustering-based analysis; DADA2,: DADA2 algorithm-based analysis; Deblur, Deblur algorithm-based analysis.

set (30% of the data), which was used in a testing phase to verify the potential of the model.

Data analysis

The 16S rRNA gene sequencing data were analyzed using the QIIME2 platform (v2020.2) (Hall and Beiko, 2018), which includes the VSEARCH software (Rognes et al., 2016) and software tools for DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017). For the OTU clustering method: in brief, primers were removed using the Cutadapt plugin, and paired reads were merged using the "join-pairs" function of the VSEARCH plugin. The merged reads were then dereplicated (using the "dereplicate-sequences" function), singletons were filtered (using the "feature-table filter-features" function), chimeras were filtered (using the "uchime-ref" function, with the Greengenes13_8 97% OTU database as a reference), and the results were clustered at 97% identity using the *de novo* clustering method (the "cluster-features-de-novo" function) via the QIIME2 VSEARCH plugin with default settings. For DADA2-based methods, reads were truncated to lengths of 290 bp and 220 bp for the forward and reverse reads, respectively, to remove low-quality bases at the end of the reads, and the DADA2 plugin was run with default settings to construct the ASV feature table. For Deblur-based methods, the joined reads from VSEARCH were input into the Deblur plugin to construct the ASV feature table with default settings, and singletons were filtered. The OTUs and ASVs were then compared against the Silva Database (v138.1, https://www.arb-silva.de, download code: qiime2 rescript get-silva-data–pversion "138.1"–p-target "SSURef_NR99") (Pruesse et al., 2007) using the "classify-sklearn" algorithm via the feature-classifier plugin (Bokulich et al., 2018). Data on read numbers are listed in Supplementary Table S1, and rarefaction plots are presented in Supplementary Figure S1. We also conducted alpha and beta diversity analyses using the diversity plugin in QIIME2.

Bacterial taxonomic analysis

Typically, 16S rRNA data are examined at the genus level in further analysis, so genus profiles were entered into the following analyses. Bacterial taxonomic analyses were carried out, and comparisons between the three methods were conducted using the Wilcoxon rank sum test (Bauer, 1972). Linear discriminant analysis effect size (LEfSe, http://huttenhower.sph.harvard.edu/lefse/) (Segata et al., 2011) was used to identify disease-associated taxonomic features that could be used to explain differences between controls and cases. These features were selected via LEfSe analysis using the Kruskal–Wallis rank sum test ($P < 0.05$), and linear discriminant analysis (LDA score > 2) was used to assess the effect size associated with each feature.

Analysis of diagnostic models

In order to differentiate CRC samples from healthy samples, a random forest (RF) model (Liu and Zhao, 2017; Yachida et al., 2019) was built using the random Forest package (v4.6) in R. Receiver operating characteristic (ROC) curves were constructed, and the area under the curve (AUC) was calculated to evaluate

the diagnostic performance of these RF models using the pROC package (v1.17.0.1) in R. Subsequently, differences between the three models (constructed based on each of the three methods) in terms of diagnostic model efficiency were evaluated using the roc.test() function in the pROC package.

Statistical analysis

The Mann–Whitney *U* test was employed to evaluate differences between groups. Permutational multivariate analysis of variance (PERMANOVA) was conducted to analyze the variance of the data generated using different methods, and the Mantel test was used to analyze the associations between these data. Spearman correlation analysis was performed to analyze the correlations between microbiota features. Plots were constructed using the ggplot2 package (v3.3.3) in R.

Results

Variation in taxonomic community composition across different methods

The number of OTUs/ASVs obtained using the three methods varied considerably, with DADA2 obtaining the most variants, followed by Deblur and OTU clustering (Figure 2A). At the taxonomic levels of genus and species, there was little difference between DADA2 and Deblur. However, OTU clustering obtained the largest number at the genus level and Deblur the smallest. In terms of alpha diversity, we found that all indices differed significantly between methods, as DADA2 produced the highest Shannon index, while OTU clustering produced the highest observed OTUs index and the highest Chao1 index (Figure 2B, Supplementary Figures S2A, B).

For exploration of the difference among the three methods in terms of taxonomic profiles generated, the genus level was selected for analysis. First, the Venn analysis indicated that a total of 429 genera could be detected by all three of the methods, accounting for 58.9% of the total number (the total number of detected genera was 729), and only 71 genera showed significant differences (Kruskal–Wallis, $P_{fdr} < 0.05$) in abundance among the three methods (Figure 2C, Supplementary Tables S2, S3). There were 163 genera that were identified by both OTU clustering and DADA2, but not Deblur, and there were significant difference between these two groups for 4 genera (Mann–Whitney test, $P_{fdr} < 0.05$) (Figure 2C, Supplementary Tables S2, S4). The genera obtained by OTU clustering and DADA2 mostly overlapped, with 592 shared genera, accounting for 81.2% of the total number. Regarding the 14 genera shared only by DADA2 and Deblur (and not OTU clustering), there was no difference between the two groups (Figure 2C, Supplementary Table S2). Finally, an interesting finding was that OTU clustering and Deblur did not identify any genera in common beyond the 429 identified by all methods (Figure 2C), indicating that the genera detected by Deblur were also detected by DADA2.

In most studies, taxa with higher abundance are more easily identified. Therefore, we also focused on comparison of the top 25 genera in terms of abundance; all these genera were detected by all three methods, with only four genera showing inter-method differences (Kruskal–Wallis, $P_{fdr} < 0.05$) (Figure 2D, Supplementary Table S6).

PCoA analysis based on Bray–Curtis dissimilarity was used to examine the sample clustering based on different methods; the results showed that the three methods could not be clearly distinguished, as most samples were clustered together, with a $P > 0.05$ for PERMANOVA (Figure 2E). In addition, a Mantel test was performed to evaluate the correlations among the taxonomic profiles obtained by the three methods; the results showed that there were significant correlations among the three profiles ($r = 0.99$, $P < 0.001$) (Figure 2F). These results indicated that, although there were some differences in the number of taxa obtained by the three methods, the taxonomic profiles obtained using the different methods were strongly correlated, and the relative abundances of the expected taxa under each method were strikingly similar.

Analysis of case–control differences

We analyzed the differences between cases and controls to evaluate whether the different algorithms produced different disease-related outcomes. First, we found that although the Shannon index was lower for CRC patient samples than for healthy samples across all three methods, only DADA2 identified a significant difference between the groups (Figure 3A). In contrast, inter-group differences on the Chao1 index and the observed OTUs index were observed under all methods (Supplementary Figures S2C, D).

According to the results of PCoA analysis based on Bray–Curtis dissimilarity (Figures 3B–D), the samples from the CRC patients and healthy controls could be clearly distinguished into two groups; the *P*-values of PERMANOVA were also significant, indicating that the different methods could produce the same conclusion.

Subsequently, LEfSe analysis was conducted to identify disease-related markers that distinguished the CRC group and the healthy group. Under this analysis, OTU clustering produced the largest number of markers (47 markers), followed by DADA2 (40 markers) and Deblur (39 markers) (Figures 3E, F, Supplementary Table S7, Supplementary Figures S3A–C). A total of 49 markers were obtained, among which 37 could be detected using all three methods (Figures 3E, F). All three methods indicated enrichment of 13 genera in the CRC group; these included *Fusobacterium*, *Gemella, Peptostreptococcus*, and *Streptococcus*, which have been reported on widely in CRC research (Kwong et al., 2018; Brennan and Garrett, 2019; Wong et al., 2019). In contrast, the different methods identified 23 genera as enriched in the healthy group; these included *Roseburia*, *Faecalibacterium*, and *Blautia*, which have been proven to have a positive effect on human health. However, the results for *Lachnoclostridium, Escherichia-Shigella,* and *Megamonas* differed between the three methods. This indicates that, though there were a small number of differences in disease-related markers, identification of most of the markers could be reproduced using different methods.

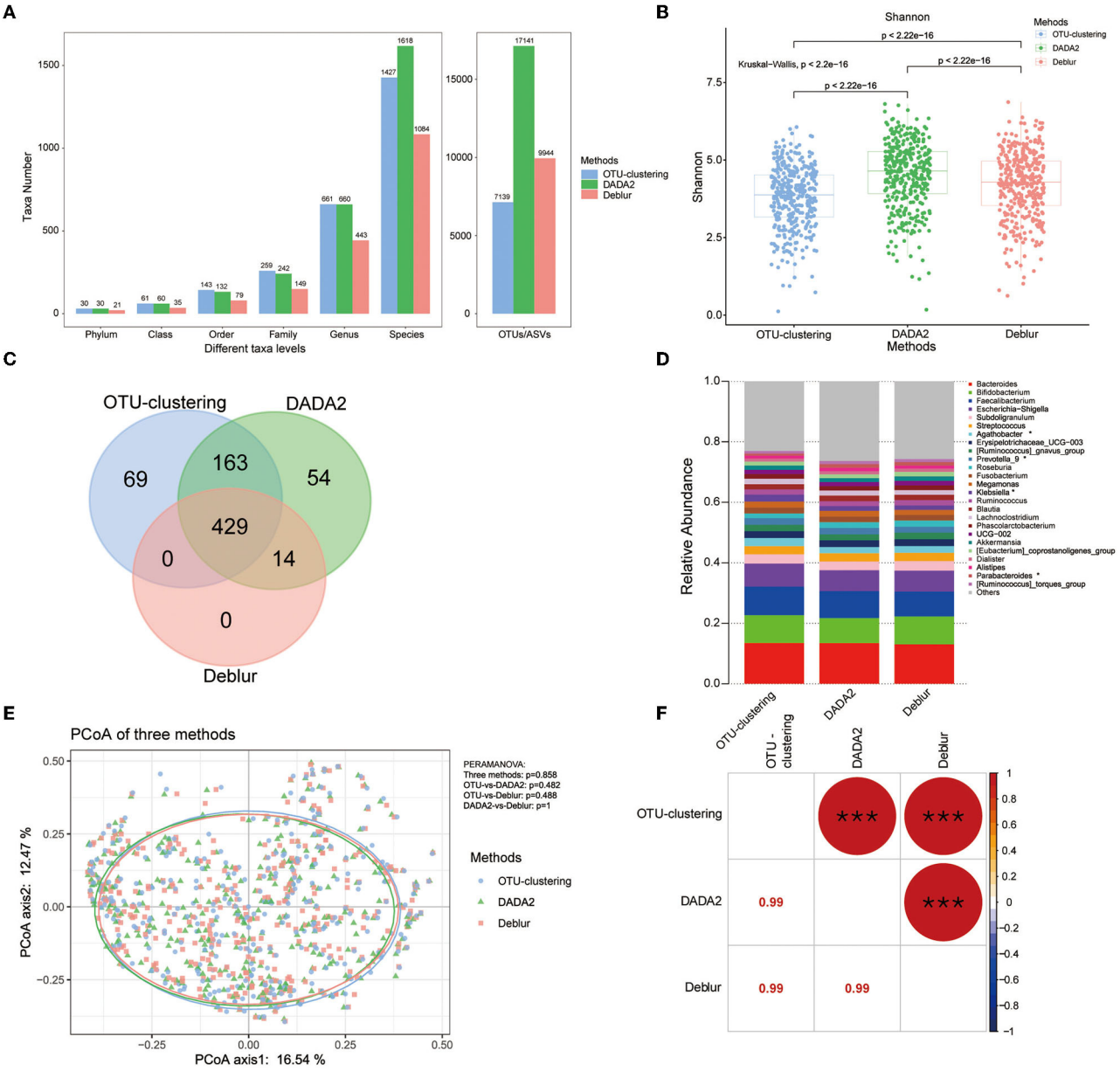

FIGURE 2
Comparison of bacterial diversity and taxonomic analyses under each of the three methods. **(A)** Number of taxa assessed at different taxonomic levels. **(B)** Shannon index across all samples calculated under each method. *P*-values for comparisons between two methods were calculated using the Mann–Whitney *U* test; all three methods were compared via the Kruskal–Wallis test. **(C)** Venn diagram showing analysis at the genus level. **(D)** Relative abundance of the top 25 genera across all samples; *indicates Kruskal–Wallis $P_{fdr} < 0.05$. **(E)** PCoA analysis among the three methods, with *P*-values as calculated by PERMANOVA. **(F)** Mantel test comparing the taxonomic profiles obtained using each of the three methods. The number in the lower cell represents the Spearman correlation coefficient; the circle in the upper cell represents the *P*-value of the correlation. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Differences between disease-diagnostic models

One of the scenarios for application of an understanding of the gut microbiome in CRC is early screening and auxiliary diagnosis; this is also the most valuable application of studies of this type. Therefore, we used the random forest algorithm to construct disease diagnosis models based on the disease-related markers identified through LEfSe analysis (Supplementary Table S7), and evaluated the diagnostic efficiency of these models on the basis of the AUC according to the associated ROC curves. All models could distinguish between CRC and healthy samples well, as the AUC values were >86%. The results indicated that the model based on DADA2 analysis exhibited the best performance (training set: AUC = 89.4%, CI 85.72–93.16%; test set: AUC = 89.7%, CI 82.71–95.44%), followed by the OTU clustering model (training set: AUC = 88.9%, CI 84.83–92.93%; test set: AUC = 88.3%, CI 81.72–94.87%) and finally the Deblur model (training set: AUC =

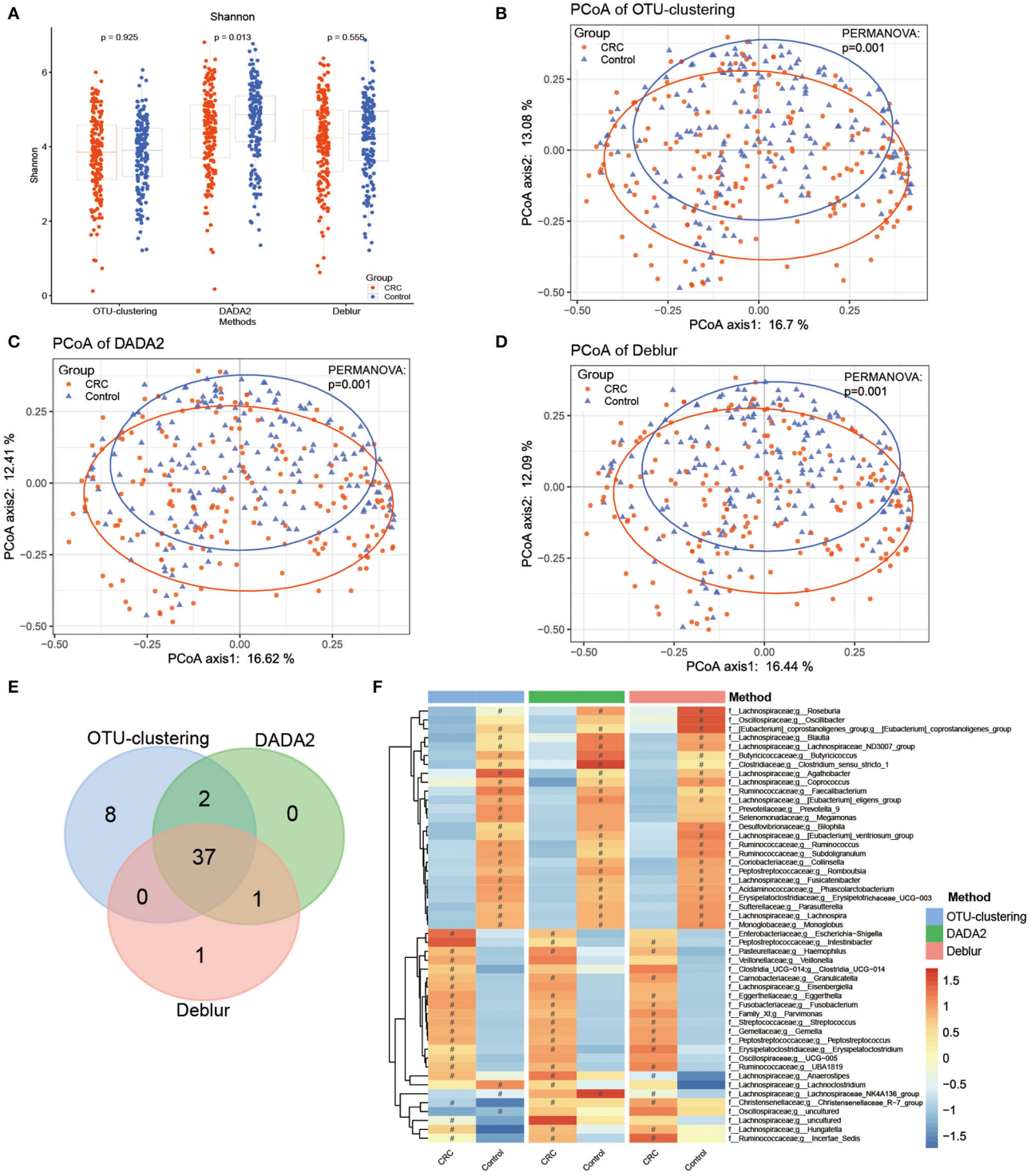

FIGURE 3
Bacterial diversity and taxonomic analysis of differences between cases and controls. **(A)** Comparison of Shannon index between cases and controls, with *P*-values as calculated for the Mann–Whitney U test. **(B–D)** PCoA analysis comparing cases and controls under **(B)** the OTU clustering method, **(C)** the DADA2 method, **(D)** the Deblur method, with *P*-values calculated by PERMANOVA. **(E)** Venn diagram showing analysis of disease-related markers based on LEfSe analysis. **(F)** Heatmap of disease-related markers based on LEfSe analysis. # represents enriched groups.

86.8%, CI 82.45–91.15%; test set: AUC = 87.5%, CI 80.68–94.23%) (Figures 4A, B, Supplementary Figures S4A–F). However, the only significant difference in AUC was between the DADA2 and Deblur models in the training set ($P < 0.05$) (Figure 4A). These results indicated that there was no significant difference between the three methods in terms of the efficiency of the resulting diagnostic model.

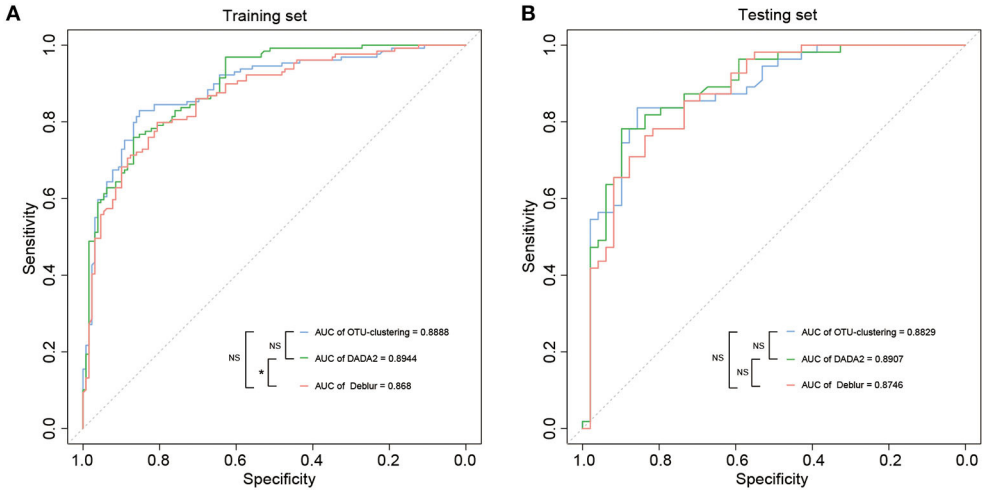

FIGURE 4
Disease classification models based on disease-associated taxa. **(A, B)** The AUC value of disease classification models using the random forest in the training set **(A)** and testing set **(B)** based on disease-associated taxa from LEfSe. The color of the lines means the data were obtained from different methods, and "*" suggested significant differences ($P < 0.05$) between models using the roc.test() function, while NS suggested that there was no significant difference ($P > 0.05$).

Discussion

We compared the results obtained by using three methods (DADA2, Deblur, and OTU clustering) to analyze data from a clinical cohort. First, we found that DADA2 and Deblur could identify more ASVs than OTUs in the community (Figure 2A). The reason for this is that, during the error-correction process of the denoising algorithms, the identity standard was 100%, and the obtained ASVs were considered to represent real individuals. In contrast, during the process of *de novo* OTU clustering, the identity standard was 97%, without no reliance on any database, and the number of OTUs was relatively low. The results under the DADA2 and Deblur methods indicated higher alpha diversity (Figure 2B, Supplementary Figures S2A, B). Comparing the two denoising methods, the number of ASVs obtained using DADA2 was much higher than that obtained using Deblur (Figure 2A), which may be because only singletons were removed from all samples by default under DADA2; in contrast, under Deblur, not only were singletons in a single sample removed, but sequences below 10 in all samples were also removed. These results suggest that DADA2 may be more sensitive to rare species. A report by Nearing et al. has also claimed that DADA2 can obtain more ASVs than Deblur in analysis of simulated and environmental microbial communities (Nearing et al., 2018).

All three methods included a process for removal of singletons with low frequencies, which not only removed some errors but also discarded some rare taxa. This requires users to make a reasonable selection of thresholds according to the purpose of their analysis. OTU clustering has clear disadvantages in this regard, as some sequences below the threshold cannot be accurately distinguished. The classic *de novo* clustering algorithm is Uparse (Edgar, 2013), whose developer, Robert Edgar, argued that the 97% threshold for taxonomic classification was too low; he proposed that the threshold for 16S sequences with full length should be 99% and that the threshold for the V4 region should be 100% in order to improve accuracy (Edgar, 2018). However, use of a higher threshold in OTU clustering may introduce identification of spurious taxonomic units without correction of sequencing errors, as employed under DADA2 and Deblur. In addition, data from studies based on OTU clustering cannot be combined for further analysis, which introduces major challenges in drawing comparisons between different studies, while DADA2 and Deblur do not have this problem.

After obtaining feature sequences (ASVs/OTUs), we compared them with the reference database containing known taxa in order to obtain taxonomic information using the classify-sklearn algorithm, which is a form of naive Bayes classifier (a classification method based on machine learning) (Bokulich et al., 2018). Previous studies have shown that the classify-sklearn algorithm could provide more accurate annotation information at the genus and species levels (Kaehler et al., 2019; Ziemski et al., 2021). The results showed that, as the annotation level increased, the gaps in the number of taxa obtained using each of the three methods gradually narrowed, especially the gap between DADA2 and OTU clustering (Figure 2A). At the species level, the largest number of taxa was obtained using DADA2 (1,618), followed by OTU clustering (1,427) and then Deblur (1,084). At the genus level and higher, the numbers obtained using DADA2 and OTU clustering were very similar, whereas Deblur consistently identified the smallest number of taxa, with a loss of approximately 30%−70%. The number of taxa identified using Deblur was the lowest at the species level and higher, which may be related to the process used under Deblur for filtering sequences below 10 in all samples, resulting in the removal of low-frequency taxa. DADA2 removed only singletons in all samples, which not only increased the number of ASVs but also made it easier to obtain more taxa, especially

at the species level. However, DADA2 and OTU obtained similar numbers of taxa at the genus level and higher, indicating that DADA2 had this advantage only at the species level. Therefore, DADA2 is recommended in cases where the researchers need more information on specifically on ASVs and at the species level.

Usually, further analysis of 16S rRNA data is conducted at the genus level, so the genera profiles were selected for subsequent analyses. Venn diagram analysis showed that the three methods obtained a total of 729 genera, of which 429 genera (58.9%) could be identified using all three methods (Figure 2C). Among these genera, 71 differed significantly among methods (Kruskal–Wallis $P_{fdr} < 0.05$) (Figure 2C). OTU clustering and DADA2 shared 81% of the detected genera, while DADA2 and Deblur shared only 60%. Regarding the 71 genera with inter-method differences, most of the genera were enriched under OTU clustering (Supplementary Table S3), possibly because OTU clustering retained more sequences. Although the number of genera identified using DADA2 and OTU clustering was similar, each also identified some unique genera that was not identified by the other, and the relative abundance of these unique genera was very low: the maximum abundance was 3.454e-06 among the 69 genera identified only by OUT clustering and 4.18e-06 among the 69 genera identified identified only by DADA2 (Figure 2C, Supplementary Table S5). Traditionally, genera with higher abundance as regarded as playing an important role in the community. These results showed that the top 25 genera could be detected by all three methods, including four significant taxa (Kruskal–Wallis, $P_{fdr} < 0.05$, Supplementary Table S6). This indicates, although the three methods will lead to identification of different genera, the differences are mainly reflected in genera with low relative abundance, and researchers need to choose appropriate methods according to their objectives.

PCoA analysis showed that the samples could not be distinguished well through use of different methods. PERMANOVA also indicated that the use of different methods had no significant effect on the microbial profile obtained ($P > 0.05$) (Figure 2E). The microbial profiles were robustly correlated across methods (Mantel test, $n = 358$, $r = 0.99$, $P < 0.001$) (Figure 2F), confirming the strong similarity of the results across the three methods.

In clinical research, disease-related bacteria are important for researchers. Therefore, we compared the results of case–control analyses based on each of the three methods. First, in alpha diversity analyses, inter-group differences in Shannon index were observed only under DADA2, while inter-group differences on the Chao1 index and the observed OTUs index were observed under all methods, with the CRC group showing significantly lower diversity than the healthy group (Figure 3A, Supplementary Figures S2C, D). This result was consistent with those of previous CRC studies, which have shown that alpha diversity is reduced in the CRC population (Yang et al., 2021); other studies have also observed no significant difference in the Shannon index (Feng et al., 2015; Wu et al., 2021). This finding suggests that researchers must select the appropriate alpha index according to their methods. PCoA analysis showed that all three methods could distinguish successfully between the CRC group and the healthy group, with $P < 0.05$ in PERMANOVA (Figures 3B–D), confirming the difference between the cases and controls as compared using different methods.

LEfSe analysis was used to explore disease-related microbial markers. A total of 49 markers (OTU-clustering: 47; DADA2: 40; Deblur: 39) were obtained using the three methods, of which 37 markers (75.5%) could be reproduced using a different method; the enrichment trend was consistent for most markers (Figure 3F, Supplementary Figures S3A–C). For example, *Haemophilus*, *Granulicatella*, *Eggerthella*, *Fusobacterium*, *Parvimonas*, *Streptococcus*, *Gemella*, *Peptostreptococcus*, and other genera (for a total of 13) were found under all three methods to be enriched in the CRC group. *Fusobacterium,* especially *Fusobacterium nucleatum,* is an important marker of CRC (Kostic et al., 2012; Yang et al., 2021) and an opportunistic pathogen in many chronic oral and intestinal diseases, such as inflammatory bowel disease (IBD) (Weng et al., 2019). *Fusobacterium nucleatum* has been reported to promote glycolysis and oncogenesis in CRC by upregulating the lncRNA ENO1-IT1 (Hong et al., 2021) and promoting CRC cell migration by modulating the long non-coding RNAs keratin7-antisense (KRT7-AS) and keratin7 (KRT7) (Chen et al., 2020). *Streptococcus* is also a common pathogenic genus that often causes inflammation and bacteremia, possibly promoting CRC (Kwong et al., 2018; McAuliffe et al., 2019). *Peptostreptococcus* is an anaerobic, gram-positive bacterium, and *Peptostreptococcus anaerobius* has been reported to promote CRC and modulate tumor immunity (Long et al., 2019). *Haemophilus* is an opportunistic pathogen that may cause hemorrhagia and acute meningitis. Finally, *Parvimonas* is a fastidious, anaerobic, gram-positive coccus that is widely found among healthy human oral and gastrointestinal flora, and previous studies have demonstrated that *Parvimonas micra* is associated with CRC (Löwenmark et al., 2020; Xu et al., 2020).

Roseburia, *Faecalibacterium*, *[Eubacterium]_eligens_group, Bilophila, Phascolarctobacterium, Butyricicoccus, Blautia*, and other genera (for a total of 23) were found to be enriched in the healthy group. *Butyricicoccus* and *Faecalibacterium* are butyric acid producers, which play an important role in intestinal and host health and act as protectors against CRC (Miquel et al., 2013; Zhou et al., 2018; Chang et al., 2020). Butyric acid is the main energy source of colonic epithelial cells and can reduce the pH in the colon, regulate human immunity, and exert anti-inflammatory effects. *Faecalibacterium* has been found in multiple studies to be significantly decreased in many diseases (Lopez-Siles et al., 2017), including Crohn's disease (CD) (Martinez-Medina et al., 2006), ulcerative colitis (UC) (Machiels et al., 2014), inflammatory bowel diseases (IBD) (Frank et al., 2007), and CRC. *Roseburia* can ferment various carbohydrates and may play a positive role in exerting anti-inflammatory effects and preventing CRC (Machiels et al., 2014). *Phascolarctobacterium* produces short-chain fatty acids (SCFAs) and plays various important roles in maintaining human health, such as enhancing gastrointestinal function, reducing inflammation levels, and influencing metabolic state and mood of the host (Wu et al., 2017). Finally, *Blautia* produces acetic acid, which not only contributes to gas emissions in the intestine but also exerts anti-inflammatory effects (Liu et al., 2021; Miyake et al., 2021).

The markers obtained using each of the three methods were not always completely consistent. For example, *Lachnoclostridium* was found to be enriched in the healthy group under OTU clustering, enriched in the CRC group under DADA2, and not significantly enriched in either group under Deblur. *Lachnoclostridium* has been reported to be enriched in colorectal adenoma and cancer (Li et al., 2020). *Escherichia-Shigella* was found to be enriched in CRC under DADA2 and OTU clustering, which is consistent with previous studies (Wang et al., 2012; Han et al., 2019). *Intestinibacter* was found to be enriched in CRC under DADA2 and Deblur; this genus has been reported to be associated with immune-mediated inflammatory diseases (IMIDs) and to be found in greater abundance in CD (Forbes et al., 2018). *Megamonas* was found to be enriched in the healthy group only under OTU clustering. *Megamonas* has previously been reported to be enriched in healthy controls compared with cachectic cancer patients (Ubachs et al., 2021). *Megamonas* can ferment many carbohydrates, producing various intestinal epithelial cell nutrients, such as acetic acid, propionic acid, and lactic acid (Tian et al., 2020; Ubachs et al., 2021). *Oscillibacter* was found to be enriched in the healthy group only under Deblur; this genus has previously been proven to be enriched in normal tissue samples compared to their respective tumor counterparts (Loke et al., 2018). These results indicate that, even though a small number of markers differed between the methods, most of the disease-related markers identified were consistent across methods.

Finally, random forest (RF) models were constructed to evaluate diagnostic efficiency based on each of the three methods. For the training set, the highest AUC was obtained for the DADA2 model, at 89.4%, followed by 88.9% for the OTU clustering model and 86.8% for the Deblur model; the only significant difference in AUC was between DADA2 and Deblur ($P < 0.05$) (Figure 4A, Supplementary Figures S4A, C, E). This trend was subsequently verified in the test set (DADA2 > OTU clustering > Deblur), but there were no significant differences among the AUCs for the three models ($P > 0.05$) (Figure 4B, Supplementary Figures S4B, D, F). These results indicate that each of the different methods can achieve good diagnostic efficiency, with DADA2 being the best.

In conclusion, although there were differences in the number of OTUs/ASVs obtained using the three methods, the differences in the numbers of taxa were smaller, especially for the comparison between DADA2 and OTU clustering at the genus level. Moreover, the microbial profiles were strongly correlated. This indicates that the results obtained using the three methods are comparable. Case–control analysis also showed that the three methods could yield similar results, with mostly consistent identification of CRC-related markers. However, it should also be noted that the three methods were performed with the default parameters, and adjusting some of the parameters in the selected method could help users to obtain their desired results.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JW and GL designed the research study. GL and TL performed bioinformatics analysis and wrote the manuscript. XZha and XZhu edited the manuscript. All authors read and approved the final manuscript.

Funding

The work was supported by grants from the Natural Science Basic Research Program of Shaanxi (grant number 2020JC-01).

Conflict of interest

GL is employed by Guangdong Hongyuan Pukong Medical Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1178744/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Rarefaction plots of the Shannon index **(A–D)** and observed OTUs index **(E–H)**.

SUPPLEMENTARY FIGURE S2
(A, B) Chao1 index **(A)** and observed OTUs index **(B)** of all samples using three methods. *P*-values comparing two methods were calculated using the Mann–Whitney U test; the Kruskal–Wallis test was used in the three-method comparison. **(C, D)** Chao1 index **(C)** and observed OTUs **(D)** for cases and controls based on three methods. *P*-values for comparisons of the two groups were calculated using the Mann–Whitney *U* test.

SUPPLEMENTARY FIGURE S3
LEfSe analysis comparing CRC patients and controls based on three methods.

SUPPLEMENTARY FIGURE S4
The performance of the disease classification models based on the three methods was evaluated using AUCs. OTU clustering training set **(A)** and test set **(B)**; DADA2 training set **(C)** and test set **(D)**; Deblur training set **(E)** and test set **(F)**. The confidence interval for the AUC value was calculated.

SUPPLEMENTARY TABLE S1
The reads were obtained per sample.

SUPPLEMENTARY TABLE S2
The number of genera from venn analysis among three methods.

SUPPLEMENTARY TABLE S3
Kruskal–Wallis analysis of genera detected by three methods.

SUPPLEMENTARY TABLE S4
Mann-Whitney test of genera detected by OTU-clustering and DADA2 alone in venn analysis.

SUPPLEMENTARY TABLE S5
The averege relative abundance of genera only detected by OTU-clustering or DADA2.

SUPPLEMENTARY TABLE S6
Kruskal–Wallis analysis of top25 genera detected by three methods.

SUPPLEMENTARY TABLE S7
LEfSe analysis between CRC and Control based on three methods.

References

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, 18. doi: 10.1186/gb-2011-12-2-r18

Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., et al. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiol.* 17 doi: 10.1186/s12866-017-1101-8

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*. 2, 16. doi: 10.1128/mSystems.00191-16

Ashktorab, H., Kupfer, S. S., Brim, H., and Carethers, J. M. (2017). Racial disparity in gastrointestinal cancer risk. *Gastroenterology*. 153, 910–23. doi: 10.1053/j.gastro.2017.08.018

Bauer, D. F. (1972). Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc*. 67, 687–690. doi: 10.1080/01621459.1972.10481279

Boers, S. A., Jansen, R., and Hays, J. P. (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *Eur. J. Clin. Microbiol. Infect. Dis*. 38, 1059–70. doi: 10.1007/s10096-019-03520-3

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2ʹs q2-feature-classifier plugin. *Microbiome*. 6, 90. doi: 10.1186/s40168-018-0470-z

Brennan, C. A., and Garrett, W. S. (2019). Fusobacterium nucleatum—symbiont, opportunist and oncobacterium. *Nat. Rev. Microbiol.* 17, 156–66. doi: 10.1038/s41579-018-0129-6

Brenner, H., Kloor, M., and Pox, C. P. (2014). Colorectal cancer. *Lancet (London, England)*. 383, 1490–502. doi: 10.1016/S0140-6736(13)61649-9

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., et al. (2017). Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science*. 358, 1443–8. doi: 10.1126/science.aal5240

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., Holmes, S. P., et al. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*. 13, 581–3. doi: 10.1038/nmeth.3869

Castellarin, M., Warren, R. L., Freeman, J. D., Dreolini, L., Krzywinski, M., Strauss, J., et al. (2012). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*. 22, 299–306. doi: 10.1101/gr.126516.111

Chang, S. C., Shen, M. H., Liu, C. Y., Pu, C. M., Hu, J. M., Huang, C. J. A., et al. (2020). gut butyrate-producing bacterium Butyricicoccus pullicaecorum regulates short-chain fatty acid transporter and receptor to reduce the progression of 1,2-dimethylhydrazine-associated colorectal cancer. *Oncol. Lett*. 20, 12190. doi: 10.3892/ol.2020.12190

Chen, S., Su, T., Zhang, Y., Lee, A., He, J., Ge, Q., et al. (2020). Fusobacterium nucleatum promotes colorectal cancer metastasis by modulating KRT7-AS/KRT7. *Gut Microbes*. 11, 511–25. doi: 10.1080/19490976.2019.1695494

Coenye, T., and Vandamme, P. (2003). Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol. Lett*. 228, 45–9. doi: 10.1016/S0378-1097(03)00717-1

Dalal, N., Jalandra, R., Bayal, N., Yadav, A., Harshulika, K., Sharma, M., et al. (2021). Gut microbiota-derived metabolites in CRC progression and causation. *J. Cancer Res. Clin. Oncol*. 147, 3141–55. doi: 10.1007/s00432-021-03729-w

Díaz-Tasende, J. (2018). Colorectal cancer screening and survival. *Rev. Esp. Enferm. Dig*. 110, 681–3. doi: 10.17235/reed.2018.5870/2018

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*. 34, 2371–5. doi: 10.1093/bioinformatics/bty113

Edgar, R. C.. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods*. 10, 996–8. doi: 10.1038/nmeth.2604

Fearon, E. R., and Vogelstein, B. A. (1990). genetic model for colorectal tumorigenesis. *Cell*. 61, 759–67. doi: 10.1016/0092-8674(90)90186-I

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L. et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature communica*. 6, 6528. doi: 10.1038/ncomms7528

Forbes, J. D., Chen, C. Y., Knox, N. C., Marrie, R. A., El-Gabalawy, H., Kievit, D. T., et al. (2018). A comparative study of the gut microbiota in immune-mediated inflammatory diseases-does a common dysbiosis exist? *Microbiome*. 6, 1. doi: 10.1186/s40168-018-0603-4

Frank, D. N., St. Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., Pace, N. R., et al. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. USA*. 104, 13780–5. doi: 10.1073/pnas.0706625104

Hall, M., and Beiko, R. G. (2018). 16S rRNA gene analysis with QIIME2. *Methods Mol. Biol*. 1849, 113–29. doi: 10.1007/978-1-4939-8728-3_8

Han, S., Pan, Y., Yang, X., Da, M., Wei, Q., Gao, Y., et al. (2019). Intestinal microorganisms involved in colorectal cancer complicated with dyslipidosis. *Cancer Biol. Ther*. 20, 81–9. doi: 10.1080/15384047.2018.1507255

Hong, J., Guo, F., Lu, S. Y., Shen, C., Ma, D., Zhang, X., et al. (2021). F. nucleatum targets lncRNA ENO1-IT1 to promote glycolysis and oncogenesis in colorectal cancer. *Gut*. 70, 2123–37. doi: 10.1136/gutjnl-2020-322780

Jones, S., Chen, W. D., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., et al. (2008). Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. USA*. 105, 4283–8. doi: 10.1073/pnas.0712345105

Kaehler, B. D., Bokulich, N. A., McDonald, D., Knight, R., Caporaso, J. G., Huttley, G. A., et al. (2019). Species abundance information improves sequence taxonomy classification accuracy. *Nat. Commun*. 10, 4643. doi: 10.1038/s41467-019-12669-6

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res*. 22, 292–8. doi: 10.1101/gr.126573.111

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol*. 12, 118–23. doi: 10.1111/j.1462-2920.2009.02051.x

Kwong, T. N. Y., Wang, X., Nakatsu, G., Chow, T. C., Tipoe, T., Dai, R. Z. W., et al. (2018). Association between bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. *Gastroenterology*. 155, 383–390.e8. doi: 10.1053/j.gastro.2018.04.028

Li, T., Nakatsu, G., Chen, Y. X., Yau, T. O., Chu, E., Wong, S., et al. (2020). A novel faecal Lachnoclostridium marker for the non-invasive diagnosis of colorectal adenoma and cancer. *Gut*. 69, 1248–57. doi: 10.1136/gutjnl-2019-318532

Liu, X., Mao, B., Gu, J., Wu, J., Cui, S., Wang, G., et al. (2021). Blautia-a new functional genus with potential probiotic properties? *Gut Microbes*. 13, 1–21. doi: 10.1080/19490976.2021.1875796

Liu, Y., and Zhao, H. (2017). Variable importance-weighted random forests. *Quant Biol (Beijing, China)*. 5, 338–51. doi: 10.1007/s40484-017-0121-6

Loke, M. F., Chua, E. G., Gan, H. M., Thulasi, K., Wanyiri, J. W., Thevambiga, I., et al. (2018). Metabolomics and 16S rRNA sequencing of human colorectal cancers and adjacent mucosa. *PLoS ONE*. 13, e0208584. doi: 10.1371/journal.pone.0208584

Long, X., Wong, C. C., Tong, L., Chu, E. S. H., Ho Szeto, C., Go, M. Y. Y., et al. (2019). Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat Microbiol*. 4, 2319–30. doi: 10.1038/s41564-019-0541-3

Lopez-Siles, M., Duncan, S. H., Garcia-Gil, L. J., and Martinez-Medina, M. (2017). Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics. *ISME J*. 11, 841–52. doi: 10.1038/ismej.2016.176

Löwenmark, T., Löfgren-Burström, A., Zingmark, C., Eklöf, V., Dahlberg, M., Wai, S. N., et al. (2020). Parvimonas micra as a putative non-invasive faecal biomarker for colorectal cancer. *Sci. Rep*. 10 doi: 10.1038/s41598-020-72132-1

Machiels, K., Joossens, M., Sabino, J., De Preter, V., Arijs, I., Eeckhaut, V., et al. (2014). A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut*. 63, 1275–83. doi: 10.1136/gutjnl-2013-304833

Martinez-Medina, M., Aldeguer, X., Gonzalez-Huix, F., Acero, D., and Garcia-Gil, L. J. (2006). Abnormal microbiota composition in the ileocolonic mucosa of Crohn's disease patients as revealed by polymerase chain reaction-denaturing gradient gel electrophoresis. *Inflamm. Bowel Dis*. 12, 1136–45. doi: 10.1097/01.mib.0000235828.09305.0c

McAuliffe, A., Bhat, V., and Cox, J. (2019). Streptococcus gallolyticus Group Bacteremia and Colonic Adenocarcinoma. *J. Am. Osteopath. Assoc*. 119, 65. doi: 10.7556/jaoa.2019.011

Miquel, S., Martín, R., Rossi, O., Bermúdez-Humarán, L. G., Chatel, J. M., Sokol, H., et al. (2013). Faecalibacterium prausnitzii and human intestinal health. *Curr. Opin. Microbiol*. 16, 255–61. doi: 10.1016/j.mib.2013.06.003

Miyake, T., Mori, H., Yasukawa, D., Hexun, Z., Maehira, H., Ueki, T., et al. (2021). The comparison of fecal microbiota in left-side and right-side human colorectal cancer. *Eur. Surg. Res*. 62, 248–54. doi: 10.1159/000516922

Muthappa, D. M., Lamba, S., Sivasankaran, S. K., Naithani, A., Rogers, N., Srikumar, S., et al. (2022). 16S rRNA based profiling of bacterial communities colonizing bakery-production environments. *Foodborne Pathog. Dis*. 19, 485–94. doi: 10.1089/fpd.2022.0014

Nearing, J. T., Douglas, G. M., Comeau, A. M., and Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*. 6, 5364. doi: 10.7717/peerj.5364

Park, J., Kim, N. E., Yoon, H., Shin, C. M., Kim, N., Lee, D. H., et al. (2021). Fecal microbiota and gut microbe-derived extracellular vesicles in colorectal cancer. *Front. Oncol*. 11, 650026. doi: 10.3389/fonc.2021.650026

Patin, N. V., Kunin, V., Lidström, U., and Ashby, M. N. (2013). Effects of OTU clustering and PCR artifacts on microbial diversity estimates. *Microb. Ecol*. 65, 709–19. doi: 10.1007/s00248-012-0145-4

Plummer, E., and Twin, J. A. (2015). Comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *J. Proteomics Bioinform*. 8, 283–91. doi: 10.4172/jpb.1000381

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*. 35, 7188–96. doi: 10.1093/nar/gkm864

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. V. (2016). Search a versatile open source tool for metagenomics. *PeerJ*. 4, 2584. doi: 10.7717/peerj.2584

Sanschagrin, S., and Yergeau, E. (2014). Next-generation sequencing of 16S ribosomal RNA gene amplicons. *J. Vis. Exp*. doi: 10.3791/51709-v

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE*. 6, 27310. doi: 10.1371/journal.pone.0027310

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol*. 12, R60. doi: 10.1186/gb-2011-12-6-r60

Shaukat, A., Kahi, C. J., Burke, C. A., Rabeneck, L., Sauer, B. G., Rex, D. K. A. C. G., et al. (2021). Clinical guidelines: colorectal cancer screening 2021. *Am. J. Gastroenterol*. 116, 458–79. doi: 10.14309/ajg.0000000000001122

Stoffel, E. M., and Murphy, C. C. (2020). Epidemiology and mechanisms of the increasing incidence of colon and rectal cancers in young adults. *Gastroenterology*. 158, 341–53. doi: 10.1053/j.gastro.2019.07.055

Tian, Y., Zuo, L., Guo, Q., Li, J., Hu, Z., Zhao, K., et al. (2020). Potential role of fecal microbiota in patients with constipation. *Therap. Adv. Gastroenterol*. 13, 1756284820968423. doi: 10.1177/1756284820968423

Ubachs, J., Ziemons, J., Soons, Z., Aarnoutse, R., van Dijk, D. P. J., Penders, J., et al. (2021). Gut microbiota and short-chain fatty acid alterations in cachectic cancer patients. *J. Cachexia Sarcopenia Muscle*. 12, 2007–21. doi: 10.1002/jcsm.12804

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., et al. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J*. 6, 320–9. doi: 10.1038/ismej.2011.109

Weng, Y. J., Gan, H. Y., Li, X., Huang, Y., Li, Z. C., Deng, H. M., et al. (2019). Correlation of diet, microbiota and metabolite networks in inflammatory bowel disease. *J. Dig. Dis*. 20, 447–59. doi: 10.1111/1751-2980.12795

Westcott, S. L., and Schloss, P. D. (2015). novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*. 3, 1487. doi: 10.7717/peerj.1487

Wong, H. K., Ho, P. L., and Lee, C. K. (2019). *Streptococcus gallolyticus* bacteremia and colorectal carcinoma. *Gastroenterology*. 156, 291–2. doi: 10.1053/j.gastro.2018.07.059

Wu Y, Jiao N, Zhu R, Zhang, Y, Wu, D. et al. (2021). Identification of microbial markers across populations in early detection of colorectal cancer. *Nat Commun*. 12:3063. doi: 10.1038/s41467-021-23265-y

Wu, F., Guo, X., Zhang, J., Zhang, M., Ou, Z., Peng, Y., et al. (2017). Phascolarctobacterium faecium abundant colonization in human gastrointestinal tract. *Exp. Ther. Med*. 14, 3122–6. doi: 10.3892/etm.2017.4878

Xi, Y., and Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Transl. Oncol*. 14, 101174. doi: 10.1016/j.tranon.2021.101174

Xu, J., Yang, M., Wang, D., Zhang, S., Yan, S., Zhu, Y., et al. (2020). Alteration of the abundance of Parvimonas micra in the gut along the adenoma-carcinoma sequence. *Oncol. Lett*. 20 doi: 10.3892/ol.2020.11967

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med*. 25, 968–76. doi: 10.1038/s41591-019-0458-7

Yang, Y., Du, L., Shi, D., Kong, C., Liu, J., Liu, G., et al. (2021). Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nat. Commun*. 12, 6757. doi: 10.1038/s41467-021-27112-y

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol*. 12, 635–45. doi: 10.1038/nrmicro 3330

Zhang, F., Zhang, Y., Zhao, W., Deng, K., Wang, Z., Yang, C., et al. (2017). Metabolomics for biomarker discovery in the diagnosis, prognosis, survival and recurrence of colorectal cancer: a systematic review. *Oncotarget*. 8, 35460–72. doi: 10.18632/oncotarget.16727

Zhou, L., Zhang, M., Wang, Y., Dorfman, R. G., Liu, H., Yu, T., et al. (2018). *Faecalibacterium prausnitzii* produces butyrate to maintain Th17/Treg balance and to ameliorate colorectal colitis by inhibiting histone deacetylase 1. *Inflamm. Bowel Dis*. 24, 1926–40. doi: 10.1093/ibd/izy182

Ziemski, M., Wisanwanichthan, T., Bokulich, N. A., and Kaehler, B. D. (2021). Beating naive bayes at taxonomic classification of 16S rRNA gene sequences. *Front. Microbiol*. 12. doi: 10.3389/fmicb.2021.644487