



OPEN Nanopore full length 16S rRNA gene sequencing increases species resolution in bacterial biomarker discovery

Pablo Aja-Macaya^{1,2,6}, Kelly Conde-Pérez^{1,6}, Noelia Trigo-Tasende^{1,4,6}, Elena Buetas³, Mohammed Nasser-Ali¹, Paula Nió¹, Soraya Rumbo-Feal¹, Susana Ladra², Germán Bou^{1,5}, Alex Mira³, Juan A. Vallejo^{1✉} & Margarita Poza^{1,4}

Discovery of disease-related bacterial biomarkers could be a useful approach for early prevention or diagnosis of various afflictions, such as colorectal cancer. This typically involves analyzing small regions of the 16S rRNA gene (e.g. V3V4) through short-read technologies like Illumina, obtaining genus-level results. However, recent developments in third-generation sequencing, such as Oxford Nanopore Technologies (ONT)'s new R10.4.1 chemistry and its improved basecalling models, are beginning to allow for a more complete and accessible species-level analysis through full-length 16S rRNA gene sequencing (spanning regions V1-V9). Thus, the goal of this study was to compare and evaluate both approaches, using colorectal cancer biomarker discovery as a representative case. This was achieved through the analysis of feces from 123 subjects, comparing both methods (Illumina-V3V4 with DADA2 and QIIME2 vs. ONT-V1V9 with Emu), multiple Dorado basecalling models (fast, hac and sup) and multiple databases (SILVA vs. Emu's Default database). Basecalling models broadly resulted in similar taxonomic output, but had significantly higher observed species and different taxonomic identification the lower the basecalling quality (p -value<0.05). Database choice with Emu influenced the identified species greatly, with Emu's Default database obtaining significantly higher diversity and identified species than SILVA (p -value<0.05). However, it overconfidently classified at times what should be an unknown species as the closest match due to its database structure. Bacterial abundance between Illumina-V3V4 and ONT-V1V9 at the genus level correlated well ($R^2 \geq 0.8$). Nanopore sequencing identified more specific bacterial biomarkers for colorectal cancer than those obtained with Illumina, such as *Parvimonas micra*, *Fusobacterium nucleatum*, *Peptostreptococcus stomatis*, *Peptostreptococcus anaerobius*, *Gemella morbillorum*, *Clostridium perfringens*, *Bacteroides fragilis* and *Sutterella wadsworthensis*. Prediction of colorectal cancer through manual feature selection and machine learning resulted in an AUC of 0.87 with 14 species or 0.82 with just 4 species (*P. micra*, *F. nucleatum*, *B. fragilis* and *Agathobaculum butyriciproducens*). Full 16S rRNA V1V9 sequencing through Oxford Nanopore and its new R10.4.1 chemistry achieved accurate species-level bacterial identification, facilitating the discovery of more precise disease-related biomarkers and increasing the taxonomic fidelity of future microbiome analyses.

Keywords Illumina, Oxford Nanopore, Colorectal cancer, 16S rRNA, Biomarkers, Gut microbiome, Metabarcoding, V1V9

¹Grupo de Investigación en Microbiología, Instituto de Investigación Biomédica de A Coruña (INIBIC), Complejo Hospitalario Universitario de A Coruña (CHUAC), Sergas, Universidade de A Coruña (UDC), As Xubias, 15006 A Coruña, Galicia, Spain. ²Database Laboratory (LBD), CITIC, Universidade da Coruña (UDC), Campus de Elviña, 15071 A Coruña, Galicia, Spain. ³Genomic and Health Department, FISABIO Foundation, Center for Advanced Research in Public Health, Benimaclet, 46020 Valencia, Comunidad Valenciana, Spain. ⁴Grupo de Microbioma y Salud, Centro Interdisciplinar de Química e Bioloxía (CICA), Facultad de Ciencias, Universidade da Coruña (UDC), Campus da Zapateira, 15008 A Coruña, Galicia, Spain. ⁵CIBER de Enfermedades Infecciosas (CIBERINFEC-ISCIII), As Xubias, 15006 A Coruña, Galicia, Spain. ⁶Pablo Aja-Macaya, Kelly Conde-Pérez and Noelia Trigo-Tasende contributed equally to this work. ✉email: juan.andres.vallejo.vidal@sergas.es

Abbreviations

ONT	Oxford Nanopore Technologies
16S rRNA	16S ribosomal RNA
CRC	Colorectal cancer
gFOBT	Guaiac-based fecal occult blood test
FIT	Fecal immunochemical test
ETBF	Enterotoxigenic strains of <i>Bacteroides fragilis</i>
qPCR	Quantitative PCR
NGS	Next generation sequencing
TGS	Third generation sequencing
AUC	Area under the receiver operating characteristic curve

Prior to the advent of the Next Generation Sequencing (NGS) and Third Generation Sequencing (TGS) platforms, the analysis of microbial compositions relied on traditional bacteriological procedures. However, the high proportion of viable but non-culturable or difficult-to-cultivate microorganisms in the biosphere complicates the analysis of microbiome compositions¹. As a matter of fact, nowadays more than 99% of the potentially 10^{11} – 10^{12} microbial species on Earth remain undiscovered². Nevertheless, high-throughput sequencing methods enabled, for the first time in history, the possibility to identify, without culturing, practically ~100% of the genetic material of all microorganisms in a given sample. For this reason, nowadays 16S rRNA gene sequencing represents the gold-standard method for microbiome studies, standing as the fundamental driver behind the increasing volume of studies shaping environmental and animal microbial communities^{3–6}.

Multiple studies have investigated the involvement of these communities in human diseases, such as colorectal cancer (CRC)^{7–12}, which stands out as one of the most commonly diagnosed tumors all over the world¹³. Particularly in Spain (Europe), CRC has the highest incidence among tumors, with over 40,000 new cases throughout 2023¹⁴. Moreover, CRC is the second leading cause of cancer mortality, with more than 15,000 deaths reported, according to the Spanish Association Against Cancer latest annual report¹⁴. The CRC assessment programs implemented worldwide predominantly utilize a guaiac-based fecal occult blood test (gFOBT) or a fecal immunochemical test (FIT), which are conducted biennially, given that colon cancer development is a slow and gradual process^{15–17}. As a result, the implementation of these screenings allows to: (a) increase the number of annual diagnostics, (b) improve the CRC survival statistics and (c) achieve lower mortality rates¹⁷. However, it is important to highlight that blood detection in fecal matter does not necessarily indicate the presence of a carcinoma, and that a colonoscopy (an invasive technique) is required to confirm or deny the FOBT or FIT result. Therefore, there is an urgent need to develop novel screening methods, more specific and non-invasive, in order to enhance the detection of intestinal lesions even at precancerous stages.

Interestingly, it is widely recognized that several gut microbes could develop harmful effects on colonocytes' integrity and homeostasis^{18,19}. Multiple in vitro and in vivo experiments demonstrated over the last years that specific bacteria, such as *pks*⁺ *Escherichia coli*, enterotoxigenic strains of *Bacteroides fragilis* (ETBF), *Parvimonas micra* or *Fusobacterium nucleatum* are extremely related with the colorectal tumorigenesis process^{20–25}. More precisely, microorganisms have the potential to directly and indirectly: (a) affect the epithelial permeability (e.g., by modulating the expression of tight junction proteins)^{26,27}, (b) promote chronic tissue inflammation (e.g., through the secretion of toxins, enhancing bacterial adherence to epithelial cells)²⁸, (c) deregulate host anti-tumoral immune activities (e.g., *Fusobacterium nucleatum* through the production of certain adhesins such as Fap2, affecting the function of natural killer T cells)²⁵, (d) trigger chromosomal instability and DNA mutagenesis (e.g., ETBF via the BFT toxin promotes the hyperproduction of reactive oxygen species or *E. coli pks*⁺, a pathogen which induces a characteristic mutational signature via the colibactin molecule)^{21,29,30}, (e) alter eukaryotic DNA methylation patterns (e.g., *Parvimonas micra* promotes hypermethylation in genes related to cytoskeleton regulation and tumor suppression)^{24,28,29} and (d) modulate several cell-signaling pathways (e.g., E-cadherin/ β -catenin, TLR4/MYD88/NF- κ B or SMO/RAS/p38 MAPK)^{18,31,32}. In turn, these deleterious activities lead to hyperproliferation, senescence, tumor growth and invasiveness, additionally inducing the initial stages of the metastatic process^{18,20,21,23,26–29,33,34}. Consequently, tumor progression and effectiveness of anti-cancer therapies could be directly correlated with the gut bacteriome established in each patient^{26,35,36}. Accordingly, multiple studies have put forth specific non-invasive microbiome-derived biomarkers for CRC during the preceding decade. These investigations have demonstrated the potential to differentiate individuals with malignant dysplasias from those without lesions through a simple analysis of their fecal material, utilizing 16S ribosomal ribonucleic acid (16S rRNA) gene sequencing or quantitative PCR (qPCR) procedures^{7–12}.

Multiple 16S rRNA sequencing options exist to study the microbial communities involved in diseases such as CRC and to generate disease-related bacterial biomarkers. For example, Oxford Nanopore Technologies (ONT) and its innovative long-read sequencing method, which can generate >10 kb sequences directly from native DNA^{37–39}, has positioned itself as a compelling alternative to short-read technologies such as Illumina or other long-read technologies like PacBio for 16S rRNA analysis³⁷. In comparison to Illumina, which is restricted to sequence small areas of the 16S rRNA such as V3V4 (~400 nt) and identification mostly at the genus level, ONT can obtain the full region^{6,39} (~1500 nt, V1V9) and identify reads at the species level more consistently. Additionally, the barrier to entry is smaller with ONT, as the starting sequencers cost much less, which is an advantage in less well-resourced environments. However, the main limitation is the relatively higher error rate ONT reads achieve in comparison to other technologies, although its chemistry and basecallers, such as Dorado, are constantly being improved and will presumably get to a similar quality eventually, having recently achieved Q20 and even Q25+ in a small proportion of reads (in contrast to Illumina and PacBio's consistent Q30+). Putting things into perspective, Q20 indicates an error rate of 1% (Q15 = ~3% error rate), which is the threshold

needed to confidently assign an OTU (Operational Taxonomic Unit) to a specific species in full length 16S rRNA⁴⁰.

This limitation has led to different bioinformatic approximations used for each technology. While Illumina and PacBio reads' quality allows for the creation of precise ASVs (Amplicon Sequence Variant), typically through DADA2^{41,42}, this method is not prepared for the current quality profile of ONT reads, which has caused the elaboration of various tools such as Emu⁴³ or NanoClust⁴⁴. Nonetheless, future advancements in ONT's technology might allow for the use of procedures typically reserved for Illumina or PacBio.

Thus, the present study evaluates the potential of ONT to improve upon Illumina's prevailing 16S rRNA analysis through the use of a considerable cohort of subjects (n = 123), examined previously with Illumina¹² and composed of colorectal cancer patients and their healthy counterparts, in order to obtain more precise CRC biomarkers. Two different sequencing systems and approaches, Illumina-V3V4 and ONT-V1V9 (R10.4.1 chemistry), three Dorado basecalling models (fast, hac or High Accuracy, sup or Supper-accurate; v4.1.0) and two databases (SILVA and Emu's Default database) were compared.

Methods

Recruitment of participants and fecal sampling

All volunteers provided informed signed consent prior to the initiation of the sample collection phase, which was performed in the University Hospital of A Coruña (HUAC; Galicia, Spain). Strict adherence to clinical guidelines and regulations was maintained throughout the recruitment period (Research Ethical Committee of Galicia, Spain: code CEIm-G 2018/609). A total of 93 CRC diagnosed subjects were enrolled in the project between 2019 and 2022. Certain inclusion criteria were followed as previously described¹²: (a) no antibiotic treatment within the last month, (b) no infectious disease, (c) no chemotherapy and/or radiotherapy treatments prior to sample collection, (d) no genetic predisposition to CRC development, (e) no intestinal inflammatory disorders, (f) no immunological diseases, (g) no medical history of transplantation, (h) not currently undergoing immunosuppressive treatment. Moreover, a total of 30 CRC cancer-free patients' companions/couples were asked to participate in the study, meeting the same inclusion requirements as the CRC group. A preliminary interview with each volunteer was conducted to gather individual data (e.g.: age, sex, weight or height, among others) and lifestyle habits (e.g.: dietary patterns or physical activity, among others). Samples (n=123), each containing approximately 20 mL of fecal material, were self-collected by each patient at home and preserved in 10 mL of RNeasy lysis reagent (Thermo Fisher Scientific, Waltham, MA, USA) in cold storage (~4°C) for 1–2 days, and were then stored at -80°C until DNA extraction. Samples were collected normally months after the colonoscopy, ensuring their microbiome profile could recover, just right before the surgery and the pre-surgery diet started. In rare cases, where surgery was urgent, they were collected at least 15 days before surgery.

DNA extraction

As previously outlined^{12,45} fecal samples underwent a brief pre-processing protocol prior to the DNA extraction procedure. The MasterPure™ Complete DNA/RNA Purification Kit (Epicentre, USA) was used in accordance with the manufacturer's instructions.

16S rRNA metabarcoding sequencing

Illumina (MiSeq™) 16S rRNA V3-V4

Two highly variable regions within the 16S rRNA gene (V3–V4) were selectively amplified through PCR, employing 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG as forward primer and 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC as reverse primer^{12,45}. Nuclease-free water was included as negative control in each PCR reaction to avoid bacterial contaminations. Subsequently, libraries were constructed by following the Illumina 16S Metagenomic Sequencing Library Preparation protocol (Illumina, San Diego, CA, USA). Library concentration was measured using a Qubit dsDNA HS Assay Kit (Invitrogen, USA) and a Qubit 2.0 fluorometer (Invitrogen, USA).

Libraries were pooled and diluted to a final concentration of 10 pM and mixed with 20% of 10 pM PhiX control (Illumina, USA). Samples were finally sequenced using a MiSeq Reagent Kit v3 (600 cycles) (Illumina, USA) and a MiSeq platform (Illumina, USA).

Oxford nanopore (MinION™) 16S rRNA V1-V9

Complete bacterial 16S rRNA hypervariable regions (V1–V9) were amplified, using the following forward and reverse primers: 5' AGMGTTYGATYMYGGCTCAG and 5' TACGGYTACCTTGTTACGACTT, respectively. Negative controls (nuclease-free water) were included to avoid contaminations. For each PCR reaction, 200 fmol of fecal DNA was used. Afterwards, DNA Oxford Nanopore libraries were constructed by following the Native Barcoding Kit 96 (SQK-NBD114.96) protocol (Oxford Nanopore, Oxford, UK). In the same way as Illumina libraries, DNA concentration was assessed by using a Qubit 2.0 fluorometer with the corresponding Qubit dsDNA HS Assay Kit (Invitrogen, USA).

Pooled libraries were loaded onto R10.4.1 Flow Cells (FLO-MIN114) and sequenced for 72 h following the manufacturer's instructions, detailed in the Native Barcoding Kit 96 protocol (Oxford Nanopore, Oxford, UK). Two sequencing runs were performed, with an average of 63 samples in each one.

Bioinformatic analysis

Oxford Nanopore Technologies (ONT) reads were first basecalled using Dorado duplex (v. 0.5.3)⁴⁶ with three models: fast, hac and sup (dna, r10.4.1, e8.2, 400bps, v4.1.0). The resulting reads were then separated into simplex, which were demultiplexed using Dorado (kit SQK-NBD114-96), and duplex reads ("consensus" of two parental simplex reads). Afterwards, duplex reads in which parental reads had different barcodes or which

lengths differed more than 50% were removed. The resulting reads were quality controlled with chopper (v. 0.5.0)⁴⁷, trimming 20 nt from the front and back and selecting reads between 1200–1900 nt with a specific minimum average quality score, depending on the basecalling model (fast: Q7; hac and sup: Q12). Additionally, Duplex Tools (v. 0.2.9)⁴⁸ was used to detect and remove reads with mid-strand adapters. Host contamination was assessed with Kraken2 using its Standard 64Gb database (<https://benlangmead.github.io/aws-indexes/k2>)⁴⁹. Lastly, reads were identified using Emu (v. 3.4.5)⁴³ with its Default database (rrnDB v. 5.6⁵⁰ combined with NCBI 16S RefSeq^{51,52}) and SILVA (v. 138.1)^{53,54}.

Illumina reads were analyzed through QIIME 2 (v. 2021.11)⁵⁵, where DADA2⁴¹ was used to trim, denoise, correct sequencing errors and remove chimeras on a per sequencing run basis, producing Amplicon Sequence Variants (ASVs), which were then classified using a feature classifier created with RESCRIPt (v. 2021.11.0)⁵⁶ and SILVA (v. 138.1). Additionally, contamination in these ASVs was also assessed with Kraken2's Standard 64Gb database.

Results from both ONT and Illumina were merged and analyzed in R (v. 4.2.0)⁵⁷, mainly through Phyloseq (v. 1.42.0)⁵⁸ for data management, ANCOM-BC (v. 2.0.1)⁵⁹ for differential abundance analysis (prevalence cutoff of 10%, adjusting significance by Holm-Bonferroni⁶⁰) and microbiome (v. 1.20.0)⁶¹ for centered log-ratio abundance normalization (CLR). In order to assess β -diversity differences, a PERMANOVA analysis through adonis2⁶², using a multi-dimensional scaling (MDS) and the Jensen-Shannon distance (JSD), was performed. Additionally, pairwise comparisons were conducted using Wilcoxon rank-sum tests (WRST), adjusting significance for multiple comparisons using Holm-Bonferroni. Significance values across analyses are represented as * ($p \leq 0.05$), ** ($p \leq 0.01$) or *** ($p \leq 0.001$).

For biomarker identification, automated feature selection was performed using the Boruta algorithm⁶³, with two prevalence cutoffs of 10% and 30%. Additionally, combinations of manually selected features were tested using a Random Forest machine learning algorithm and evaluated using the leave-one-out cross-validation method. The performance of the model was expressed through the area under the receiver operating characteristic curve (AUC) value.

Results

Quality control of Illumina reads resulted in a median of 32104 reads per sample with 90% of those being Q30 and 95% being Q20, generating ASVs with a median length of 418 nt. Meanwhile, quality control of ONT reads provided a total median of ~109k reads per sample and median length of 1480 nt. Median average quality varied across basecalling models, with the sup model achieving Q18 and some reads above Q30 (Fig. 1A). Duplex rate for sup was on average $8 \pm 2.73\%$, out of which $3.76 \pm 1.61\%$ were filtered. Rarefaction curves were closed at the species level on both ONT and Illumina (data not shown).

Comparison of ONT-V1V9 samples regarding β -diversity analysis (MDS and Jensen-Shannon Divergence) revealed no differences between basecalling models (Fig. 1B, p -value ≥ 0.8 in PERMANOVA), but did show significant differences between databases (p -value < 0.001 in PERMANOVA). Analysis of α -diversity was significantly different for the fast model, showing higher values of observed features (Fig. 1C, p -value < 0.05 in WRST). Additionally, Emu's Default database also resulted in significantly higher observed features when compared to SILVA (Fig. 1C, p -value < 0.05 in WRST). Moreover, significant differences were observed in the percentage of reads identified differently in respect to the sup model at each taxonomic level (p -value < 0.001 in WRST). For example, at species level, using the SILVA database, the identification of fast reads differed a median of 9.3% from sup, while hac differed 3.47% (Supplementary Fig. S1). Differences at higher levels such as family were smaller, having a median of 3.67% and 1.20% for fast and hac, respectively. Interestingly, Emu's Default database had higher standard deviation compared to SILVA in this measurement. Consequently, the sup model (Q18) was chosen for the following analyses.

When comparing Illumina-V3V4 and ONT-V1V9 with the sup basecalling model, both using the SILVA database, α -diversity at the genus level was not influenced, as shown in Fig. 2A. Regarding β -diversity, samples overlapped at the genus level and not at the species level (Fig. 2B and Fig. 2C), although PERMANOVA analysis indicated significantly different results in both cases (p -value < 0.001). Normalized CLR abundance at the genus level correlated well on average between Illumina-V3V4 and ONT-V1V9 (Pearson correlation: ≥ 0.8 , Fig. 2D), showing few taxa present in only one of the two approaches, which in most cases was due to slight differences in the taxonomy caused by the use of different classifiers. The percentage of feature counts identified as a known species (e.g. at species level and not as uncultured, unidentified, unclassified, Taxon NA...) varied significantly (p -value < 0.001 in WRST), obtaining a median of 16.75% with Illumina-V3V4, 27.74% with ONT-V1V9 (SILVA) (Supplementary Fig. S2) and, interestingly, 100% with ONT-V1V9 (Default).

In Fig. 3A the abundance of three important genera (*Fusobacterium*, *Parvimonas* and *Peptostreptococcus*), potential indicators of the presence of colorectal tumors along the large bowel, is shown across technologies (Illumina-V3V4 and ONT-V1V9) and databases (Default Emu database and SILVA). In all cases these genera are significantly more abundant in fecal samples from CRC patients than in the control group (p -value < 0.001 in WRST) and their abundances between approaches are similar. The percentage of samples with these taxa can be seen in Fig. 3B, with very similar proportions between methods except in the case of ONT-V1V9 (SILVA) for *Parvimonas*, where false positives were observed in controls (these occurrences were confirmed manually, identified as *Parvimonas* sp., not *P. micra*, at the species level).

When studying differential abundance analysis through ANCOM-BC (Fig. 4A, B) and differences in CLR abundance (Fig. 4C) using ONT-V1V9, both databases concur in multiple CRC biomarkers, such as *Parvimonas micra*, *Fusobacterium nucleatum*, *Peptostreptococcus stomatis*, *Peptostreptococcus anaerobius*, *Gemella morbillorum*, *Sutterella wadsworthensis*, *Clostridium perfringens*, *Bacteroides fragilis* and *Dialister pneumosintes*. In contrast, biomarkers for healthy controls are not shared between databases, with Emu's Default database

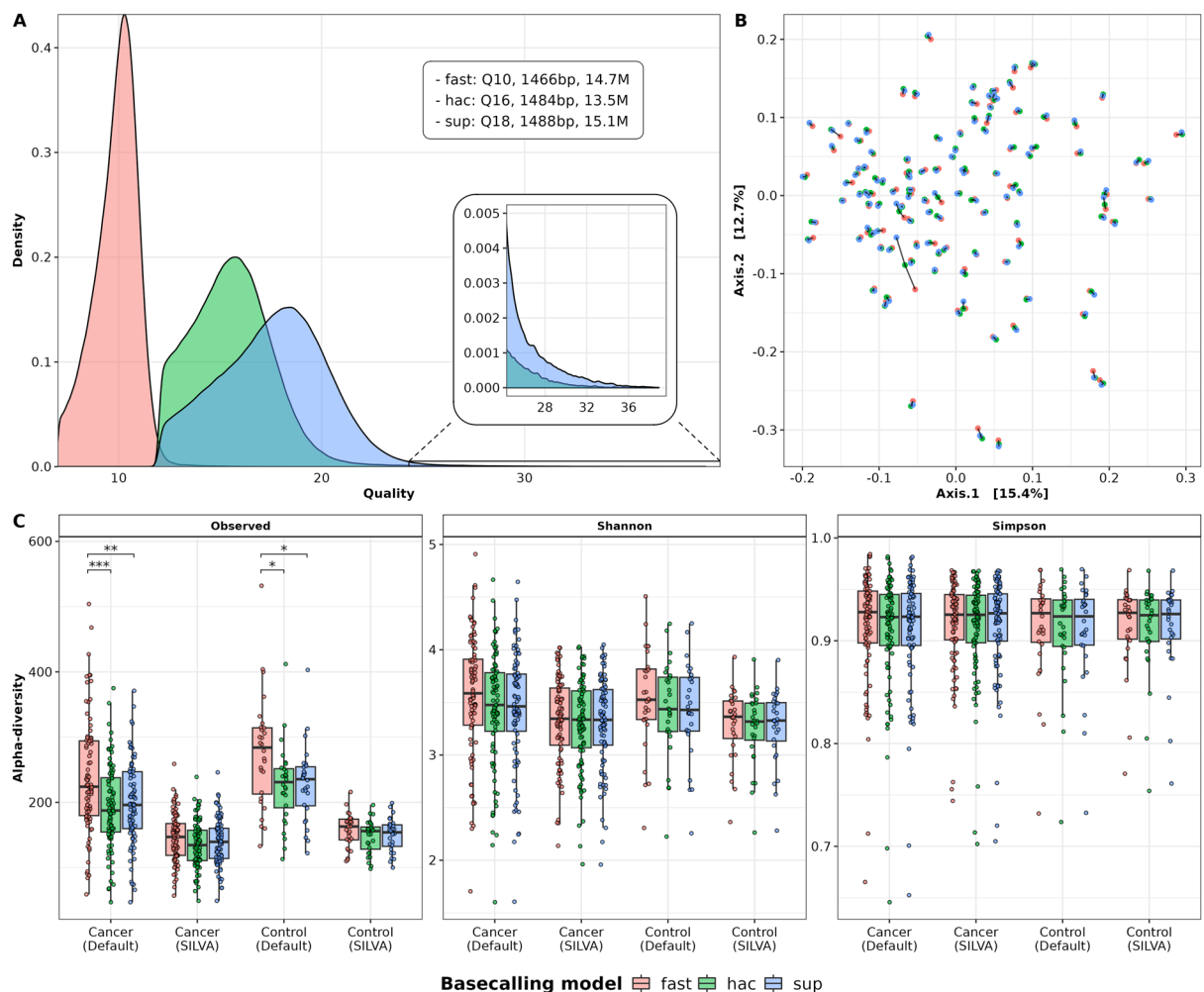


Fig. 1. Comparison of ONT-V1V9 basecalling models. **(A)** Distribution of the average quality of reads per basecalling model. **(B)** β -diversity analysis using SILVA database and MDS+JSD (Multidimensional Scaling and Jensen-Shannon Divergence). The same samples with different models are connected by lines. **(C)** α -diversity analysis using SILVA and Emu's Default database. Significance levels are given through Wilcoxon rank sum tests and comparisons between groups are not shown.

indicating more, such as: *Agathobaculum butyriciproducens*, *Romboutsia ilealis*, *Anaerostipes rhamnosivorans* and *Anaerocolumna cellulositytica*.

In order to identify useful combinations of CRC microbial biomarkers, multiple machine learning models were created, using both automatic and manual feature selection. Models using automatically selected features are shown in Table 1, with the Default Emu database and two different prevalence levels (10% or 30%). All of them obtained at least 0.9 AUC (Area Under Curve) with just 10 features. For example, using the Default database at 10% prevalence the most important features according to the Boruta algorithm were *P. micra*, *A. cellulositytica*, *A. rhamnosivorans*, *P. stomatis*, *A. butyriciproducens*, *P. anaerobius*, *Prevotella stercorea*, *Candidatus Saccharibacteria bacterium oral taxon 957* (also named *Candidatus Nanosynbacter featherlites* (TM7) in NCBI), *Olsenella timonensis* and *Raoultibacter timonensis* (Table 1).

However, further evaluation of these taxa at the read level showed low identity to their reference. Specifically, *A. cellulositytica* and *A. rhamnosivorans*, which were big contributors to these models, had a maximum of 90–92% identity, which should be classified as unknown *Lachnospiraceae*. For this reason, manual combinations of taxa with high read identity and coverage, significant differences between groups (using ANCOM-BC and CLR abundance) or identified as relevant by the Boruta algorithm were tested. These combinations and their AUC are shown in Figure 4D (Full list is available in Supplementary Table 1 and additional information or comments of each species is in Supplementary Table 2) and compared to previously described combinations with Illumina-V3V4¹². The usage of *P. micra* and *F. nucleatum* provides an AUC of 0.71, increasing to 0.76 by adding *B. fragilis*, to 0.82 by adding *A. butyriciproducens* and, lastly, obtaining a maximum AUC of 0.87 with a total of 14 features. All comparable combinations to Illumina-V3V4 obtained slightly higher AUC (e.g. *P. micra*+*F. nucleatum* vs. *Parvimonas*+*Fusobacterium*).

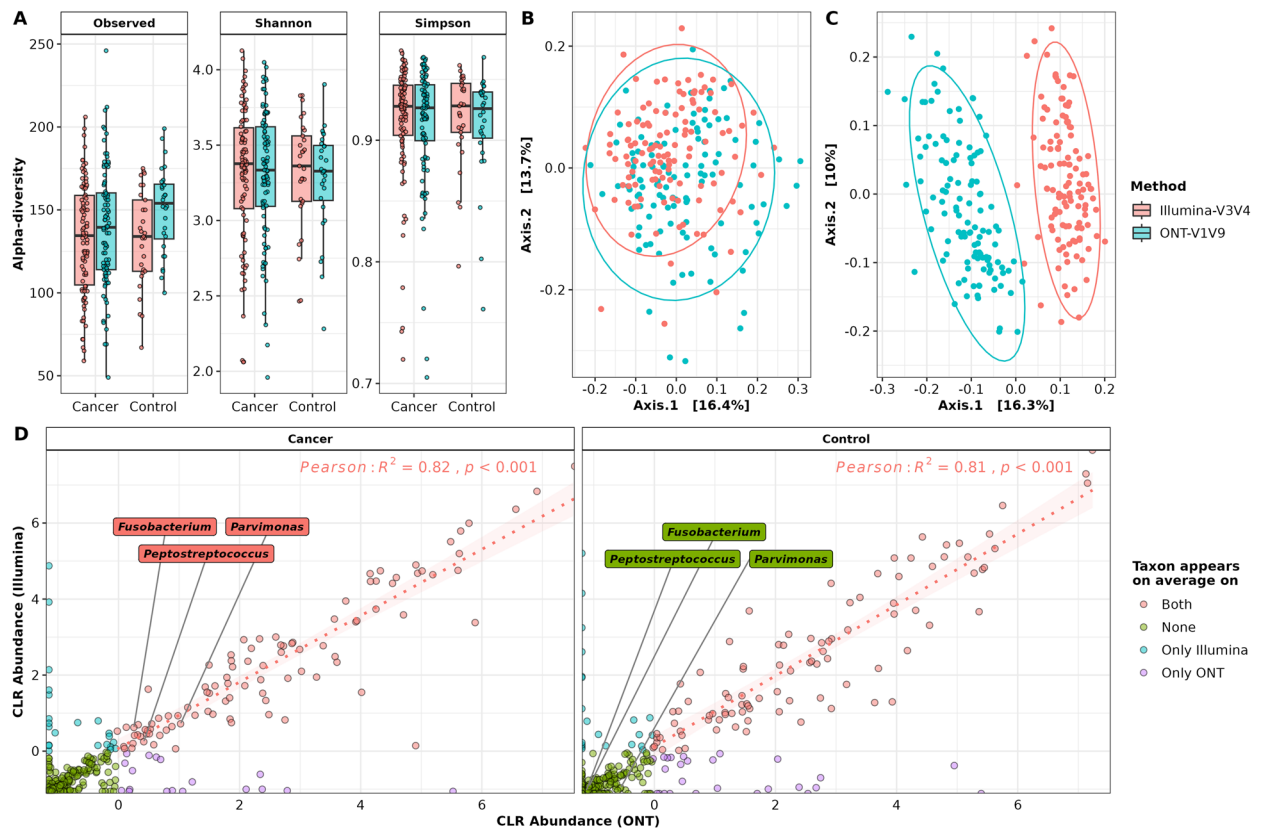


Fig. 2. Comparison of Illumina-V3V4 and ONT-V1V9 approaches using SILVA. **(A)** α -diversity indexes at the genus level, comparing the two volunteer groups (cancer vs. control). **(B)** β -diversity analysis at the genus level. **(C)** β -diversity analysis at the species level. **(D)** Mean Centered Log Ratio (CLR) abundance correlation between approaches and for each group (cancer vs. control). Each point represents a different genus, and color indicates if that taxa appear, on average, on both, none or only one of the approaches. Three relevant genera, which contain multiple species related to colorectal cancer are highlighted (*Fusobacterium*, *Parvimonas* and *Peptostreptococcus*).

Discussion

Sequencing of the ubiquitous 16S rRNA gene is a useful approach for microorganism characterization in complex and fastidious samples such as the ones used in this study, where their composition consists of hundreds of species that cannot be easily cultured under laboratory conditions. This technique allows for the discovery of disease-related bacterial biomarkers, which could be a useful approach for early prevention or diagnosis of various afflictions, such as colorectal cancer. In this procedure, small regions, ~400–500 nt out of ~1500 nt (e.g. V3V4), are typically sequenced with paired-end short-read technologies (e.g. Illumina). Unfortunately, this is mainly effective for genus level identification, while long-read technologies, such as ONT or PacBio, provide the possibility of sequencing the full gene (~1500 nt, V1V9)^{42,64}, allowing for a more accurate identification at the species level. In this study, two approaches (Illumina-V3V4 and ONT-V1V9), three ONT Dorado basecalling models (fast, hac, sup; v4.1.0), and two databases (SILVA and Emu's Default database) were compared in feces samples obtained from two groups of volunteers (colon cancer patients and healthy subjects) in order to assess ONT-V1V9's capabilities and to improve upon a previously described set of microbial biomarkers described by our research team for the early diagnosis of CRC¹².

Currently, out of these technologies, Illumina and PacBio HiFi reads provide the highest accuracy consistently^{42,65}. However, ONT's sequencing chemistry and basecalling models are constantly improving and will, presumably, eventually get to a similar accuracy routinely, as a small proportion of them can already get up to Q30. For now, ONT reads are restricted to be analyzed with tools that take into account their relatively high error rate, such as Emu⁴³, while popular algorithms for ASV creation like DADA2⁴¹ cannot overcome their properties. In this study, three of the latest models from Dorado (for 4KHz data, v4.1.0) were compared, obtaining a median quality of Q18 with the sup model. More recent models like sup v5.0.0+, which are incompatible with the current data, appear to consistently achieve Q22+. Therefore, it is probable that future basecalling models will allow for ONT reads to be analyzed with tools typically reserved for Illumina or PacBio, but it remains to be seen.

Focusing on our results, the three different ONT-V1V9 basecalling models tested in this work did not influence β -diversity or α -diversity, except in the case of observed features, where the fast model had significantly higher features than the rest. As explained before, this could be due to false positives derived from low accuracy sequences. Additionally, the percentage of reads identified discordantly in regard to sup was significantly

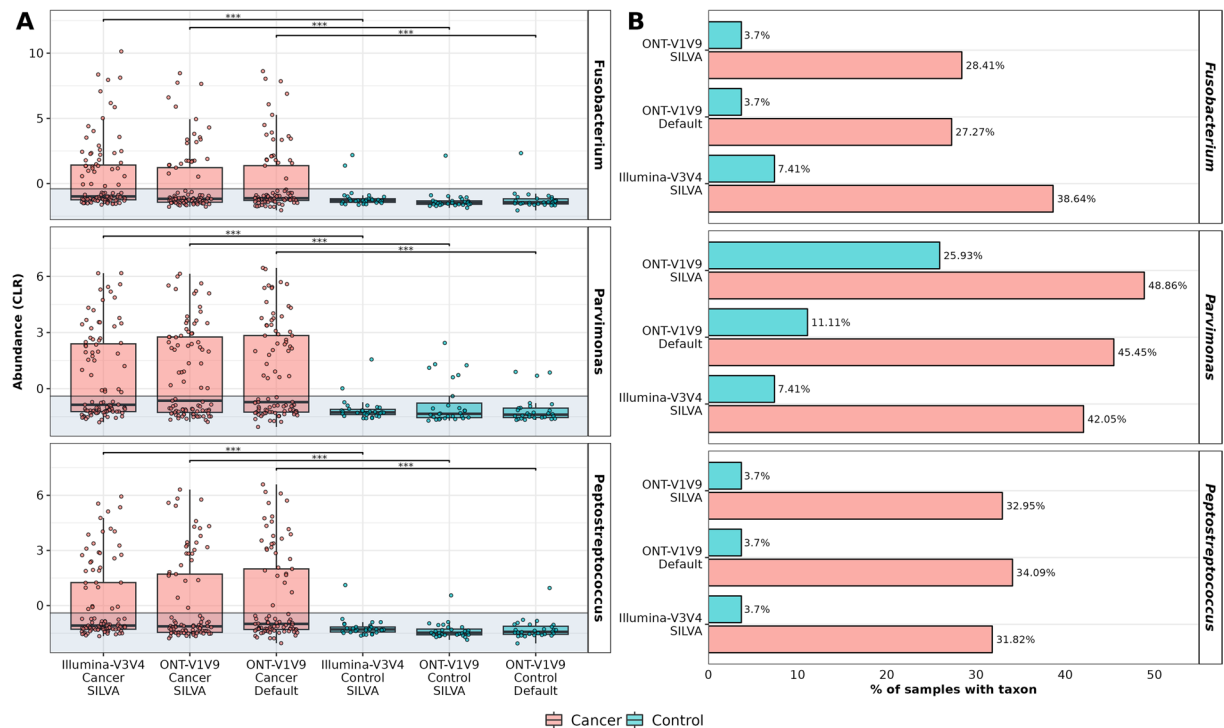


Fig. 3. Comparison of Illumina-V3V4 and ONT-V1V9 approaches with both databases for specific genera in each subject group. Three important genera, which contain multiple species associated with colorectal cancer are highlighted (*Fusobacterium*, *Parvimonas* and *Peptostreptococcus*). (A) Centered Log Ratio (CLR) abundance of each genus per group using Wilcoxon rank sum tests to assess significance. (B) Percentage of samples where each genus is present per group.

different when comparing the fast and hac models. According to our observations, the taxonomic and diversity results can be affected by the selected basecalling model and its read quality, therefore the best model (sup v4.1.0 in this case) should be chosen, even if it is very computationally expensive. Additionally, the use of duplex reads in amplicons with high homology and similar length might not be completely appropriate, as we observed (and removed) false duplex sequences, deriving from two unrelated simplex reads with different barcodes.

Comparing Illumina-V3V4 and ONT-V1V9 (sup) approaches with SILVA resulted in no significant differences when looking at α -diversity. In regard to β -diversity, although samples did overlap at the genus level, PERMANOVA analysis indicated that there were significant differences at both genus and species level⁶⁶. The differences at the genus level could be attributed to slight differences in taxonomy caused by the use of two different classifiers, while at the species level they are expected and likely due to a more precise identification with ONT. The CLR abundance and presence of three genera, which contain species related to colorectal cancer was similar and the percentage of feature counts identified as a known species was significantly higher in ONT-V1V9 (27.74%) than in Illumina V3V4 (16.75%). Thus, these results could indicate that both approaches are similar at the genus level, although not identical, and that ONT-V1V9 increases the amount of species identified, as expected and previously described^{64,66}.

Interestingly, when using Emu's Default database, the percentage of feature counts identified as a known species increased to 100%. Closer examination of both databases revealed that SILVA was seven times larger than Emu's Default database (NCBI 16S RefSeq + rrnDB), and contained sequences identified at multiple taxonomic levels, including unclassified or uncultured taxa, while Emu's Default database had all its sequences identified at the species level. A smaller database, such as Emu's Default database, might function better with higher error reads that really belong to a well-characterized species, but the lack of unknown microorganisms in Emu's Default database leads to some reads being assigned to a species even when their identity is 90%. Ideally, this kind of reads should be assigned to higher taxonomic levels such as family, reserving the species level for reads with $\geq 98\%$ identity⁴⁰, but it is a difficult task, as described by Emu authors⁴³. This is specially evident in specific families with many unknown or uncultured species, such as the gut bacteria inside the *Lachnospiraceae* family. This is complicated further by the use of relatively high error reads where Q18, the median accuracy obtained here, equals to an error rate of $\sim 2\%$ (Q15 would equal to $\sim 3\%$), although this is expected to improve in the near future, as commented previously. The consequence of this may be significant, especially in the search for microbial biomarkers applied to the diagnosis of human disorders, so it is important to be cautious when choosing the reference database to use with Emu and to study its results at the read level.

Selection of colorectal cancer biomarkers through ANCOM-BC and CLR abundance analyses revealed multiple species with high read identity and increased abundance in CRC patients such as *Parvimonas micra*, *Fusobacterium nucleatum*, *Peptostreptococcus stomatis*, *Peptostreptococcus anaerobius*, *Gemella morbillorum*,

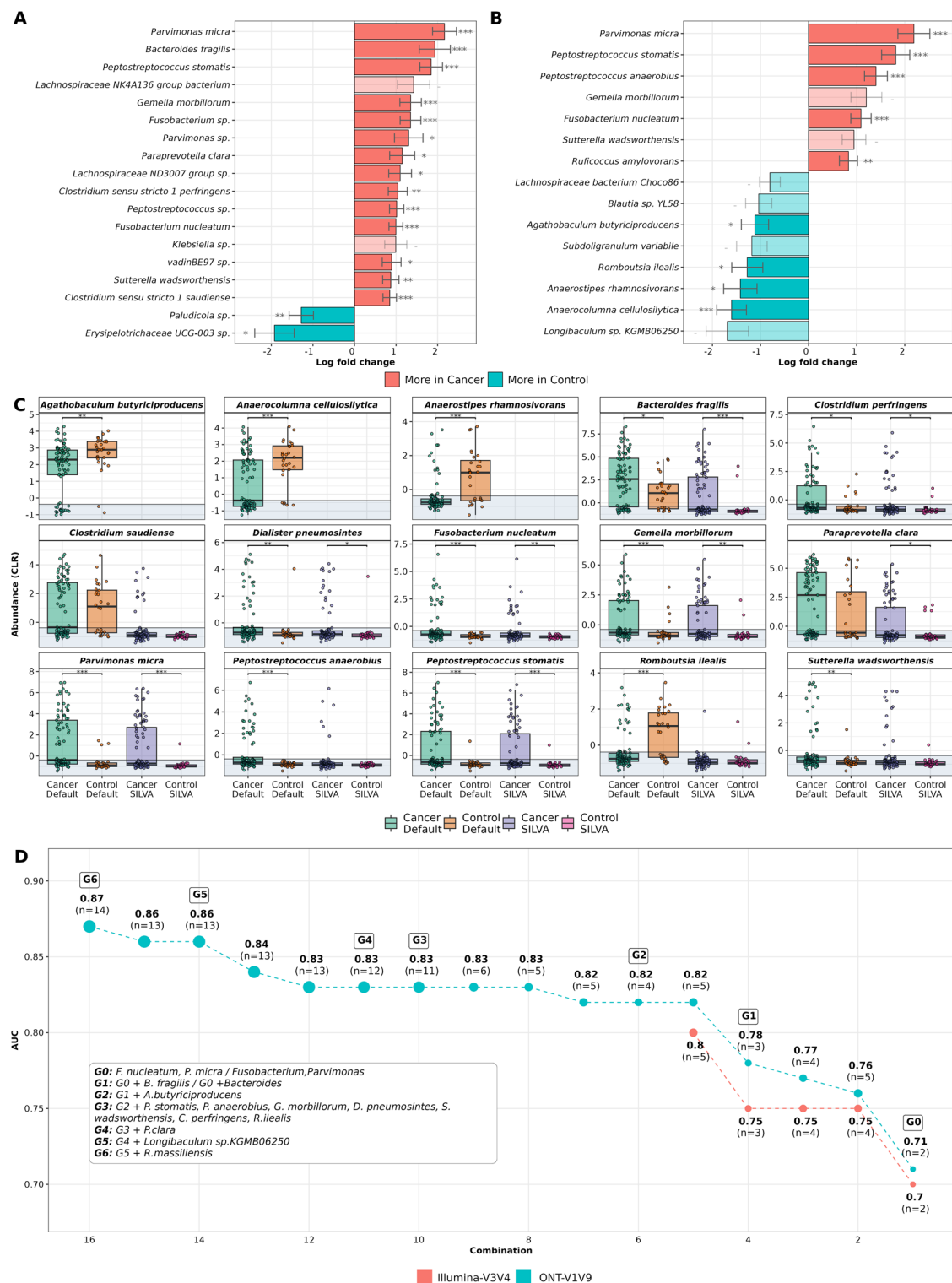


Fig. 4. Colorectal cancer biomarkers obtained with ONT-V1V9 and both databases. **(A)** Differential abundance analysis (DAA) through ANCOM-BC using ONT-V1V9 and SILVA. Control subjects are the reference group, meaning a higher Log Fold Change (LFC) indicates higher abundance of a taxon in the cancer group. **(B)** DAA through ANCOM-BC using ONT-V1V9 and Emu's Default database. Control subjects are the reference group, meaning a higher Log Fold Change (LFC) indicates higher abundance of a taxon in the cancer group. **(C)** Centered Log Ratio (CLR) abundance of species with significant differences between cancer and control groups, indicated through Wilcoxon rank sum tests. **(D)** AUC of two models (Illumina-V3V4 or ONT-V1V9) using manually selected features based on read identity to reference, ANCOM-BC and CLR abundance. Sample size (n) refers to the number of features included in each model. A complete list of species and genera included in each combination is found in Supplementary Table 1.

Selection	Features	Feature count	AUC
Boruta top 10 (30%)	<i>P. micra</i> , <i>A. butyrificiproducens</i> , <i>A. cellulositytica</i> , <i>O. timonensis</i> , <i>A. bacterium</i> , <i>R. timonensis</i> , <i>Streptococcus</i> sp. A12, <i>B. luti</i> , <i>Clostridium</i> sp. BNL1100, <i>S. variabile</i>	10	0.92
Boruta top 10 (10%)	<i>P. micra</i> , <i>A. cellulositytica</i> , <i>A. rhamnosivorans</i> , <i>P. stomatis</i> , <i>A. butyrificiproducens</i> , <i>P. anaerobius</i> , <i>P. stercorea</i> , <i>Candidatus Saccharibacteria bacterium</i> oral taxon 957, <i>O. timonensis</i> , <i>R. timonensis</i>	10	0.91
Manual top 4	<i>F. nucleatum</i> , <i>P. micra</i> , <i>B. fragilis</i> , <i>A. butyrificiproducens</i>	4	0.82
Manual top 14	<i>F. nucleatum</i> , <i>P. micra</i> , <i>B. fragilis</i> , <i>A. butyrificiproducens</i> , <i>P. stomatis</i> , <i>P. anaerobius</i> , <i>G. morbillorum</i> , <i>D. pneumosintes</i> , <i>S. wadsworthensis</i> , <i>C. perfringens</i> , <i>R. ilealis</i> , <i>P. clara</i> , <i>Longibaculum</i> sp. KGMB06250, <i>R. massiliensis</i>	14	0.87

Table 1. Area under the curve (AUC) for the prediction of cancer/control with different machine learning models using ONT-V1V9 and Emu’s Default database, based on if feature selection is automatic (Boruta, with two prevalence thresholds of 10% and 30%) or manual.

Sutterella wadsworthensis, *Clostridium perfringens*, *Bacteroides fragilis* and *Dialister pneumosintes* when compared to control samples from healthy volunteers. These species have been previously associated to CRC through a variety of methods^{12,18–20,23–25}. Other biomarkers for healthy controls were indicated by Emu’s Default database, such as *Agathobaculum butyrificiproducens*, a butyrate-producing species⁶⁷, a molecule associated to a healthy gut through its anti-inflammatory properties⁶⁸. Thus, the species level identification of these bacteria improves upon Illumina-V3V4, where the only clearly defined colorectal cancer biomarker at the species level was *B. fragilis*¹², providing higher definition.

Machine learning models using automatic feature selection achieved good results for the prediction of colorectal cancer, with an AUC of 0.9 with just 10 features. However, closer evaluation of some of the selected taxa when using Emu’s Default database revealed low identity to their reference, due to previously commented issues. This led to manual feature selection of taxa with high read identity and coverage, significant differences between groups or identified as important by Boruta. The maximum AUC obtained was 0.87 with a total of 14 features, with the highest contributors being *P. micra*, *F. nucleatum*, *B. fragilis* and *A. butyrificiproducens* (AUC 0.82). The comparable combinations to Illumina-V3V4 were slightly better in ONT-V1V9, but not overwhelmingly better. This could be due to the fact that, even if Illumina-V3V4 was only able to confidently observe genera, the underlying species were few (or just one in the case of *Parvimonas*), which ends up providing similar abundances between both techniques.

Overall, these results bode well for ONT-V1V9 and for the diagnosis of CRC using the described biomarkers. However, a limitation of the present study is the sample size (n = 123) and specially the uneven amount of subjects in each group (~25% healthy subjects), which could be greatly influencing feature selection for colorectal cancer prediction. Thus, said biomarkers must be thoroughly assessed in a more geographically diverse, larger and equilibrated cohort. Another limitations is not knowing the actual ground truth of these samples. Nonetheless, as previous studies have reported before, the usage of 16S rRNA V1V9 sequencing with ONT presents clear advantages when comparing to short-read technologies information-wise (sequencing a longer segment provides better identification), and against other long-read technologies cost-wise (if including the price of the sequencer)^{42,66}. The technique itself, 16S rRNA metabarcoding, is indeed cost-efficient for bacterial characterization in complex samples, although it is expected that, as sequencing and computation costs keep diminishing, the better alternative will be to use metagenomic sequencing. This approach will provide a more accurate classification and, ideally, a comprehensive perspective of each microbial genome, which in the case of colorectal cancer related species could show specific virulence factors that influence the disease.

Data availability

Raw sequence reads generated from 16S rRNA gene sequencing by MiSeq (Illumina-V3V4) and MinION (ONT-V1V9, Dorado sup v4.1.0) are available in the NCBI SRA database under bioproject PRJNA911189. Used scripts and additional documentation are available at <https://github.com/Pablo-Aja-Macaya/CRC-ONT16S>.

Received: 20 November 2024; Accepted: 7 July 2025

Published online: 21 July 2025

References

1. Austin, B. The value of cultures to modern microbiology. *Antonie Van Leeuwenhoek* **110**(10), 1247–1256 (2017).
2. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci.* **113**(21), 5970–5975 (2016).
3. Větrovský, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PloS one* **8**(2), 57923 (2013).
4. Mandlik, J. S., Patil, A. S. & Singh, S. Next-generation sequencing (NGS): Platforms and applications. *J. Pharm. Bioallied Sci.* **16**(Suppl 1), 41–45 (2024).
5. Gao, B. et al. An introduction to next generation sequencing bioinformatic analysis in gut microbiome studies. *Biomolecules* **11**(4), 530 (2021).
6. Leggett, R. M. et al. Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nat. Microbiol.* **5**(3), 430–442 (2020).
7. Baxter, N. T., Ruffin, M. T., Rogers, M. A. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 1–10 (2016).
8. Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S. & Frizelle, F. A. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci. Rep.* **7**(1), 11590 (2017).

9. Bosch, S. et al. The potential of fecal microbiota and amino acids to detect and monitor patients with adenoma. *Gut Microbes* **14**(1), 2038863 (2022).
10. Clos-García, M. et al. Integrative analysis of fecal metagenomics and metabolomics in colorectal cancer. *Cancers* **12**(5), 1142 (2020).
11. Eklöf, V. et al. Cancer-associated fecal microbial markers in colorectal cancer detection. *Int. J. Cancer* **141**(12), 2528–2536 (2017).
12. Conde-Pérez, K. et al. *The multispecies microbial cluster of Fusobacterium, Parvimonas* (Bacteroides and Faecalibacterium as a precision biomarker for colorectal cancer diagnosis, Molecular Oncology, 2024).
13. International Agent for Research on Cancer, World Health Organization: Global Cancer Observatory (2024). <https://gco.iarc.fr/en> Accessed 01 Apr 2024
14. Spanish Association Against Cancer: Dimensions of Cancer (2024). <https://observatorio.contraelcancer.es/explora/dimensiones-del-cancer> Accessed 01 Apr 2024
15. European Commission: European guidelines on colorectal cancer screening and diagnosis (2024). <https://cancer-screening-and-care.jrc.ec.europa.eu/en/ecicc/european-colorectal-cancer-guidelines> Accessed 01 Apr 2024
16. Vacante, M., Ciuni, R., Basile, F. & Biondi, A. Gut microbiota and colorectal cancer development: A closer look to the adenoma-carcinoma sequence. *Biomedicine* **8**(11), 489 (2020).
17. Servizo Galego de Saúde (SERGAS): Programa galego de detección precoz do cancro colorrectal. Resultados 2013-2021 (2021). https://www.sergas.es/Saude-publica/Documents/536/Informe_anual_2013_2021_PGDPCC.pdf Accessed 15 Apr 2024
18. Kim, J. & Lee, H. K. Potential role of the gut microbiome in colorectal cancer progression. *Front. Immunol.* **12**, 807648 (2022).
19. Zackular, J. P. et al. The gut microbiome modulates colon tumorigenesis. *MBio* **4**(6), 10–1128 (2013).
20. Boleij, A. et al. The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* **60**(2), 208–215 (2015).
21. Cuevas-Ramos, G. et al. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc. Natl. Acad. Sci.* **107**(25), 11537–11542 (2010).
22. Arthur, J.C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J.M., Fan, T.-J., Campbell, B.J., Abujamel, T., Dogan, B., Rogers, A.B. et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**(6103), 120–123 (2012).
23. Cheng, W. T., Kantilal, H. K. & Davamani, F. The mechanism of Bacteroides fragilis toxin contributes to colon cancer formation. *Malays. J. Med. Sci. MJMS* **27**(4), 9 (2020).
24. Bergsten, E. et al. Parvimonas micra, an oral pathobiont associated with colorectal cancer, epigenetically reprograms human colonocytes. *Gut Microbes* **15**(2), 2265138 (2023).
25. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res.* **30**(6), 492–506 (2020).
26. Ewaschuk, J.B., Diaz, H., Meddings, L., Diederichs, B., Dmytrash, A., Backer, J., Looijer-van Langen, M., & Madsen, K.L. Secreted bioactive factors from Bifidobacterium infantis enhance epithelial cell barrier function. *Am. J. Physiol.-Gastrointest. Liver Physiol.* **295**(5), 1025–1034 (2008).
27. Anderson, R. C. et al. Lactobacillus plantarum MB452 enhances the function of the intestinal barrier by increasing the expression levels of genes involved in tight junction formation. *BMC Microbiol.* **10**, 1–11 (2010).
28. Potrykus, M., Czaja-Stolc, S., Stankiewicz, M., Kaska, Ł. & Małgorzewicz, S. Intestinal microbiota as a contributor to chronic inflammation and its potential modifications. *Nutrients* **13**(11), 3839 (2021).
29. Cougnoux, A. et al. Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* **63**(12), 1932–1942 (2014).
30. Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., Hoeck, A., Wood, H.M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., & Stege, P.B. et al. Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. *Nature* **580**(7802), 269–273 (2020).
31. Li, S. et al. Tumorigenic bacteria in colorectal cancer: Mechanisms and treatments. *Cancer Biol. Med.* **19**(2), 147 (2022).
32. Jasemi, S. et al. Toxigenic and non-toxigenic patterns I, II and III and biofilm-forming ability in Bacteroides fragilis strains isolated from patients diagnosed with colorectal cancer. *Gut Pathogens* **12**, 1–7 (2020).
33. Chen, Y., Liu, B., Wei, Y. & Kuang, D.-M. Influence of gut and intratumoral microbiota on the immune microenvironment and anti-cancer therapy. *Pharmacol. Res.* **174**, 105966 (2021).
34. Scott, N., Whittle, E., Jeraldo, P. & Chia, N. A systemic review of the role of enterotoxigenic Bacteroides fragilis in colorectal cancer. *Neoplasia* **29**, 100797 (2022).
35. Wei, P. et al. Microbiota in colorectal cancer related to liver metastasis. *Chin. J. Cancer Res.* **36**(1), 17 (2024).
36. Yadav, D. et al. Gut microbiome-colorectal cancer relationship. *Microorganisms* **12**(3), 484 (2024).
37. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **82**(11), 801–811 (2021).
38. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**(11), 1348–1365 (2021).
39. Oxford Nanopore Technologies: Microbiome sequencing & analysis (2024). <https://nanoporetech.com/applications/research-area-s/microbiome/technology-comparison> Accessed 22 Apr 2024
40. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**(14), 2371–2375 (2018).
41. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**(7), 581–583 (2016).
42. Buetas, E. et al. Full-length 16S rRNA gene sequencing by PacBio improves taxonomic resolution in human microbiome samples. *BMC Genomics* **25**(1), 310 (2024).
43. Curry, K. D. et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nat. Methods* **19**(7), 845–853 (2022).
44. Rodríguez-Pérez, H., Ciuffreda, L. & Flores, C. Nanoclust: A species-level analysis of 16s rRNA nanopore sequencing data. *Bioinformatics* **37**(11), 1600–1601 (2021).
45. Conde-Pérez, K., Buetas, E., Aja-Macaya, P., Martín-De Arribas, E., Iglesias-Corrás, I., Trigo-Tasende, N., Nasser-Ali, M., Estévez, L.S., Rumbo-Feal, S., Otero-Alén, B. et al. Parvimonas micra can translocate from the subgingival sulcus of the human oral cavity to colorectal adenocarcinoma. *Mol. Oncol.* (2022).
46. Oxford Nanopore PLC.: Dorado. GitHub (2023)
47. De Coster, W. & Rademakers, R. NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**(5), 311 (2023).
48. Oxford Nanopore PLC.: Duplex Tools. GitHub (2023)
49. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with kraken 2. *Genome Biol.* **20**, 1–13 (2019).
50. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. & Schmidt, T. M. rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* **43**(D1), 593–598 (2015).
51. O’Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**(D1), 733–745 (2016).
52. Schoch, C. L. et al. NCBI taxonomy: A comprehensive update on curation, resources and tools. *Database* **2020**, 062 (2020).
53. Pruesse, E. et al. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**(21), 7188–7196 (2007).

54. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**(D1), 590–596 (2012).
55. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**(8), 852–857 (2019).
56. Robeson, M. S. et al. RESCRIPt: Reproducible sequence taxonomy reference database management. *PLoS Comput. Biol.* **17**(11), 1009581 (2021).
57. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). R Foundation for Statistical Computing. <https://www.R-project.org/>
58. McMurdie, P. J. & Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS one* **8**(4), 61217 (2013).
59. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**(1), 3514 (2020).
60. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 65–70 (1979)
61. Lahti, L. & Shetty, S. Microbiome R Package (2012–2019)
62. Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H.B.A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.-H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C.J.F., & Weedon, J. vegan: Community Ecology Package. (2022). R package version 2.6–4. <https://CRAN.R-project.org/package=vegan>.
63. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
64. Matsuo, Y. et al. Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. *BMC Microbiol.* **21**, 1–13 (2021).
65. Pacific Biosciences: HiFi Sequencing: HiFi reads for highly accurate long-read sequencing (2024). <https://www.pacb.com/technology/hifi-sequencing/> Accessed 26 Aug 2024
66. Yeo, K. et al. A comparison between full-length 16S rRNA Oxford nanopore sequencing and Illumina V3–V4 16S rRNA sequencing in head and neck cancer tissues. *Arch. Microbiol.* **206**(6), 248 (2024).
67. Ahn, S., Jin, T.-E., Chang, D.-H., Rhee, M.-S., Kim, H.J., Lee, S.J., Park, D.-S., & Kim, B.-C. *Agathobaculum butyriciproducens* gen. nov. sp. nov., a strict anaerobic, butyrate-producing gut bacterium isolated from human faeces and reclassification of *Eubacterium desmolans* as *Agathobaculum desmolans* comb. nov. *Int. J. Syst. Evolut. Microbiol.* **66**(9), 3656–3661 (2016)
68. Liu, H. et al. Butyrate: A double-edged sword for health?. *Adv. Nutr.* **9**(1), 21–29 (2018).

Acknowledgements

We would like to thank all the participants for collaborating with us during the development of this project and to the surgical nurse Gema Carro Díaz for their assistance during patients' recruitment and sampling phase.

Author contributions

K.C.-P., J.A.V., M.P., S.L., A.M. and G.B. conceived and designed the study. M.P. and J.A.V. supervised the study. K.C.-P., J.A.V., S.R.-F., M.N.-A., and M.P. obtained all samples. Only M.P. managed the informed consents of volunteers. K.C.-P., N.T.-T. and M.N.-A. processed samples. K.C.-P., N.T.-T. and P.N. performed 16S rRNA sequencing in both Illumina and Oxford Nanopore platforms. P.A.-M. and E.B. made bioinformatic and statistical analysis. K.C.-P. and P.A.-M. designed and created tables and figures. P.A.-M., K.C.-P. and N.T.-T. wrote the main manuscript text. All authors reviewed the manuscript. All authors agree to the publication of the results obtained in this biomedical research project.

Funding

This study has been funded by the Instituto de Salud Carlos III (ISCIII), through the projects PI20/00413 and PI23/00696, cofunded by the European Union (EU) to MP. Also, the work was funded by the project IN607A2024/09 from Xunta de Galicia (Spain) to GB. Additionally, it has been supported by the project RTI2018-102032-B-I00 from the Spanish Ministry of Science and Innovation to AM and by CIBER INFEC ISCIII (CB21/13/00055) to GB and MP. KC-P was financially supported by the Spanish Association against Cancer (AECC) from December 2019 to December 2023. Additionally, this work, integrated into the framework of PERTE for Vanguard Health, has been co-financed by the Spanish Ministry of Science, Innovation and Universities with funds from the European Union NextGenerationEU, from the Recovery, Transformation and Resilience Plan (PRTR-C17.I1) and from the Autonomous Community of Galicia within the framework of the Biotechnology Plan Applied to Health to SL.

Declarations

Competing interests

All the authors declare no competing interests.

Ethics approval and consent to participate

This study adhered to the standards of clinical practice and research regulations (Law of Biomedical Research 14/2007), in agreement with the Declaration of Helsinki and the Convention on Human Rights and Biomedicine. Compliance with the protection of non-public personal data of all those involved within the RGPD – UE 2016/679, LOPDGD 3/2018 Law 41/2002 and its implementing regulations, Royal Decree 1720/2007, were enforced. This project (PI20/00413), granted by Carlos III Health Institute (ISCIII; Spain), was supervised by the local ethics committee, the Research Ethical Committee of Galicia (code CEIm-G 2018/609, Galicia, Spain), and by the Spanish Agency for Medicines and Healthcare Products (AEMPS) for the use of CRC patients' samples from CHUAC (A Coruña, Galicia, Spain). The experimental protocols were approved by the Ethical Committee of Galicia (CEIm-G 2018/609, Galicia, Spain). Informed consent for Biobank (CHUAC, A Coruña, Galicia, Spain, UNE-EN ISO 9001-2015) was signed by all individuals grouped in this study. Anonymized clinical data used during the study for CRC patients was obtained from the Galician Health

Service (SERGAS). All individuals recruited in this project (cancer patients and controls) signed a formal consent form for the publication of scientific and clinical results in scientific articles.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-10999-8>.

Correspondence and requests for materials should be addressed to J.A.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025