Vietnam Journal of Biotechnology **24**(1): 29-41, 2026. DOI: 10.15625/vjbt-23336

Research Article

COMPARISON OF GUT MICROBIAL SIGNATURES ASSOCIATED WITH COLORECTAL CANCER ACROSS TWO DIFFERENT SAMPLE COLLECTION DATASETS

Thi Thu Cong Ha[1,#], Thao Hien Nguyen[1,2,#], Thi Tuyet Nhung Pham[3,4] and Thi Thanh Tam Tran[1,✉]

[1] *Department of Life Sciences, University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam*

[2] *Hanoi University of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam*

[3] *108 Military Central Hospital, 1 Tran Hung Dao, Hai Ba Trung, Hanoi, Vietnam*

[4] *Hanoi Medical University, 1 Ton That Tung, Kim Lien, Hanoi, Vietnam*

[#] *Authors contributed equally to this work*

✉To whom correspondence should be addressed. Email: tran-thi-thanh.tam@usth.edu.vn

Received: 13 November 2024; Accepted: 28 November 2025; Published online: 27 January 2026

ABSTRACT

Colorectal cancer (CRC) is among the most prevalent cancers globally and in Vietnam. Diagnosing CRC is challenging due to difficulties in tumor detection, leading to thousands of deaths annually. CRC has been demonstrated to be linked with alterations in gut microbial composition and function. Various bacterial taxa have been recognized as potential biomarkers of CRC and suspected to play a crucial role in colon carcinogenesis. This pilot study explores consistent and divergent bacterial signatures in the fecal microbiota of 15 CRC patients compared to 12 healthy individuals in two different sample collection datasets. Both datasets proceeded with the same DNA extraction method, followed by amplification of the 16S rRNA gene's hypervariable regions (V3-V4), and then sequenced with Illumina MiSeq sequencing platform at two different time points. Our findings show that the gut microbiota's alpha and beta diversity did not differ statistically significantly between the healthy individuals and the CRC patients in either dataset. We observed 12 genera in the first dataset and 13 genera in the second dataset that exhibited significant differential abundance between CRC patients and healthy controls. However, due to the small sample size, after adjustment for multiple testing, only *Peptostreptococcus* and *Fusobacterium* in Dataset 1, and *Parvimonas* in Dataset 2, remained significantly associated with CRC according to ANCOM-BC2. Notably, *Parvimonas* was also detected in CRC patients but not in healthy controls in Dataset 1. This genus may potentially be used as fecal biomarker for CRC detection in Vietnamese patients. Additionally, our study underscores the importance of validating fecal bacterial biomarkers across different sample collection datasets to improve the accuracy and effectiveness of CRC diagnosis and treatment, potentially advancing personalized treatment approaches for Vietnamese patients.

Keywords: 16S rRNA metagenomics, colorectal cancer, gut microbial signatures, sample collection dataset, Vietnamese patients.

INTRODUCTION

Colorectal cancer (CRC), consisting of colon and/or rectum cancer, which are parts of the large intestine and the digestive system, is one of the world's most commonly diagnosed cancers and one of the deadliest cancers, with approximately 903,859 deaths reported across 185 countries in 2022 (Bray *et al.*, 2024). It is the second most prevalent cause of cancer-related death worldwide and one of the most often diagnosed cancers. While early detection can greatly enhance the prognosis, the rate of early diagnosis remains limited as many patients either lack typical symptoms or exhibit vague signs during the initial stages (Duan *et al.*, 2022). In 2022, there were 16,835 new cases reported in Vietnam, ranking fourth among common cancers (9.7%) and fifth in terms of cancer-related deaths (Bray *et al.*, 2024). This disease also accounts for 5% of cancer cases in Vietnam across 63 provinces/cities of Vietnam and 40% of underreported cases due to an underdeveloped epidemiologic system (Le and Dao, 2020). The exact cause of CRC is still unclear, but it is believed to result from the interplay of multiple factors. Apart from age, factors such as an unhealthy diet, smoking, obesity, family history, high alcohol consumption, and physical inactivity can contribute to the development of CRC (Siegel *et al.*, 2020). Research in the last decade has shown that the gut microbial diversity, composition, and function are highly variable in response to environmental factors, including antibiotic usage and habitual dietary patterns, and are correlated with various diseases, including diabetes mellitus, respiratory diseases, inflammatory bowel disease, brain disorders, CRC, etc. (Hou *et al.*, 2022). Specifically, gut microbiota dysbiosis has been implicated in disease onset through multiple pathogenic mechanisms, including intestinal barrier dysfunction, induced inflammation, immune dysregulation, and metabolic disturbances (Shen *et al.*, 2025).

Each human being hosts a vast and complex community of gut microorganisms, collectively referred to as the gut microbiota, whose total cell count is now estimated to be approximately equal to that of human cells (Sender *et al.*, 2016). Previous studies have shown alterations in the gut microbial diversity and abundance of several specific bacterial taxa in patients with CRC compared to healthy controls (Flemer *et al.*, 2017; Tito *et al.*, 2024; Yu *et al.*, 2017). Among several bacterial candidates, *Fusobacterium nucleatum, Peptostreptococcus stomatis, Parvimonas micra*, and *Solobacterium moorei* were identified as universal fecal microbial signatures for CRC, consistently enriched in CRC patients across four distinct cohorts (Yu *et al.*, 2017). Similarly, a previous meta-analysis conducted on 768 CRC patients using shotgun metagenomic data confirmed the enrichment of *Fusobacterium, Porphyromonas, Parvimonas, Peptostreptococcus, Gemella, Prevotella*, and *Solobacterium* in CRC (Wirbel *et al.*, 2019). A pilot study involving 10 Vietnamese patients with newly diagnosed CRC and 5 healthy individuals found a significant increase in the abundance of *Gemella, Parvimonas*, and *Peptostreptococcus* in the CRC group (Nhung *et al.*, 2023). In a subsequent study involving Vietnamese patients, an increase in *P. micra*, *P. stomatis*, and *Prevotella intermedia*, along with a decrease in several health-associated species such as *Lactobacillus johnsonii*, *Bifidobacterium longum*, *Butyricicoccus pullicaecorum*, and *Ruminococcus* species, etc., was observed in patients with CRC (Nhung *et al.*, 2024). In this work, we carried out a comparative analysis of two Vietnamese metagenomic studies on CRC patients, focusing on gut microbiota diversity and composition. Our goal was to identify variations in microbial biomarkers between the two sample collection datasets and consistent biomarkers that could be used for the diagnosis of CRC patients in Vietnam.

MATERIALS AND METHODS

Sample collection datasets

Our study included 16S metagenomic data and clinical data, including age, gender, weight, height, and comorbidity of 27 individuals, which were selected from two different sample collection datasets. Raw sequencing data of samples in the two datasets were retrieved from Nhung *et al.* (2023) and Nhung *et al.* (2024). The first collection dataset (Dataset 1) includes 7 CRC patients and 4 healthy subjects, while the other dataset (Dataset 2) has 8 CRC patients and 8 healthy subjects. All participants were recruited from the 108 Central Military Hospital, Vietnam. Genomic DNA was isolated from fecal samples using the DNeasy® PowerSoil® Pro Kit (Qiagen, USA), followed by amplicon sequencing of the 16S rRNA gene's hypervariable regions (V3-V4) on the Illumina® MiSeq platform as previously described (Nhung *et al.*, 2023; Nhung *et al.*, 2024). Ethical clearance was granted by the Hanoi Medical University Institutional Review Board, under approval number 503/GCN-HĐĐĐNCYSH-ĐHYHN dated May 10, 2021.

Bioinformatics analysis of 16S rRNA metagenomic data

The two datasets' raw sequencing data were jointly re-processed using Quantitative Insights into Microbial Ecology 2 (QIIME 2) v2024.5 following the workflow described by Nguyen *et al.* (2025). First, the quality of the datasets from both runs was assessed using the *FastQC* tool (Andrews, 2010). Raw forward and reverse reads were imported into a QIIME artifact, and primers were trimmed from reads with *cutadapt* v4.9 implemented in QIIME 2. The trimmed reads were further processed with the *dada2 denoise-paired* command to merge paired-end reads into contigs, group sequences into amplicon sequence variants (ASVs), filter out chimera sequences (sequences created artificially from two or more different sources) and singletons. For sequence classification, we used the *qiime feature-classifier classify-sklearn* command against the SILVA v138 reference database (Cole *et al.*, 2009) with a default confidence threshold of 70%. A phylogenetic tree was constructed from ASV sequences using *qiime phylogeny align-to-tree-mafft-fasttree*. The *qiime diversity core-metrics-phylogenetic* script was applied to calculate alpha and beta diversity based on a rarefied ASV table adjusted to 37,569 reads per sample, corresponding to the minimal read count across all samples.

Statistical analysis

Downstream analysis of the 16S rRNA data was conducted with R v4.2.1 software. Differences in alpha diversity indices and taxonomic abundance between two groups (CRC patients vs. healthy controls or Dataset 1 vs. Dataset 2) were evaluated by the non-parametric statistical Mann-Whitney U test. In addition, group differences based on absolute abundance data were evaluated with ANCOM-BC2 (Lin and Peddada, 2024). To account for multiple testing, false discovery rate (FDR) correction was performed using the Benjamini–Hochberg (BH) method.

Unweighted and weighted UniFrac distances, which measure beta diversity, were displayed with principal coordinates analysis (PCoA). Group differences in beta diversity were assessed using the Permutational Multivariate Analysis of Variance test (PERMANOVA).

RESULTS

Clinical characteristics of two sample collection datasets

A total of 27 subjects participated in the 2 different sample collection datasets (15 CRC samples and 12 control samples, as shown in Table 1). No cases of type 2 diabetes were reported among either the CRC or healthy subjects across the two datasets. Both datasets included both male and female participants with a BMI ranging from 18.5 kg/m^2 to 30 kg/m^2, with the average BMI for each group presented in Table 1. Statistical significance in age at diagnosis and BMI between the CRC and healthy groups within each dataset, as well as in the combined datasets, was detected with the Mann-Whitney U test. No bias was observed in either age or BMI between the groups across the datasets ($p > 0.05$).

Table 1. Baseline characteristics of two sample collection datasets.

Characteristics		**Dataset 1**		**Dataset 2**	
		CRC	HC	CRC	HC
Number of subjects		7	4	8	8
Gender	Male, no. (%)	4 (57.1%)	1 (25.0%)	3 (37.5%)	3 (37.5%)
	Female, no. (%)	3 (42.9%)	3 (75.0%)	5 (62.5%)	5 (62.5%)
Age at diagnosis*		56 ± 10	61 ± 13	64 ± 8.0	62 ± 7.0
BMI* (kg/m^2)		21.5 ± 1.5	21.4 ± 1.4	21.3 ± 1.2	23.4 ± 2.4

*HC: Healthy control, CRC: Colorectal cancer. *Data are presented as mean ± standard deviation.*

Gut microbiota diversity in two sample collection datasets

Faith's phylogenetic diversity (Faith's PD) and Shannon's index were used to measure the related similarity of taxa and the combined species richness and evenness, respectively, in each sample. No significant differences were observed between the two datasets for either Faith's PD (p = 0.25) or Shannon's index (p = 0.94) (Figure 1A). Within each dataset, comparisons between the CRC and healthy groups also showed no statistically significant differences in Faith's PD (Dataset 1: p = 0.78; Dataset 2: p = 0.32). Similarly, no significant difference in Shannon's index was reported between the CRC and healthy groups in either dataset (Dataset 1: p = 0.92; Dataset 2: p = 0.1) (Figure 1B). Beta diversity was measured using unweighted UniFrac to evaluate the presence or absence of bacterial composition between communities, and weighted UniFrac to account for taxonomic abundance. A clear separation between the two datasets was observed with unweighted UniFrac (PERMANOVA: R^2 = 0.06; p = 0.02), whereas weighted UniFrac showed no significant separation between the two datasets (PERMANOVA: R^2 = 0.01; p = 0.97) (Figure 1A). In both datasets, CRC and healthy samples formed overlapping clusters, indicating limited separation in gut community composition. The PERMANOVA tests also yielded *p*-values above 0.05, confirming that no statistically significant differences existed between CRC and healthy groups in either dataset (Figure 1B).

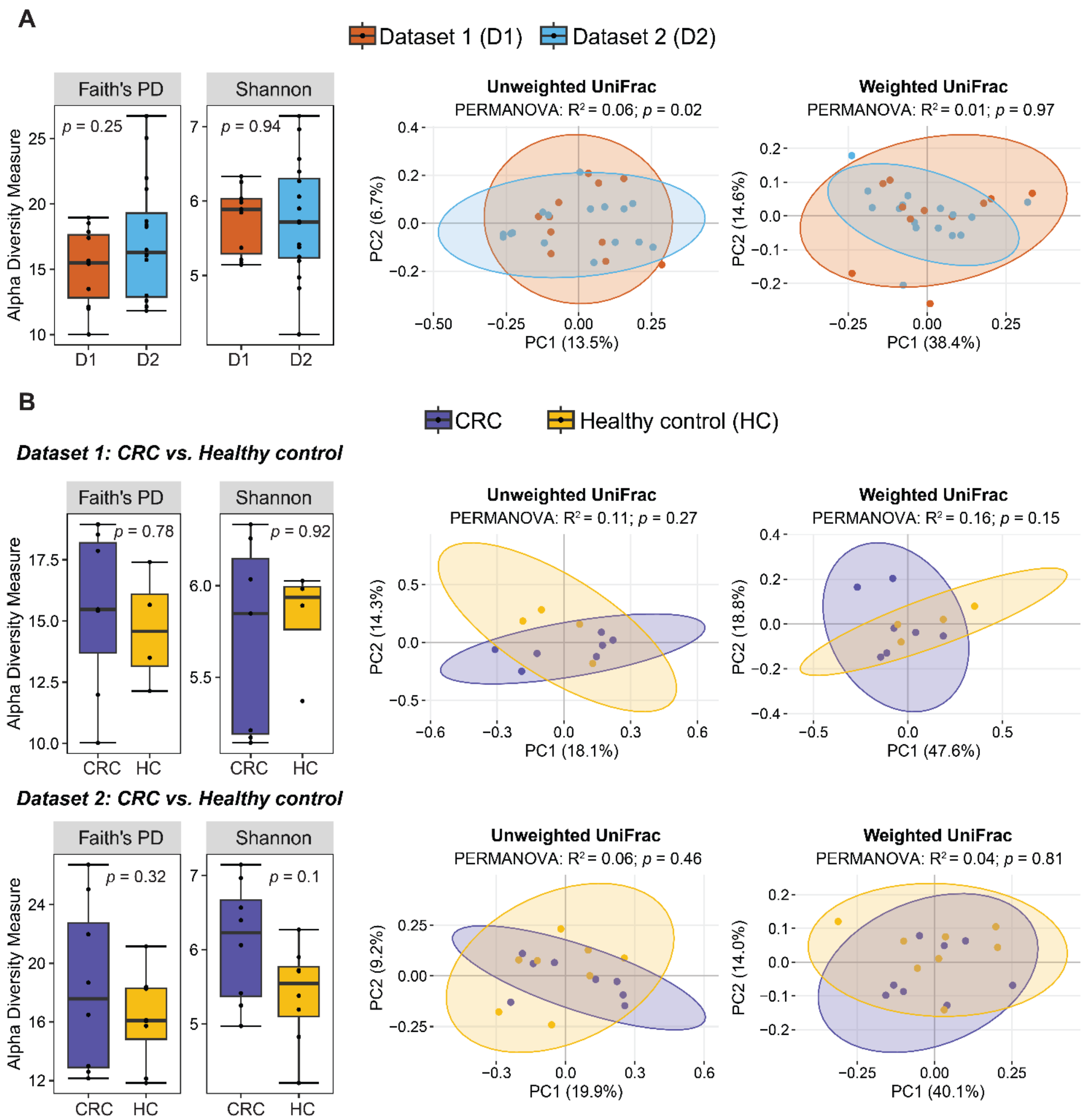

Figure 1. Comparison of gut microbiota diversity was performed between the two sample collection datasets (A), and between CRC and healthy groups within each dataset (B). The box plots represent the median and interquartile range of Faith's phylogenetic diversity (Faith's PD) and Shannon's index for CRC and healthy groups, with individual points representing each person's alpha diversity. Differences in alpha diversity indices between the two groups were evaluated by the Mann-Whitney U test. Unweighted and weighted UniFrac distances were used to assess beta diversity, which was illustrated using PCoA. Group differences in beta diversity were tested using PERMANOVA, with the following models: (A) distance ~ dataset and (B) distance ~ disease for each dataset. For all PERMANOVA analyses, the effect size (R^2) and statistical significance (*p*-values) are shown in the plots.

Differences in taxonomic composition between the two sample collection datasets

We identified 12 classified phyla, 79 classified families, and 247 classified genera across both datasets. The most abundant genera across all samples included *Blautia* (15.3% ± 9.7%), *Bacteroides* (8.6% ± 9.2%), *Faecalibacterium* (5.5% ± 4.5%), *Escherichia/Shigella* (4.2% ± 8.4%), and *Romboutsia* (3.1% ± 3.7%), which were prevalent in both the CRC and healthy groups. Additionally, the relative abundance of these genera, along with other detected taxa, varied between datasets and among individual subjects (Figure 2A). Next, differences in genus-level abundance between the two sample collection datasets were assessed using the Mann–Whitney U test, revealing significant differences in 6 classified genera (nominal p-values < 0.05). Among these, *Achromobacter* was more prevalent in Dataset 1, while 5 genera, including *Megamonas*, *Ligilactobacillus*, *Negativibacillus*, *Pseudomonas*, and *Lactobacillaceae* HT002, were more dominant in Dataset 2. Except for *Megamonas* and *Negativibacillus*, the remaining four genera remained significantly different after FDR correction (BH-adjusted p < 0.05). Notably, *Lactobacillaceae* HT002 and *Pseudomonas* were present in Dataset 2 but absent in Dataset 1, whereas *Achromobacter* was observed only in Dataset 1 (Figure 2B). These results suggested that, apart from inter-individual variation, technical factors substantially contributed to the differences in gut microbiota composition observed across microbiome studies.

Discrepancies in the differential abundance of genera between CRC patients and healthy individuals across two sample collection datasets

Differences in genus abundance between CRC patients and healthy controls were identified with the Mann-Whitney U test and ANCOM-BC2. In Dataset 1, a total of 12 genera showed significant differences between the CRC and healthy groups. Of these, three and 11 genera were identified by the Mann–Whitney U test and ANCOM-BC2, respectively. Two genera, *Peptostreptococcus* and *Parabacteroides*, were consistently detected by both methods, with *Peptostreptococcus* significantly increased and *Parabacteroides* significantly decreased in CRC compared with healthy controls. *Parvimonas* was identified as significantly decreased based on the Mann-Whitney U test (p = 0.023), while ANCOM-BC2 could not be applied because this taxon was entirely absent in the healthy group. Instead, a Fisher's exact test confirmed its group-specific presence in CRC samples (p = 0.015). None of the taxa remained significant in the Mann-Whitney U test after FDR correction. Similarly, most taxa also lost significance in ANCOM-BC2, with the exception of *Peptostreptococcus* (BH-adjusted p = 0.046) and *Fusobacterium* (BH-adjusted p = 0.009), which remained significant (Table 2).

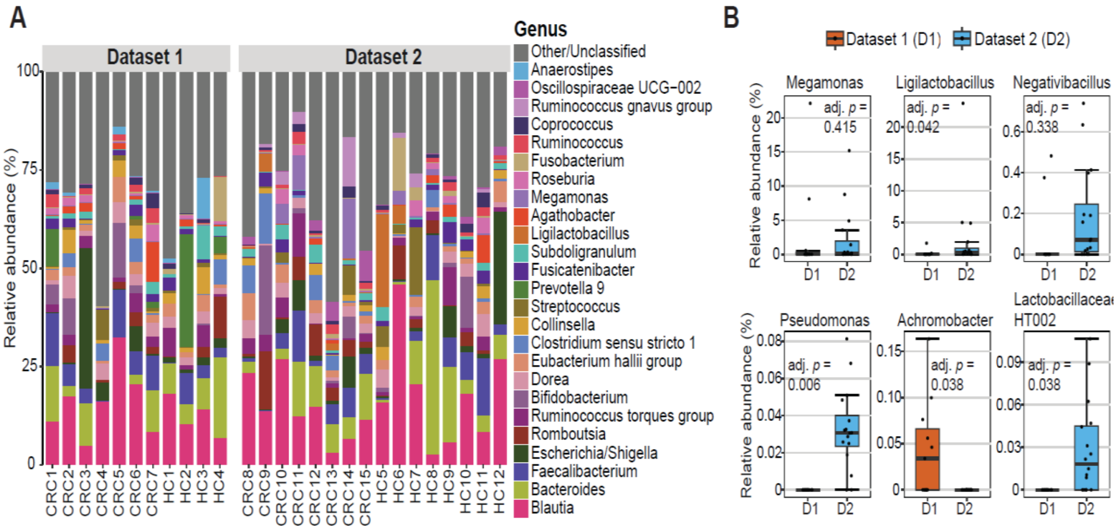

Figure 2. Comparison of gut taxonomic composition at the genus level between Dataset 1 and Dataset 2. (A) Bar plots displaying the relative abundance of microbial genera in CRC (colorectal cancer) and healthy control (HC) samples from both datasets. Only genera with a mean relative abundance greater than 1% are displayed, while low-abundance and unclassified taxa are grouped as "Other/Unclassified". (B) Differences in genus abundance between Dataset 1 and Dataset 2 were assessed using the Mann–Whitney U test. Only genera with *p*-values below 0.05 are included in the bar plots. Benjamini–Hochberg–adjusted *p*-values (adj. *p*) are displayed in the figures.

In Dataset 2, 13 genera differed significantly between the CRC and healthy groups, with 6 detected by the Mann–Whitney U test and 12 identified by ANCOM-BC2. *Parvimonas*, *Lactococcus*, and *Oscillospiraceae* UCG-003 were significantly elevated in the CRC group according to both methods. In contrast, *Lactiplantibacillus* was enriched in healthy individuals based on the Mann-Whitney U test (p = 0.013), and its complete absence in the CRC group was further supported by Fisher's exact test (p = 0.026). After FDR correction, however, only *Parvimonas* remained significant in ANCOM-BC2 (BH-adjusted p = 0.029) (Table 2). Among the significant genera detected in both datasets, *Parvimonas, Peptostreptococcus, Lactococcus, Gemella*, *Lachnospiraceae* CAG-56, and *Fusobacterium* were shared between Dataset 1 and Dataset 2. Four of these (*Parvimonas, Peptostreptococcus, Lactococcus,* and *Gemella*) were enriched in the CRC groups across both datasets, whereas *Fusobacterium* showed depletion. On the other hand, *Lachnospiraceae* CAG-56 exhibited opposite patterns between the two datasets (Table 2 and Figure 3).

Table 2. Differentially abundant genera between CRC patients and healthy individuals in Dataset 1 and Dataset 2.

Genus	**Mann-Whitney U test**		**ANCOM-BC2**			
	P-value	***Adjusted p-value***	***Log2FC***	***Standard error***	***P-value***	***Adjusted p-value***
Dataset 1						
_Parvimonas_*$	0.023	0.818	_Absence in the healthy group_			
_Peptostreptococcus_*$	0.033	0.818	2.663	0.406	0.001	0.046
_Lactococcus_*	0.372	0.909	2.555	0.579	0.012	0.167
_Gemella_*	0.058	0.818	1.758	0.574	0.018	0.167
Lachnospiraceae CAG-56*	0.562	0.950	2.121	0.510	0.009	0.167
_Fusobacterium_*	0.924	1.0	-5.019	0.567	< 0.001	0.009
Oscillospiraceae UCG-002	0.562	0.950	1.904	0.493	0.012	0.167
Ruminococcus	0.164	0.818	2.078	0.735	0.022	0.167
Desulfovibrio	0.766	1.0	-1.960	0.580	0.028	0.190
Holdemanella	0.536	0.950	-1.707	0.475	0.037	0.231
Parabacteroides$	0.024	0.818	-2.221	0.759	0.019	0.167
Bilophila	0.333	0.909	-2.029	0.545	0.014	0.167
Dataset 2						
_Parvimonas_*$	0.025	0.566	3.171	0.483	< 0.001	0.029
_Peptostreptococcus_*	0.056	0.694	2.665	0.523	0.004	0.116
_Lactococcus_*$	0.012	0.566	1.338	0.463	0.018	0.253
_Gemella_*	0.025	0.566	0.640	0.529	0.266	0.763
Lachnospiraceae CAG-56*	0.116	0.725	-1.800	0.572	0.020	0.253
_Fusobacterium_*	0.787	0.967	-3.246	0.730	0.002	0.099
Lactiplantibacillus$	0.013	0.566	_Absence in the CRC group_			
Lachnospiraceae UCG-010	0.424	0.947	1.196	0.507	0.043	0.394
Family XIII UCG-001	0.701	0.947	1.323	0.442	0.020	0.253
Oscillospiraceae UCG-003$	0.033	0.599	1.590	0.480	0.013	0.253
Oscillospiraceae NK4A214	0.298	0.913	1.682	0.574	0.022	0.253
Agathobacter	0.040	0.620	-1.218	0.814	0.160	0.635
Butyricicoccus	0.226	0.769	-1.585	0.687	0.042	0.394

Differences in genus abundance were assessed by Mann–Whitney U (relative abundance) and ANCOM-BC2 (absolute abundance). P-values are reported as nominal and Benjamini–Hochberg (BH)–adjusted values. Log2FC represents the log₂ fold change in CRC patients compared with healthy controls in ANCOM-BC2. *Genera detected in both datasets. $Genera identified as significant by both the Mann–Whitney U test and ANCOM-BC2, or absent in one group.

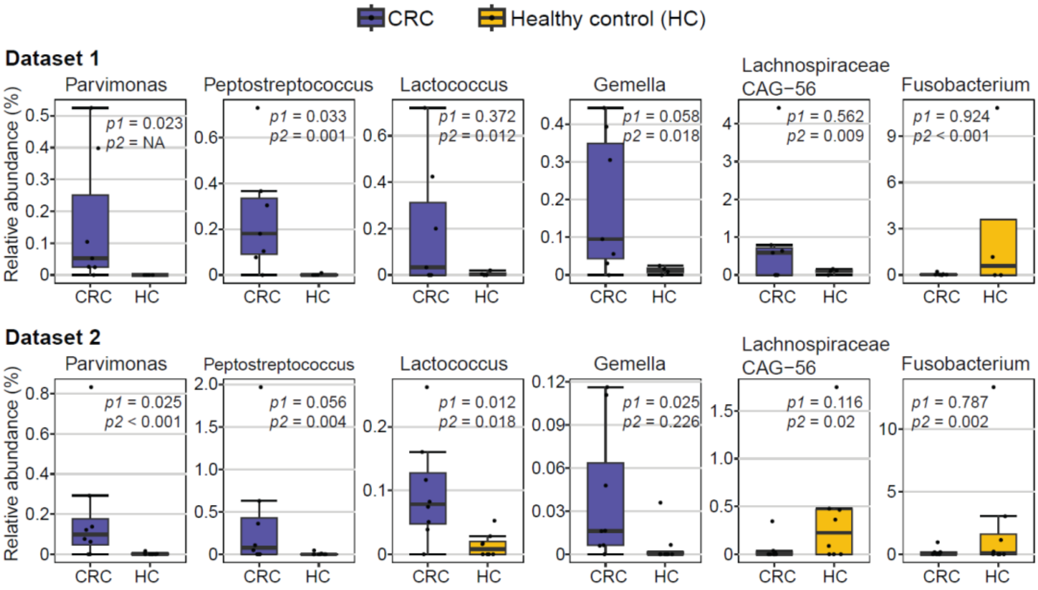

Figure 3. Box plots show the relative abundances for bacterial genera that significantly differ between CRC patients and healthy individuals that were identified in Dataset 1 and Dataset 2. Only genera that were detected in both datasets were included in the plots. Nominal *p*-values are reported from the Mann–Whitney U test (*p1*) and ANCOM-BC2 (*p2*). NA: not applicable as these taxa are absent in one group.

DISCUSSION

Through this pilot study, we confirmed differentiated gut microbial communities in CRC patients and healthy individuals, as well as the variability between the two sample collection datasets. This variability may partly explain the inconsistencies reported across different studies. In addition to regional factors like country or continent, and biological factors such as age, disease status, gender, and BMI, which significantly influence microbiome composition and disease-associated signatures as noted by Ghosh *et al.* (2020), our findings indicate that technical variables may also contribute to alterations in microbial diversity and composition. Indeed, both datasets in this study were recruited from the same hospital and proceeded with identical DNA extraction and sequencing protocols. Despite the absence of significant differences in alpha diversity (Faith's PD and Shannon's index) and beta diversity based on weighted Unifrac distance, the weighted Unifrac distance revealed notable distinctions between the two datasets. Because unweighted UniFrac reflects community membership rather than relative abundance, the observed pattern suggests that low-abundance taxa are more sensitive to technical or batch effects, whereas the lack of separation in weighted UniFrac indicates that overall abundance structure remained relatively stable across datasets. These findings support existing literature on the potential impact of DNA extraction batch and sequencing run batch on gut microbiota profiles (Crane *et al.*, 2022; Jokela *et al.*, 2023). By using the Mann–Whitney U test and ANCOM-BC2 to identify significant genera between CRC and healthy groups in each dataset, we found 12 and 13 significant genera in Dataset 1 and Dataset 2, respectively, prior to FDR correction. Among these, only 5 genera, including *Parvimonas, Peptostreptococcus, Lactococcus, Gemella,* and *Fusobacterium*, exhibited consistent trends across both sample collection datasets. Apart from sequencing bias, it is worth noting that individuals from Dataset 1 and Dataset 2 come from several cities in the Northern regions, such as Hanoi, Hai Phong, Nam Dinh, Hoa Binh, Lang Son, etc. Therefore, the regional factor may have contributed to variations in gut microbiota as in agreement with a previous report (Ghosh *et al.*, 2020). Another study by Suchandra *et al.* (2024) examined the gut microbiota of 200 healthy individuals living in a rural area in India and found an association between microbiota composition and lifestyle factors, including diet, smoking status, water intake, and quality of sleep. Since our participants experienced different lifestyles, this could be another biological factor for variation between the two datasets. In addition to biological and technical factors, bioinformatic pipeline choices may also influence microbiome results (Siegwald *et al.*, 2019). Steps such as quality filtering thresholds, ASV inference, taxonomic assignment databases, and normalization strategies can affect diversity metrics and differential abundance outcomes. Although samples from both datasets in this study were analyzed using the same analytical pipeline to minimize analytical bias, batch-specific noise cannot be entirely excluded.

Of the genera shared by both datasets, *Parvimonas* was the only genus that remained significant in Dataset 2 using ANCOM-BC2 after FDR correction, and it was also absent in healthy controls in Dataset 1. This reproducibility across independently sequenced cohorts strengthens confidence that *Parvimonas*, particularly *Parvimonas micra*, represents a robust CRC-associated signal rather than a dataset-specific artifact. These findings are consistent with those reports from international meta-analyses (Thomas *et al.*, 2019; Tito *et al.*, 2024) and with our previous observations in Vietnamese CRC patients (Nhung *et al.*, 2023; Nhung *et al.*, 2024), further supporting its potential role in colorectal carcinogenesis. Additionally, the potential use of *Parvimonas micra* as a non-invasive fecal biomarker was evaluated using quantitative PCR in fecal samples from a Sweden cohort of 238 CRC patients and 94 controls, yielding specificity and sensitivity values of 87.3% and 60.5%, respectively, for CRC detection (Löwenmark *et al.*, 2020).

A limitation of this study was that we could not evaluate the contribution of each technical variable, such as DNA extraction batch or sequencing run batch, to microbiome variations in CRC patients and healthy controls separately. It is important to repeat the DNA extraction and sequencing protocols on the same set of samples to ensure consistency and reliability of results throughout a microbiome study. Second, the cohort sizes (11 individuals in Dataset 1 and 16 in Dataset 2) are modest for 16S rRNA case-control analyses, limiting statistical power, particularly for detecting modest effects after adjusting for multiple testing. In fact, a large sample size is essential in gut microbiome studies to ensure adequate statistical power (Agronah and Bolker, 2025). Moreover, this limited sample size precluded adjustment for potential confounders such as age, sex, and BMI. The absence of such adjustments (e.g., via ANCOM-BC2's covariate options) may weaken the robustness of the conclusions, even though no significant group differences were observed for these confounders. Power calculations indicate that detecting differences in microbial diversity and taxonomic abundance with a medium effect size (Cohen's d = 0.5) at a two-tailed significance level of α = 0.05 between two independent groups using the Mann–Whitney U test would require 134 samples (67 per group) to achieve 80% power, as estimated using G*Power 3 (Faul *et al.*, 2007). Therefore, future studies should include larger and equally sized sample groups to enable more robust and meaningful comparisons between different sample collection datasets. Despite its limitations, this study offers valuable knowledge to researchers in evaluating potential biomarkers for diagnostic purposes.

CONCLUSION

Our comparative analysis of two distinct sample collection datasets provides insights into both shared and dataset-specific microbial signatures associated with CRC. These findings highlight the importance of cross-dataset validation when evaluating candidate microbial biomarkers. While individual datasets may yield variable results due to technical or sampling differences, taxa that show consistent enrichment across datasets, such as *Parvimonas,* are more likely to represent biologically meaningful and clinically relevant biomarkers

ACKNOWLEDGMENTS

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 108.04-2021.22.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Andrews S. (2010). FastQC a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Agronah M. and Bolker B. (2025). Investigating statistical power of differential abundance studies. *PLoS One, 20*(4), e0318820. https://doi.org/10.1371/journal.pone.0318820

Bray F., Laversanne M., Sung H., Ferlay J., Siegel R. L., Soerjomataram I., *et al.* (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 74*(3), 229-263. https://doi.org/10.3322/caac.21834

Bull M. J., and Plummer N. T. (2014). Part 1: The human gut microbiome in health and disease. *Integrative Medicine (Encinitas), 13*(6), 17-22.

Cole J. R., Wang Q., Cardenas E., Fish J., Chai B., Farris R. J., *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research, 37* (Database issue), D141-145. https://doi.org/10.1093/nar/gkn879

Crane R. J., Parker E. P. K., Fleming S., Gwela A., Gumbi W., Ngoi J. M., *et al.* (2022). Cessation of exclusive breastfeeding and seasonality, but not small intestinal bacterial overgrowth, are associated with environmental enteric dysfunction: A birth cohort study amongst infants in rural Kenya. *EClinicalMedicine, 47*, 101403. https://doi.org/10.1016/j.eclinm.2022.101403

Duan B., Zhao Y., Bai J., Wang J., Duan X., Luo X., *et al.* (2022). Colorectal cancer: An overview. In J. A. Morgado-Diaz (Ed.), *Gastrointestinal Cancers*. Brisbane (AU).

Faul F., Erdfelder E., Lang A. G., and Buchner A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior*

Research Methods, 39(2), 175-191. https://doi.org/10.3758/bf03193146

Flemer B., Lynch D. B., Brown J. M., Jeffery I. B., Ryan F. J., Claesson M. J.*, et al.* (2017). Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut, 66*(4), 633-643. https://doi.org/10.1136/gutjnl-2015-309595

Ghosh T. S., Das M., Jeffery I. B., and O'Toole P. W. (2020). Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *Elife, 9*. https://doi.org/10.7554/eLife.50240

Hou K., Wu Z. X., Chen X. Y., Wang J. Q., Zhang D., Xiao C.*, et al.* (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy, 7*(1), 135. https://doi.org/10.1038/s41392-022-00974-4

Jokela R., Ponsero A. J., Dikareva E., Wei X., Kolho K. L., Korpela K.*, et al.* (2023). Sources of gut microbiota variation in a large longitudinal Finnish infant cohort. *EBioMedicine, 94*, 104695. https://doi.org/10.1016/j.ebiom.2023.104695

Le N. T. and Dao H. V. (2020). Colorectal cancer in Vietnam. *Colorectal Cancer*. IntechOpen. https://doi.org/10.5772/intechopen.93730

Lin H. and Peddada S. D. (2024). Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. *Nature Methods, 21*(1), 83-91. https://doi.org/10.1038/s41592-023-02092-7

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. *Scientific Reports, 10*(1), 15250. https://doi.org/10.1038/s41598-020-72132-1

Nguyen B. N., Nguyen L. T. N., Trinh D. T. M., Nguyen H. T., and Tran T. T. T. (2025). Preliminary insights into the gut microbiota of patients with rheumatoid arthritis in Vietnam. *PeerJ, 13*, e20521. https://doi.org/10.7717/peerj.20521

Nhung P. T. T., Hang L. T. T., and Tam T. T. T. (2023). Preliminary assessment of gut microbiota diversity in colorectal cancer patients using 16S rRNA sequencing. *Journal of 108 - Clinical Medicine and Phamarcy, 18*(8). https://doi.org/10.52389/ydls.v18i8.2095.

Nhung P. T. T., Le H. T. T., Nguyen Q. H., Huyen D. T., Quyen D. V., Song L. H.*, et al.* (2024). Identifying fecal microbiota signatures of colorectal cancer in a Vietnamese cohort. *Frontiers in Microbiology, 15*, 1388740. https://doi.org/10.3389/fmicb.2024.1388740

Sender R., Fuchs S., and Milo R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biology, 14*(8), e1002533. https://doi.org/10.1371/journal.pbio.1002533

Shen Y., Fan N., Ma S. X., Cheng X., Yang X., and Wang G. (2025). Gut microbiota dysbiosis: pathogenesis, diseases, prevention, and therapy. *MedComm, 6*(5), e70168. https://doi.org/10.1002/mco2.70168

Siegel R. L., Miller K. D., Goding Sauer A., Fedewa S. A., Butterly L. F., Anderson J. C.*, et al.* (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians, 70*(3), 145-164. https://doi.org/10.3322/caac.21601

Siegwald L., Caboche S., Even G., Viscogliosi E., Audebert C., and Chabe M. (2019). The impact of bioinformatics pipelines on microbiota studies: does the analytical "microscope" affect the biological interpretation? *Microorganisms, 7*(10), 393. https://doi.org/10.3390/microorganisms7100393

Suchandra G., Manisha K., and Sandhya K. (2024). Exploring the gut microbiota of rural region of Haryana (India): sociodemographic, socioeconomic factors and lifestyle. *Clinical Epidemiology and Global Health* 30**,** 101806. https://doi.org/10.1016/j.cegh.2024.101806.

Thomas A. M., Manghi P., Asnicar F., Pasolli E., Armanini F., Zolfo M., *et al.* (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine, 25*(4), 667-678. https://doi.org/10.1038/s41591-019-0405-7

Tito R. Y., Verbandt S., Aguirre Vazquez M., Lahti L., Verspecht C., Llorens-Rico V.*, et al.* (2024). Microbiome confounders and quantitative profiling challenge predicted microbial targets in colorectal cancer development. *Nature Medicine, 30*(5), 1339-1348. https://doi.org/10.1038/s41591-024-02963-2

Wirbel J., Pyl P. T., Kartal E., Zych K., Kashani A., Milanese A.*, et al.* (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine, 25*(4), 679-689. https://doi.org/10.1038/s41591-019-0406-6

Wirbel J., Pyl P. T., Kartal E., Zych K., Kashani A., Milanese A.*, et al.* (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine, 25*(4), 679-689. https://doi.org/10.1038/s41591-019-0406-6

Yu J., Feng Q., Wong S. H., Zhang D., Liang Q. Y., Qin Y.*, et al.* (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut, 66*(1), 70-78. https://doi.org/10.1136/gutjnl-2015-309800