frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 17 March 2025
DOI 10.3389/fbinf.2025.1484113

OPEN ACCESS

EDITED BY
David W. Ussery,
University of Arkansas for Medical Sciences, United States

REVIEWED BY
Maryam Omrani,
San Raffaele Hospital, Italy
Chen Li,
St Jude Children Hospital, United States

*CORRESPONDENCE
Amy S. Graham,
grhamy001@myuct.ac.za

†These authors have contributed equally to this work and share last authorship

RECEIVED 21 August 2024
ACCEPTED 11 February 2025
PUBLISHED 17 March 2025

CITATION
Graham AS, Patel F, Little F, van der Kouwe A, Kaba M and Holmes MJ (2025) Using short-read 16S rRNA sequencing of multiple variable regions to generate high-quality results to a species level.
Front. Bioinform. 5:1484113.
doi: 10.3389/fbinf.2025.1484113

Using short-read 16S rRNA sequencing of multiple variable regions to generate high-quality results to a species level

Amy S. Graham[1,2]*, Fadheela Patel[3], Francesca Little[4], Andre van der Kouwe[5,6], Mamadou Kaba[3†] and Martha J. Holmes[1,2,7,8†]

[1]Imaging Sciences, Neuroscience Institute, University of Cape Town, Cape Town, South Africa, [2]Department of Human Biology, Division of Biomedical Engineering, University of Cape Town, Cape Town, South Africa, [3]Department of Pathology, Division of Medical Microbiology, University of Cape Town, Cape Town, South Africa, [4]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa, [5]Athinoula A. Martinos Centre for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA, United States, [6]Department of Radiology, Harvard Medical School, Boston, MA, United States, [7]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC, Canada, [8]ImageTech, Simon Fraser University, Surrey, BC, Canada

Introduction: Short-read amplicon sequencing studies have typically focused on 1-2 variable regions of the 16S rRNA gene. Species-level resolution is limited in these studies, as each variable region enables the characterisation of a different subsection of the microbiome. Although long-read sequencing techniques can take advantage of all 9 variable regions by sequencing the entire 16S rRNA gene, short-read sequencing has remained a commonly used approach in 16S rRNA research. This work assessed the feasibility of accurate species-level resolution and reproducibility using a relatively new sequencing kit and bioinformatics pipeline developed for short-read sequencing of multiple variable regions of the 16S rRNA gene. In addition, we evaluated the potential impact of different sample collection methods on our outcomes.

Methods: Using xGen™ 16S Amplicon Panel v2 kits, sequencing of all 9 variable regions of the 16S rRNA gene was carried out on an Illumina MiSeq platform. Mock cells and mock DNA for 8 bacterial species were included as extraction and sequencing controls respectively. Within-run and between-run replicate samples, and pairs of stool and rectal swabs collected at 0–5 weeks from the same infants, were incorporated. Observed relative abundances of each species were compared to theoretical abundances provided by ZymoBIOMICS. Paired Wilcoxon rank sum tests and distance-based intraclass correlation coefficients were used to statistically compare alpha and beta diversity measures, respectively, for pairs of replicates and stool/rectal swab sample pairs.

Results: Using multiple variable regions of the 16S ribosomal Ribonucleic Acid (rRNA) gene, we found that we could accurately identify taxa to a species level and obtain highly reproducible results at a species level. Yet, the microbial profiles of stool and rectal swab sample pairs differed substantially despite being collected concurrently from the same infants.

Conclusion: This protocol provides an effective means for studying infant gut microbial samples at a species level. However, sample collection approaches need to be accounted for in any downstream analysis.

KEYWORDS

microbiome, 16S rRNA sequencing, short-read, multiple variable regions, species-level

1 Introduction

Our ability to describe and appreciate the complexities of the human microbiome has been radically improved by next-generation sequencing tools (Bharti and Grimm, 2021; Ji and Nielsen, 2015; Rogers and Bruce, 2010). Although the use of second-generation, short-read sequencing platforms allows high read depths to be rapidly sequenced (Hu et al., 2021; Tucker et al., 2009), it is limited in terms of the assembly of contiguous sequences (Li et al., 2010; Zerbino and Birney, 2008). Various third-generation sequencing techniques exist and allow long reads to be sequenced, yet a greater number of errors have previously been found to result with these techniques (Amarasinghe et al., 2020; Midha et al., 2019; Sedlazeck et al., 2018; Van Dijk et al., 2018; Quail et al., 2012). As short-read sequencing approaches are still frequently used, however, there is a need to consider alternative ways to improve these approaches.

The 16S rRNA gene has been identified as a particularly useful target of research as it is common to all bacteria (Acinas et al., 2004; Patel, 2001). The gene consists of regions of DNA in which the sequence is conserved across all bacteria, while in other regions there is variation according to the individual bacterial species (Wang and Qian, 2009; Lane et al., 1985). As such, targeted amplicon sequencing can be done, comparing variable region sequences to a database of known taxa, to identify which bacterial species are present in a sample (Wang and Qian, 2009). In the past, amplicon sequencing studies have typically focused on one or two variable regions at a time (Claassen-Weitz et al., 2018; Gao et al., 2018; Yu et al., 2017; Hosgood III et al., 2014; Caporaso et al., 2011; Zhou et al., 2011). Yet, certain variable regions are better for enabling classification to lower taxonomic levels and each variable region favours classification of specific taxa (Bukin et al., 2019; Guo et al., 2013; Chakravorty et al., 2007). Consequently, this approach limits the ability to obtain accurate species-level resolution when focusing only on a single short fragment of the 16S rRNA gene. Using the entire 16S rRNA sequence is expected to provide better classification potential to a species level (Johnson et al., 2019).

The use of short-read sequencing techniques to study multiple variable regions of the 16S rRNA gene has captured the interest of researchers. There has been a rapid development in sequencing kits and bioinformatics pipelines to process multiple variable region 16S rRNA sequencing data (Callahan et al., 2021; Fuks et al., 2018; Schriefer et al., 2018; Wang et al., 2016; Amir et al., 2013). The xGen™ 16S Amplicon Panel v2 kits (Integrated DNA Technologies, Coralville, IA, United States) are an example, having been developed to amplify all nine variable regions of the 16S rRNA gene. Furthermore, a complementary bioinformatics pipeline known as the Swift Normalase Amplicon Panels APP for Python 3 (SNAPP-py3), was developed specifically for the analysis of sequencing data obtained using these kits (Chai, 2021).

Being relatively new, there are only a few publications in which the SNAPP-py3 pipeline has been used to analyse data sequenced with the xGen kits (Nuccio et al., 2023; Bennato et al., 2022). However, neither of these studies took advantage of the species-level classification that can be achieved with the xGen kits and SNAPP-py3 pipeline. Although Bennato et al. (2022) included a control containing DNA for 20 known bacterial species, they only reported the ability to pick up these bacteria at a genus level. To our knowledge, the combined ability of these kits and pipeline to obtain accurate species-level classification has not been assessed. Therefore, we sought to establish a protocol in which the SNAPP-py3 pipeline and additional processing steps were utilised to analyse short-read multiple variable region 16S rRNA data following sequencing with xGen amplicon panel kits.

The accuracy of sequencing protocols can be evaluated in a few ways using mock controls. Firstly, researchers can calculate the proportion of expected species that have been detected down to a species level when using a given protocol and for select regions of the 16S rRNA gene (Johnson et al., 2019; Fouhy et al., 2016). F-scores can be calculated based on the precision and sensitivity with which these species are identified (Özkurt et al., 2022). The classification process can also be assessed according to the percentage of overall reads that are classified as belonging to one of the expected control species (Szoboszlay et al., 2023; Urban et al., 2021). Furthermore, accuracy can be assessed by comparing observed relative abundances to expected abundances (provided by suppliers) for each taxon in a control (Maki et al., 2023; Szoboszlay et al., 2023; Drengenes et al., 2021; Laursen et al., 2017; Caporaso et al., 2011). This can be done at different taxonomic levels and gives an indication of whether the amplification or sequencing processes have introduced bias by favouring certain species over others.

As stool collection is not always possible due to various factors, rectal swab collection has become a common sampling method for studying the gut microbiome (Bassis et al., 2017). Storage of rectal swabs differs to that of stool, as swabs generally need to be placed in a medium (CDC, 2015), for example, PrimeStore (Flygel et al., 2020). The results obtained from sequencing rectal swab samples can be inconsistent in terms of numbers of bacteria detected (Chanderraj et al., 2022).

Previous studies have explored whether rectal swab samples can provide a reliable alternative to stool samples (Radhakrishnan et al., 2023; Bokulich et al., 2019; Reyman et al., 2019; Bassis et al., 2017; Freedman et al., 2017). Although pairs of stool and rectal swab samples collected concurrently from the same individual generally display similar diversity and functional profiles (Radhakrishnan et al., 2023; Reyman et al., 2019;

Bassis et al., 2017), there have been other studies that suggest that these samples are not equivalent for detecting specific taxa (Jones et al., 2018; Freedman et al., 2017; Goldfarb et al., 2014). In particular, rectal swab samples have been found to be more effective for detecting a greater number of harmful species in children with gastrointestinal infections (Freedman et al., 2017; Goldfarb et al., 2014). When sequencing the meconium (first stool) sample passed by newborn infants, rectal swab samples have been found to provide a less accurate representation of the microbiome compared to stool samples (Graspeuntner et al., 2023). Moreover, rectal swab samples provide a poorer representation of the microbiome if they are sequenced after greater than 48 h at room temperature (Bokulich et al., 2019). As a result, the interchangeability of stool and rectal swab samples needs to be assessed for new protocols. Moreover, sample collection approach is an important variable to consider with regards to the research objectives of a study.

The first aim of this research involves assessing the accuracy of extraction and sequencing protocols to achieve classification at the species level. This will be done by sequencing and analysing mock controls containing either whole cells or already-extracted DNA from eight known bacterial species, assessing the precision and sensitivity with which species were identified and how well their relative abundances matched theoretical abundances. Secondly, our goal is to evaluate the within-run and between-run reproducibility of species-level analysis. Technical replicate samples sequenced on either the same plate (within-run) or across different sequencing plates and runs (between-run), will be compared to determine this. Finally, we aim to identify whether there are differences at a species level between different sample collection approaches. To achieve this, we will compare pairs of stool and rectal swab samples collected from the same participants at the same time point (0–5 week-old newborns). We hypothesise that by sequencing multiple variable regions of the 16S rRNA gene and using the SNAPP-py3 pipeline, we could obtain accurate species-level resolution and achieve reproducible results.

2 Methods

2.1 Extraction controls, sequencing controls and technical replicates

To assess the reproducibility and accuracy of DNA extraction and sequencing steps, mock controls and technical replicates were included on each plate (Figure 1). A ZymoBIOMICS™ Microbial Community Standard (catalog number ZR D6300), consisting of eight known bacterial species including both gram-positive and gram-negative bacteria, was included on each plate as a DNA extraction control. Each plate included a ZymoBIOMICS™ Microbial Community DNA Standard (catalog number ZR D6305), which contains already extracted DNA for eight known bacterial species. This served as a sequencing control. Seventeen within-run and 8 between-run technical replicate pairs were also included across the sequencing plates. These included samples that were collected across several different time points during infancy, specifically 0–5 weeks, 3 months, 6 months, 9 months and 12 months.

2.2 Stool and swab sample collection

Twenty six pairs of stool and rectal swab samples were collected at the age of 0–5 weeks (baseline samples), for infants born between 37 and 42 weeks gestational age. Swabs were stored in Primestore solution (PrimeStore® Molecular Transport medium). All stool and swab samples were transferred to a −80°C freezer for long-term storage.

Stool samples were thawed, and half of a pea-sized scoop was collected from the side/centre of the sample. This was placed in a tube with 750 µL of lysis buffer. For rectal swab samples, 400 µL of sample in Primestore was placed in a tube with 400 µL of lysis buffer. These then underwent off-board lysis, using the QT Qiagen bead beater, prior to DNA extraction.

2.3 DNA extraction, preparation of sequencing library and illumina sequencing

Manual DNA extraction of the stool and rectal swab samples and mock extraction controls was carried out using the Quick-DNA™ Fecal/Soil Microbe Microprep Kit (ZymoBIOMICS catalog number D6012). Prior to carrying out polymerase chain reaction (PCR), Qubit™ (Thermo Fisher Scientific, Waltham, MA, United States) was done to check the starting DNA concentrations. xGen™ 16S Amplicon Panel v2 kits (Integrated DNA Technologies, Coralville, IA, United States) were used in library preparation for sequencing. Kits included primer pairs for amplification of all nine hypervariable regions of the 16S rRNA gene. Additionally, the primers for these kits have dual indices to allow greater numbers of samples to be run together in a single flow cell. Moreover, the xGen kits include Normalase™ which could be used to enzymatically normalise library sizes prior to sequencing. Finally, qPCR was performed following the Normalase step to quantify the final library size prior to sequencing.

Negative controls, including Primestore, Milli-Q water, elution buffer and Tris EDTA/Nuclease free water, were added to each plate (Figure 1) together with prepared libraries from the stool and rectal swab samples. Mock controls and technical replicates, as described above, were also included on each plate. Sequencing was conducted across two sequencing runs on seven plates. The combined 16S library per run was subjected to paired-end sequencing on the Illumina® MiSeq™ platform, employing the MiSeq Reagent v3 kit with 600 cycles (Illumina, San Diego, CA, United States).

2.4 Bioinformatics processing of sequencing data

Ethics approval for this research was provided by the Human Research Ethics Committees at the University of Cape Town (801/2016 and 557/2020) and at Stellenbosch University (M16/10/041). Following sequencing, preprocessing steps were carried out for quality control and to prepare the data for statistical analysis (Figure 2). Raw sequencing data was run through FastQC (Andrews, 2010) to assess the quality of the reads. Following this quality control step, forward and reverse reads for each sample were processed using the SNAPP-py3 pipeline (Chai, 2021). Sequencing

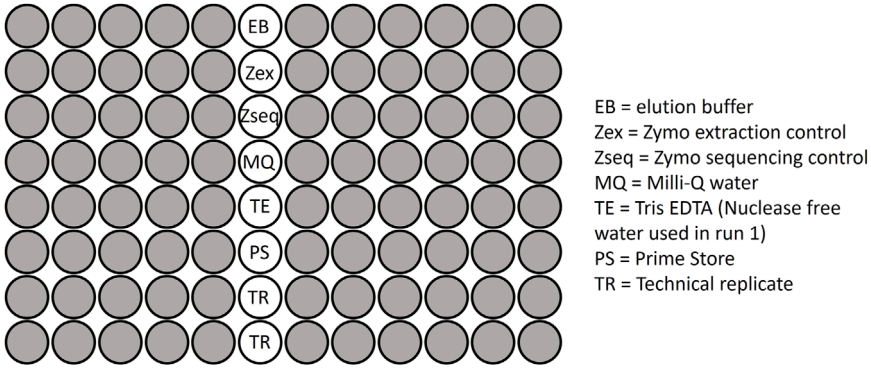

FIGURE 1
The plate layout used for sequencing microbial samples, including the negative controls, positive (mock) controls and replicates.

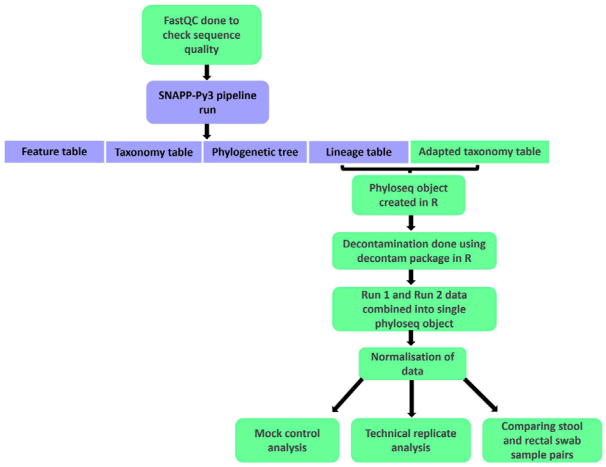

FIGURE 2
A flowchart summarising the processing and analysis steps carried out using raw forward and reverse read sequencing files. Boxes in green indicate the steps included in addition to the main SNAPP-py3 pipeline.

data from run 1 and run 2 were processed separately. Of the four main output files from the pipeline, the lineage table and an adapted taxonomy table were used for further analysis.

The remaining processing and analysis were carried out in R version 4.2.1 (R Core Team, 2022). This stage of processing began with creating a phyloseq object (McMurdie and Holmes, 2015; McMurdie and Holmes, 2013). Using the decontam package (Davis et al., 2018), decontamination was carried out separately for each plate, using plate-specific negative controls. A combined frequency and prevalence approach was used, selecting a threshold of 0.1 for the prevalence component. Phyloseq objects from runs 1 and 2 were then combined into a single phyloseq object for further downstream processing and analysis.

A normalisation step to account for different library sizes was implemented by determining the median library size and normalising each sample accordingly (Balle et al., 2020; The Jackson Laboratory, 2019). Finally, subsetting into various phyloseq objects was done to prepare for downstream statistical analysis. We ultimately had separate phyloseq objects containing the mock extraction controls, mock sequencing controls, within-run repeats, between-run repeats and baseline pairs of stool and rectal swab samples. Excel spreadsheets containing this data, as well as the corresponding code for importing the files into R as phyloseq objects, are provided in the Supplementary Datasheets S1, S3, respectively.

Batch effect correction was done using MMUPHin (Ma, 2022; Ma et al., 2022). This data was compared to data in which no batch effect correction was carried out, to determine the necessity of accounting for batch effects.

2.5 Statistical analysis

Relative abundances for controls, replicates and stool/rectal swab sample pairs were visualised using QIIME2 software (Bolyen et al., 2019). All microbiome analysis was done at a species level in R. Genus-level and phylum-level analyses were additionally included in select steps to provide additional insights.

Performance measures were calculated as outlined by Özkurt et al. (2022) using data for sequences classified to a species-level. Precision and sensitivity were calculated based on the number of correctly identified species expected to be in the mock control [true positives (TP)], the number of expected species that were not detected [false negatives (FN)] and the number of non-expected species classified as being in the control [false positive (FP)]. F-scores could then be calculated based on these values. The calculations used were as follows:

$$\text{Precision} = TP/(TP + FP)$$

$$\text{Sensitivity} = TP/(TP + FN)$$

F-score = 2*precision*sensitivity/(precision + sensitivity) (Özkurt et al., 2022).

The percentage relative abundances of each of the eight expected species were determined for mock cell (extraction) and mock DNA

(sequencing) controls on each sequencing plate. For each control, the total percentage of sequences that were correctly classified as an expected species, was calculated. Furthermore, observed relative abundances were compared to the theoretical abundances provided by ZymoBIOMICS for each species in the mock controls by calculating Observed/Expected (O/E) ratios (Maki et al., 2023).

Functions from the phyloseq package in R (McMurdie and Holmes, 2015; McMurdie and Holmes, 2013), specifically the estimate_distance and distance function, were used to calculate alpha and beta diversity measures for replicates and stool/rectal swab samples. The alpha diversity measures included are observed richness (Fisher et al., 1943), Shannon's index (Shannon, 1948) and Simpson's index (Simpson, 1949). Bray Curtis (Bray and Curtis, 1957) and Jaccard's (Ludwig and Reynolds, 1988) distances were the beta diversity measures included in our analysis. Plot_richness and plot_ordination functions were used to plot alpha and beta diversity measures, respectively. Paired Wilcoxon tests were used to compare alpha diversity measures between pairs of within-run replicates and to identify differences between pairs of between-run replicates. In order to determine whether beta diversity measures were reproducible between technical replicate pairs, distance-based intraclass correlation coefficients (dICCs) were calculated separately for within-run and between-run replicates (Chen and Zhang, 2022). Paired Wilcoxon tests and dICCs were similarly used to compare pairs of stool and rectal swab samples collected from the same infants.

3 Results

Analysis of mock controls and technical replicates was carried out to assess the use of the xGen Amplicon kits and the SNAPP-py3 pipeline as a multivariate 16S rRNA sequencing approach for achieving accurate and reproducible species-level resolution. Furthermore, the similarity of samples collected using different sample collection techniques was investigated by comparing pairs of baseline stool and rectal swab samples from the same participants.

3.1 DNA extraction reliability

All eight expected bacterial species were detected in four of our seven mock extraction controls (Table 1). *Bacillus subtilis* was not detected to the species level in two controls (Table 2; Figure 3A), however classification to a genus level (*Bacillus*) was achieved (Supplementary Table S1). *Listeria monocytogenes* was not detected even at a genus level in the control from run 2, plate 1 (Supplementary Table S1). For the three controls in which we were unable to detect all eight species, the total percentage of sequencing data correctly classified as expected mock species was consequently lower (Table 2). Sensitivity scores were over 0.88 for all controls. Precision scores were lower – particularly for the control on run 1, plate 1, which had a score of 0.32. This was driven by a high number of false positive results in this control. A median F-score of 0.84 (range of 0.47–1.00) was obtained for the mock extraction controls (Table 1).

The percentage abundances of species in these controls did not accurately follow the order of theoretical abundances for some species (Table 2; Figure 4A). In particular, the relative abundances of *L. monocytogenes* were well below the theoretical abundances suggested by ZymoBIOMICS at both a species and genus level. This is emphasised by the low median Observed/Expected (O/E) ratio of 0.32 at a species level. Similarly, *Enterococcus faecalis* and *Staphylococcus aureus* had O/E ratios well below the value of 1. The relative abundances of *Escherichia coli* and *Salmonella enterica* were greater than expected, with a range of O/E ratios all lying well above the value of 1.

3.2 Sequencing reliability

Among the mock sequencing controls, the eight anticipated bacterial species were detected in five of the seven controls (Table 3; Figure 3B). The overall percentages of sequences correctly classified as expected species were slightly lower for these controls compared to the mock extraction controls (Table 4). For the run 2 plate 1 control the prevalence of *S. enterica* was particularly low, and this was not resolved at the genus level (Supplementary Table S2). *B. subtilis* again was not detected at a species level for two of these controls (Table 3). Precision scores for the mock sequencing controls ranged from 0.50 and up, while sensitivity was greater than 0.88 for all controls. F-scores had a median of 0.80 (range of 0.67–0.94).

The relative abundances of species in the mock sequencing controls more closely matched the expected abundances than was observed for the mock extraction controls (Table 4; Figure 4B), as seen by the O/E ratios being closer to 1. In these controls the abundance of *Limosilactobacillus fermentum* was well below the theoretical threshold expected. The O/E ratios for this species and *S. aureus* were consistently less than 1. Whereas *E. faecalis*, *E. coli* and *L. monocytogenes* had O/E ratio ranges above 1.

Although the inclusion of a batch effect correction step was trialled, substantial differences in the relative abundances of the various species across the controls were observed compared to when no batch effect correction was done (Supplementary Figure S1). Similarly, stricter decontamination thresholds led to poorer reproducibility in mock controls.

3.3 Within-run and between-run reproducibility

Similar patterns in the relative abundance of species could be seen when comparing pairs of within-run and between-run repeats (Supplementary Figures S2, S3). When comparing alpha diversity of within-run repeats using paired Wilcoxon tests, we found no evidence to indicate differences at a species level in the Observed richness [95% confidence interval (CI) (−3.50, 5.50); p = 0.587], Shannon's index [95% CI (−0.09, 0.12); p = 0.782] or Simpson's index [95% CI (−0.02, 0.01); p = 0.487] between these pairs of technical replicates (Figure 5). Furthermore, when comparing the beta diversity distance matrices of these within-run technical replicate pairs (Figure 6), we observed a good level of reproducibility for Bray Curtis as seen by a distance-based intraclass correlation coefficient (dICC) value of 0.940 and Jaccard's distance showed good, albeit lower, reproducibility with a dICC of 0.762. Similarly,

TABLE 1 A summary of performance and accuracy measures at a species level for mock cell controls containing eight known bacterial species.

	R1P1 Zymoex	R1P2 Zymoex	R1P3 Zymoex	R1P4 Zymoex	R2P1 Zymoex	R2P2 Zymoex	R2P3 Zymoex
True positives	7	8	8	7	7	8	8
False positives	15	3	3	0	3	0	2
False negatives	1	0	0	1	1	0	0
Precision	0.32	0.73	0.73	1.00	0.70	1.00	0.80
Sensitivity	0.88	1.00	1.00	0.88	0.88	1.00	1.00
F-score	0.47	0.84	0.84	0.93	0.78	1.00	0.89

R#, run number; P#, plate number.

we found no differences between these technical replicates when looking at their alpha and beta diversity measures at a genus level.

A comparison of alpha and beta diversity measures for between-run technical replicates similarly found no clear differences between these pairs at a species level (Figure 7). Paired Wilcoxon tests comparing Observed species [95% CI (−2.00, 13.00); p = 0.362], Shannon's index [95% CI (−0.05, 0.49); p = 0.195] and Simpson's index [95% CI (−0.004, 0.11); p = 0.148] did not motivate for the existence of differences in alpha diversity measures. Moreover, dICC results showed good reproducibility for Bray Curtis (dICC = 0.899) and moderate reproducibility for Jaccard's distance (dICC = 0.597) (Figure 8).

3.4 Interchangeability of stool and rectal swab samples

To assess whether stool and rectal swab samples may be combined for analysis, we compared 26 sample pairs collected at 0–5 weeks after birth. The relative abundances of species did not display similar patterns across pairs of samples (Supplementary Figure S4) and alpha diversity was found to differ when running paired Wilcoxon tests. Stool and swab samples from the same participants differed at a species level in terms of Observed species ($p < 0.001$) and Shannon's index ($p = 0.027$), with no differences in Simpson's index ($p = 0.394$) found (Figure 9). A comparison of beta diversity measures at a species level found that while Bray Curtis measures were moderately reliable between the pairs of stool and swab samples (dICC = 0.684), there was poor reliability when comparing their Jaccard's distances (dICC = 0.310) (Figure 10).

4 Discussion

We have outlined a short-read sequencing protocol which can be used to carry out species-level analysis. Our findings, following analysis of mock controls and technical replicates, indicate that the kits and analytical pipelines used in this study can effectively enable species-level classification and provide reproducible results within and across sequencing plates.

When assessing data from different sample collection approaches using this protocol, our results indicate that care needs to be taken as stool and swab samples collected from the same participant are not comparable at a species level.

4.1 Multiple variable regions of 16S rRNA enable adequate species-level analysis

We were able to obtain relatively good species-level resolution in terms of sensitivity using a nearly complete 16S rRNA sequence, which the SNAPP-py3 pipeline developers refer to as a 'consensus' sequence. The eight species expected to be found in the ZymoBIOMICS mock controls were not consistently detected to a species level in all mock controls across the seven sequencing plates, suggesting that there is still room for improvement in terms of reproducibility. The main species which were not detected in all controls are *B. subtilis* and *L. monocytogenes*. These species are gram-positive, containing a strong wall of peptidoglycan which can be a challenge to lyse, as has been presented in previous literature (Claassen-Weitz et al., 2020). ZymoBIOMICS have intentionally developed these mock controls to include both gram-negative and gram-positive species to enable researchers to identify inconsistencies and to optimise their lysis protocols (ZymoBIOMICS, 2024). Moreover, research indicates that the primers used in library preparation may have a greater affinity for some species compared to others (Klindworth et al., 2013), which may also contribute to false negatives in some controls. Our results indicate that the methods used are capable of detecting all species, as seen in several of our mock controls, however future optimisation of the protocol will be required.

A previous study comparing the use of different variable regions in Illumina Miseq sequencing, found that using 1-2 variable regions could at best identify 16 of 20 (80%) mock control species correctly (Fouhy et al., 2016). In five mock controls, we detected 7 of 8 mock species (88%), yet for the remaining nine controls we detected all the expected species (100%). Thus, our results would indicate that using all nine variable regions of the 16S rRNA gene improves accuracy compared to methods that look at only a couple of variable

TABLE 2 The theoretical abundances and relative abundances of the eight expected bacterial species in mock cell controls given as percentages. The median observed/expected (O/E) ratios are also provided. The total percentage of sequences correctly classified as one of the expected species is given for each control.

Species	Theoretical abundance (%)	R1P1 Zymoex	R1P2 Zymoex	R1P3 Zymoex	R1P4 Zymoex	R2P1 Zymoex	R2P2 Zymoex	R2P3 Zymoex	O/E ratio [median (range)]
Limosilactobacillus fermentum	18.4	18.21	18.27	17.99	17.64	18.60	16.60	18.02	0.98 (0.90–1.01)
Bacillus subtilis	17.4	0.00	19.55	19.71	0.00	18.76	17.63	18.55	1.07 (0.00–1.13)
Staphylococcus aureus	15.5	10.82	9.13	9.13	11.47	10.96	9.31	10.75	0.69 (0.59–0.74)
Listeria monocytogenes	14.1	4.41	4.47	4.59	4.97	0.00	4.58	4.67	0.32 (0.00–0.35)
Salmonella enterica	10.4	17.58	18.90	17.61	16.37	10.95	20.39	19.22	1.69 (1.05–1.96)
Escherichia/Shigella coli	10.1	17.06	18.29	18.26	18.89	20.57	19.26	20.50	1.87 (1.69–2.04)
Enterococcus faecalis	9.9	5.39	4.53	4.53	5.28	5.36	4.96	4.81	0.50 (0.46–0.54)
Pseudomonas aeruginosa	4.2	7.37	6.52	7.77	4.50	4.63	7.25	3.40	1.55 (0.81–1.85)
Other	0	19.16	0.34	0.42	20.89	10.18	0.02	0.09	-
% Correctly classified	100	80.84	99.66	99.58	79.11	89.82	99.98	99.91	-

R#, run number; P#, plate number.

10.3389/fbinf.2025.1484113

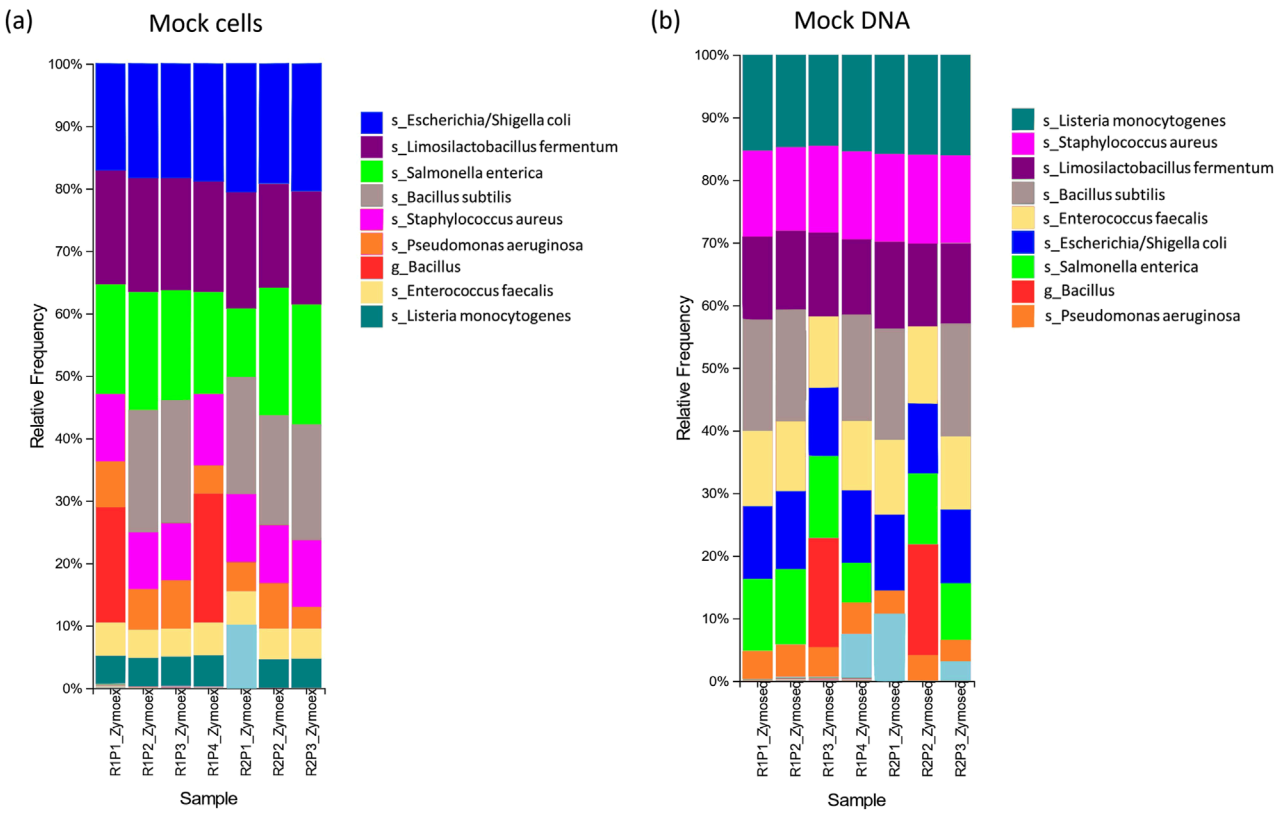

FIGURE 3
Relative abundances for **(A)** mock cell and **(B)** mock DNA controls; Legend: s = species, g = genus level classification. R# = run number; P# = plate number. The legend only includes taxa of interest to the species or genus level.

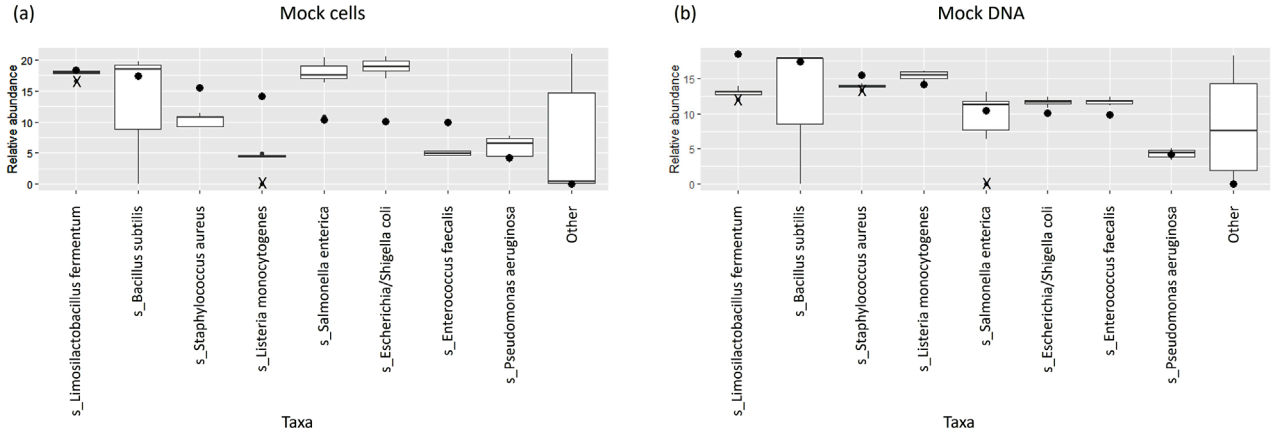

FIGURE 4
Box and whisker plots showing the relative abundance percentages of the eight expected bacterial species across **(A)** mock extraction controls and **(B)** mock sequencing controls; theoretical abundances shown as dots; outliers shown as crosses.

regions. Other short-read multiple variable region methods have been able to identify all taxa within mock controls containing a greater number of species (Fuks et al., 2018; Schriefer et al., 2018), however in these studies there were also mock controls in which not all species were identified. Schriefer et al. (2018) suggested that the depth to which sequencing was done may play a role, however an increase of sequencing depth by 10-fold ultimately had no impact on their results. Therefore, the approach we present requires further optimisation to ensure that 100% sensitivity is consistently achieved and to assess whether this approach can perform at a similar level to other multiple variable region tools when using more complex mock controls.

TABLE 3 A summary of performance and accuracy measures at a species level for mock DNA controls containing eight known bacterial species.

	R1P1 Zymoseq	R1P2 Zymoseq	R1P3 Zymoseq	R1P4 Zymoseq	R2P1 Zymoseq	R2P2 Zymoseq	R2P3 Zymoseq
True positives	8	8	7	8	8	7	8
False positives	4	8	4	6	1	0	3
False negatives	0	0	1	0	0	1	0
Precision	0.67	0.50	0.64	0.57	0.89	1.00	0.73
Sensitivity	1.00	1.00	0.88	1.00	1.00	0.88	1.00
F-score	0.80	0.67	0.74	0.73	0.94	0.93	0.84

R#, run number; P#, plate number.

When focusing only on sequences that were classified to a species level, sensitivity was high in both the mock extraction and mock sequencing controls. However, our results indicate that high sensitivity comes at the expense of obtaining poorer precision in some controls. The overall performance for mock extraction and mock sequencing controls were median F-scores of 0.84 and 0.80 respectively. These results are comparable with other research that obtained F-scores greater than 0.80 using a bioinformatics tool for analysing sequencing data, for which the authors described their results as being indicative of good performance (Özkurt et al., 2022). Poor precision scores in our case were often driven by false positive taxa that were present in very low relative abundances. Thus, removing rare taxa could yield even better results. Schriefer et al. (2018) and Fuks et al. (2018), also reported false positives at low abundances for their multiple variable region approaches.

As our measure of precision is based on the presence or absence of expected taxa, it is quite strict in how it penalises false positives. It is not weighted according to the abundances at which these occur. Therefore, it is important for us to look at the precision scores in conjunction with our abundance-based measures, to get a complete picture of the performance of the methodology used in this paper. The R1P1 mock extraction control displayed poor precision due to having a particularly high number of false positives. However, as these occurred at such low abundances, false positives had little influence on our abundance-related analysis. Rather, we found that the inability to detect all expected species (i.e., false negative results) in our controls was the main factor responsible for differences in observed abundances compared to expected abundances.

The overall percentages of sequences that were classified as expected taxa and to a species level, ranged between 79.11% and 99.98% for mock extraction controls, and between 81.80% and 99.58% for mock sequencing controls. In a study by Szoboszlay et al. (2023) the authors found that between 58.9%–68.9% of reads obtained from Illumina sequencing of the V4 region could be correctly classified to a species level. This improved if rare taxa were excluded. Nanopore sequencing of the entire 16S rRNA gene could yield classification of over 81% of reads to a species level (Szoboszlay et al., 2023). Therefore, although we did not remove rare taxa, our protocol could compete with a full-length sequencing approach and perform better than using a single variable region for short-read sequencing. Ultimately, our results indicate that the RDP classifier is able to accurately identify the expected taxa in our controls.

The relative abundances of species in our mock extraction controls sometimes diverged from the theoretical abundances outlined by the suppliers (ZymoBIOMICS, 2022), indicating that there may be some bias introduced during the extraction process. O/E ratios provided a quantitative means of assessing the bias introduced in each control. In particular, they assisted with identifying the species for which observed abundances differed most from theoretical abundances and showed that there was greater bias in the mock extraction controls.

There are various stages in 16S rRNA sequencing where prejudice can arise, favouring certain taxa or altering the relative composition of a sample (Nearing et al., 2021). Studies have found that DNA extraction kits and methods can differ in terms of their ability to extract DNA from gram negative and gram positive bacteria, indicating that the results of a study can be influenced depending on which kit and extraction method are used (Videnska et al., 2019; Yuan et al., 2012). The mock sequencing controls (for which already-extracted DNA was obtained from ZymoBIOMICS) more closely resembled the expected abundances. *L. fermentum* differed substantially from the theoretical abundance in the sequencing controls – unlike the extraction controls for which this particular species had followed the expected abundance more closely. This implies that, for the most part, there is minimal bias introduced at the sequencing step. However, there may be some bias specifically in sequencing *L. fermentum*. Our findings complement other studies which have shown that the extraction and amplification of DNA is particularly prone to prejudicing results, while there is less bias introduced at the sequencing stage of microbial research (Brooks et al., 2015; Lee et al., 2012). Similar to our findings, previous research using a short-read multiple variable region approach, also found that the observed abundance of taxa did not always accurately match the expected abundances (Fuks et al., 2018). The authors also suggested that bias may be introduced during the amplification stage of library preparation.

Our findings suggest that using this multivariate analysis approach might provide a means to improve the ability of short-read 16S rRNA sequencing studies to study microbial samples at a species

TABLE 4 The theoretical abundances and relative abundances of the eight expected bacterial species in mock DNA controls are given as percentages. The median observed/expected (O/E) ratios are also provided. The total percentage of sequences correctly classified as one of the expected species is given for each control.

Species	Theoretical abundance (%)	R1P1 Zymoseq	R1P2 Zymoseq	R1P3 Zymoseq	R1P4 Zymoseq	R2P1 Zymoseq	R2P2 Zymoseq	R2P3 Zymoseq	O/E ratio [median (range)]
Limosilactobacillus fermentum	18.4	13.08	12.60	13.32	11.94	13.88	13.16	12.86	0.71 (0.65–0.75)
Bacillus subtilis	17.4	17.89	17.84	0.00	17.05	17.81	0.00	17.99	1.02 (0.00–1.03)
Staphylococcus aureus	15.5	13.84	13.34	13.83	14.04	13.95	14.22	14.03	0.90 (0.86–0.92)
Listeria monocytogenes	14.1	15.27	14.76	14.54	15.41	15.85	15.95	16.05	1.09 (1.03–1.14)
Salmonella enterica	10.4	11.49	12.04	13.12	6.32	0.01	11.23	9.06	1.08 (0.00–1.26)
Escherichia/Shigella coli	10.1	11.59	12.40	10.84	11.62	12.18	11.20	11.64	1.15 (1.07–1.23)
Enterococcus faecalis	9.9	11.98	11.18	11.50	11.12	11.89	12.40	11.76	1.19 (1.12–1.25)
Pseudomonas aeruginosa	4.2	4.44	5.10	4.64	4.94	3.72	4.09	3.39	1.06 (0.81–1.21)
Other	0	0.42	0.74	18.20	7.54	10.72	17.74	3.21	-
% Correctly classified		99.58	99.26	81.80	92.46	89.28	82.26	96.79	-

R#, run number; P#, plate number.

10.3389/fbinf.2025.1484113

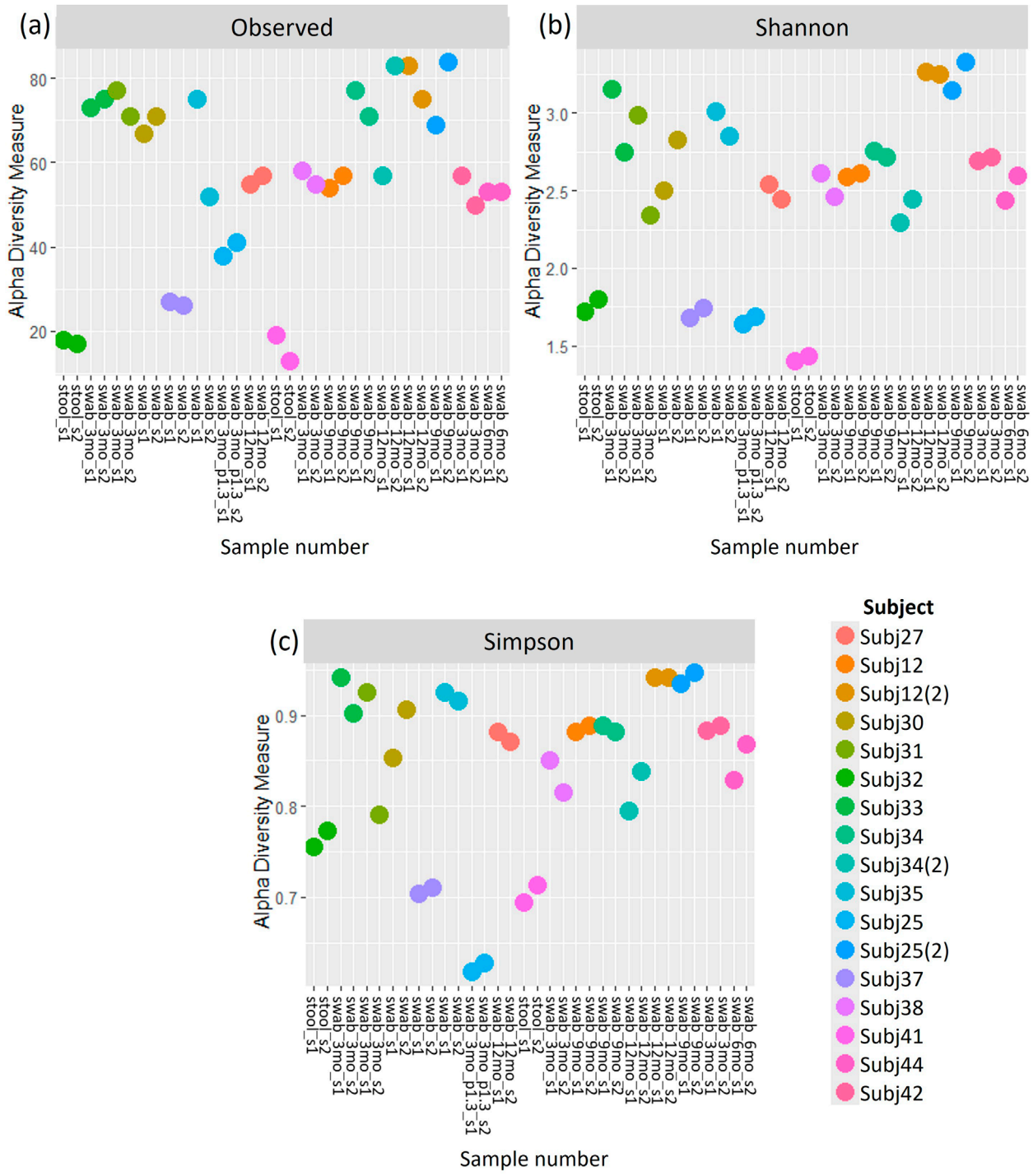

FIGURE 5
(A) Observed richness, **(B)** Shannon's index and **(C)** Simpson's index for within-run technical replicates at species-level classification. s1, sample 1; s2, sample 2.

level. This approach would benefit from additional ways to minimise bias – particularly in the DNA extraction process.

We considered including a batch effect correction step in the processing of our data. The goal was to reduce bias introduced due to samples being sequenced on different plates and in different runs. However, when visualising the relative abundances of mock sequencing controls, we found that batch effect correction led to far less consistent results across the controls. Davis et al. (2018) reported that the decontam package corrects for batch effects. As a result, it may be redundant to carry out both decontamination and batch effect correction. Moreover, as our protocol included a plate-wise decontamination step, batch effects would have already been taken

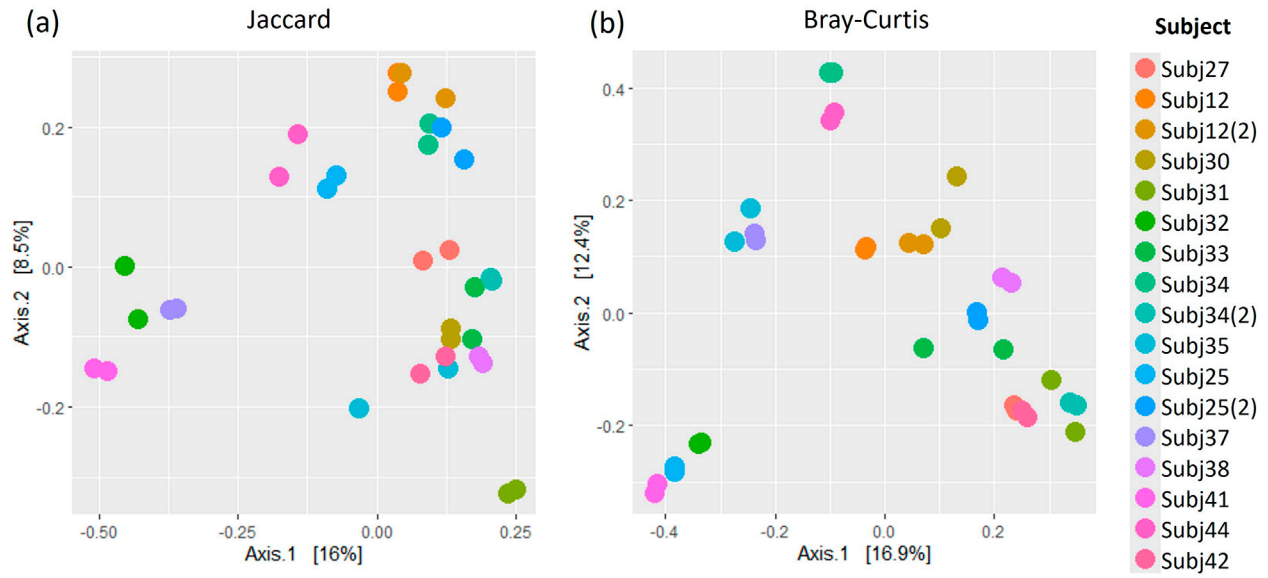

FIGURE 6
A comparison of **(A)** Jaccard and **(B)** Bray Curtis beta diversity measures between pairs of within-run technical replicates at species-level classification.

into account here. Thus, based on our findings, we considered an additional batch effect correction step excessive and problematic – leading us to exclude this step. Similarly, we found that using a very stringent threshold for decontamination also had a negative impact on our analysis. As such, we would suggest that care needs to be taken when doing additional processing of data, to ensure that bias is not introduced to the data by overcorrecting for contaminants and batch effects. Researchers should take care in deciding whether batch effect correction is necessary if their pipeline includes a decontamination step, which removes contaminants according to batches.

4.2 Short-read multiple variable region analysis can achieve good reproducibility

Observed richness provides an indication of the number of different taxa within samples, while Shannon's and Simpson's indices additionally account for how equally represented these taxa are within samples (evenness) (Kers and Saccenti, 2021). Based on our results we have no reason to reject the null hypothesis that there is no difference in alpha diversity measures between within-run technical replicate pairs.

Jaccard's distance quantifies how diversity varies between samples based on whether or not taxa are present, while Bray Curtis factors in the abundance of taxa (Kers and Saccenti, 2021; Schroeder and Jenkins, 2018). More confidence is usually placed in Bray Curtis compared to Jaccard's distance (Schroeder and Jenkins, 2018). The similarity between these beta diversity measures for replicate sample pairs was assessed in terms of dICC. dICC is a distance measure which has been established specifically for microbiome data, building on the concept of intraclass correlation coefficients, to look at the similarity between replicate samples (Chen and Zhang, 2022). Higher ICC values are indicative of strong similarity between measures (Koo and Li, 2016), in our case diversity measures for microbial sample pairs. An ICC of 0.5, however, shows moderate reproducibility and a value of <0.5 suggests poor similarity (Koo and Li, 2016). Thus, our results in which we get dICC values of 0.940 and 0.762 for Bray Curtis and Jaccard respectively, indicate that we get good beta diversity reproducibility when running samples in duplicate on the same plate. Likewise, the diversity of between-run technical replicates was similar with dICC values of 0.899 and 0.597. This suggests that we do not have significant batch effects across sequencing plates of the same run, or across different runs, when following the protocols set out in this study for sequencing and processing short-read multivariate 16S rRNA data. It should be noted that we only had one pair of replicates across runs and this was grouped with the between-plate replicates in what we referred to as our "between-run replicates." Collectively our findings indicate that there is no batch effect introduced. The sequencing and processing of infant stool and rectal swab samples is consistent, providing confidence in the analysis of the microbiome datasets generated using this protocol.

4.3 Stool and rectal swab collection approaches are not interchangeable

The implementation of a new kit and pipeline for analysing both stool and rectal swab samples in this study warranted investigation. Stool and rectal swab samples collected from the same infants and at the same time point were found to differ for several diversity measures. Contrary to our findings, studies have suggested that rectal swab samples can be used in place of stool samples and provide a reliable representative measure (Radhakrishnan et al., 2023; Reyman et al., 2019; Bassis et al., 2017). However, a couple of studies exploring diversity and functional roles of the microbiome have previously reported differences between stool and swab samples (Short et al., 2021; Sun et al., 2021). Short et al. (2021) specifically found that beta diversity differed between these sample types, while Sun et al. (2021) found that rectal swab samples varied from stool

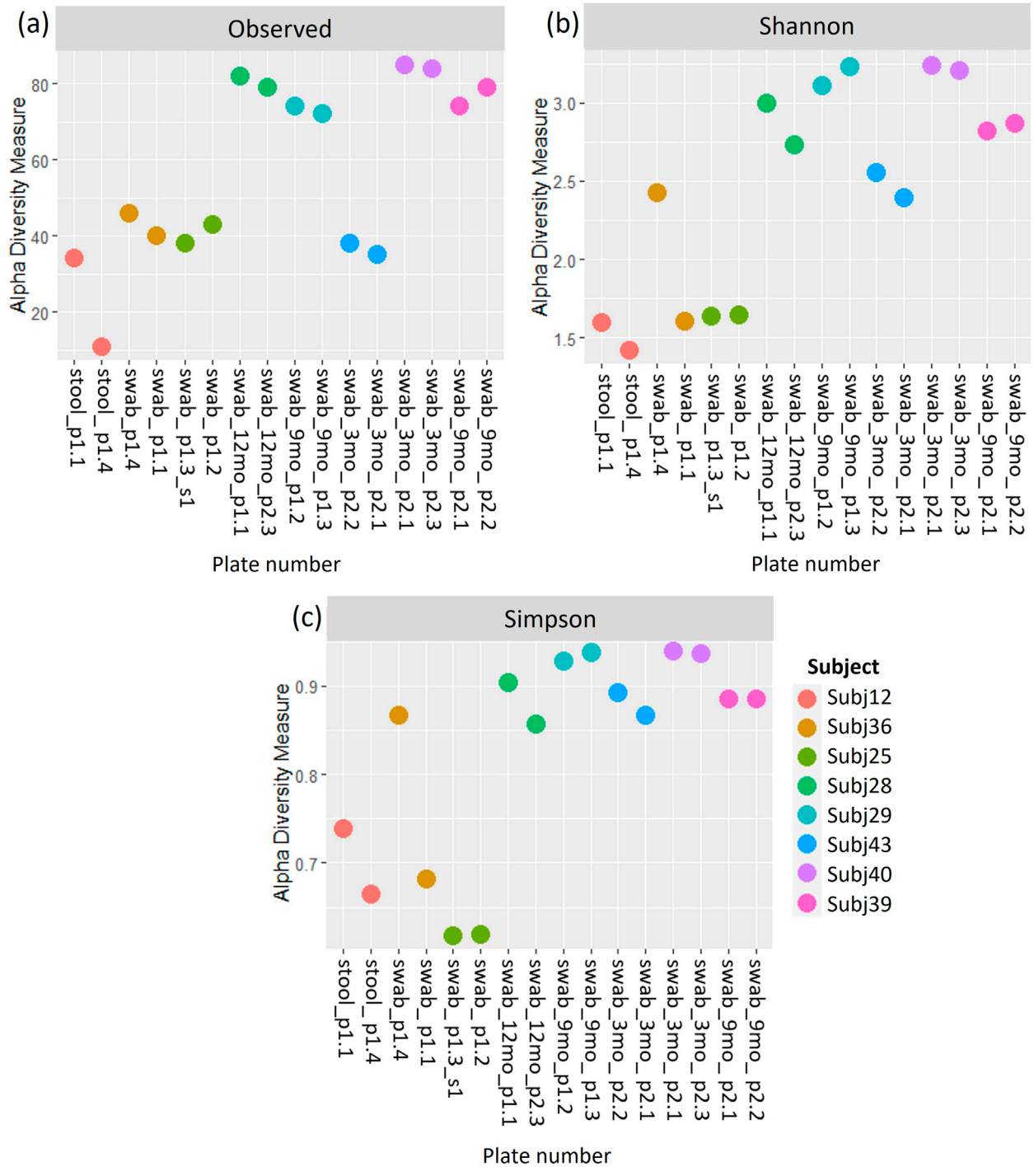

FIGURE 7
(A) Observed richness, **(B)** Shannon's index and **(C)** Simpson's index for between-run technical replicates at species-level classification. p# = run and plate number.

in terms of both alpha and beta diversity measures. Moreover, there have been indications that differences in the relative abundances of specific taxa may occur between different sample types (Jones et al., 2018). This would be particularly important to consider when comparing the prevalence of individual taxa between groups, such as when using differential abundance analysis.

Furthermore, a study by Bokulich et al. (2019) found that the length of time between the collection and processing of samples is an important factor to take into consideration when using rectal swab samples in sequencing studies. Although samples in their study were ultimately frozen, swabs were posted to the laboratory, while stool samples were immediately placed on ice until they could be

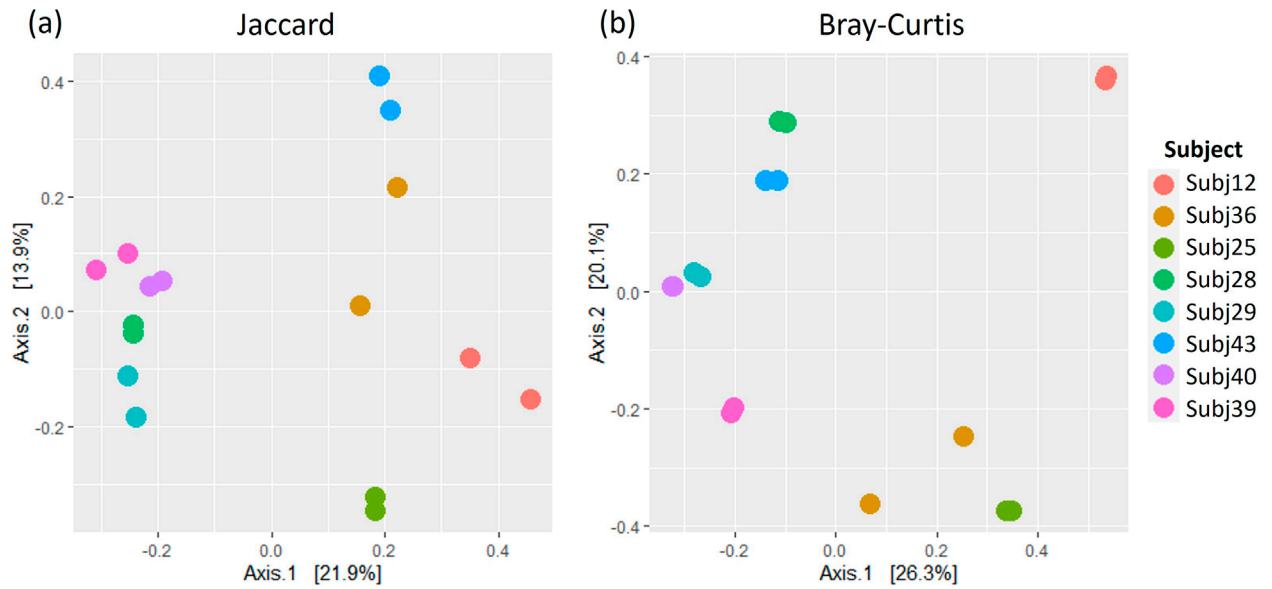

FIGURE 8
A comparison of **(A)** Jaccard and **(B)** Bray Curtis beta diversity measures between pairs of between-run technical replicates at species-level classification.

collected and taken to the laboratory. Rectal swab samples received and processed after 48 h following sample collection did not accurately match stool samples – with a higher proportion of *Enterobacteriaceae* being favoured (Bokulich et al., 2019). The stool and rectal swab samples in our study were collected together and stored at −80°C until DNA extraction and sequencing were done, and therefore we do not face the same challenges as Bokulich et al. (2019) in terms of the length of time for which samples remained at room temperature. However, it may be important to consider factors such as the time for which samples are stored in freezers, and the time between DNA extraction and sequencing in future work involving both stool and swab samples.

The fact that swabs were stored in Primestore, while stool was not, would suggest a substantial difference which requires consideration when using a combination of stool and swab samples in carrying out analysis. Several studies have found that bacterial compositions differ across different regions of entire stool samples (Zreloff et al., 2023; Huson et al., 2017; Gorzelak et al., 2015). Spectroscopic analysis has shown that patterns of metabolites can also vary substantially across different regions of stool (Liang et al., 2020; Gratton et al., 2016). Therefore, using only part of the sample does not provide a good characterisation of the entire sample. These studies highlight the importance of homogenising the entire stool sample in order to carry out sequencing (Zreloff et al., 2023; Liang et al., 2020; Gratton et al., 2016). As sequencing of rectal swab samples is not limited to only a component of the collected sample, this sample would be homogenous. This may explain the differences observed between stool and rectal swab sample pairs.

4.4 Limitations and future work

The work presented included only one mock extraction and one mock sequencing control on each sequencing plate. Future work could include duplicates of controls on each plate to better assess reproducibility of the controls. It would also expand the options available for statistically comparing these controls.

A limitation of this study is the inconsistent detection of all eight species in each mock control. We were able to detect all eight species in several controls, suggesting that the methods used for library preparation, sequencing and processing of the data are all capable of achieving consistent detection. Given that gram positive species particularly are not always detected (Claassen-Weitz et al., 2020), and these are more challenging to lyse, future work to optimise the lysis protocol should be done – for example, assessing whether carrying out mechanical lysis for longer enables the consistent detection of all species. A study involving sequencing of the entire genome, has previously shown that sequencing depth may play a crucial role in whether or not all expected taxa are detected (Pereira-Marques et al., 2019). Sequencing depth is a variable that differed for each control and may explain the inconsistencies in our results. Sequencing was done based on the Illumina guidelines which recommend a library depth of at least 100,000 reads (ILLUMINA. 16S Metagenomic Sequencing Library Preparation, 2024). However, as we are using a pipeline which combines multiple variable region reads into consensus sequences, and as other short-read multiple variable region 16S rRNA studies have utilised greater sequencing depths (Fuks et al., 2018; Schriefer et al., 2018), future work could seek to optimise the sequencing depth required for this protocol in order to better identify all species in mock controls.

In this study we evaluated a single pipeline and protocol for analysing data from multiple variable region 16S rRNA sequencing, however in future this protocol would benefit from comparisons to other tools and pipelines. Moreover, different DNA extraction kits and amplicon sequencing kits could be compared to identify the best kits for

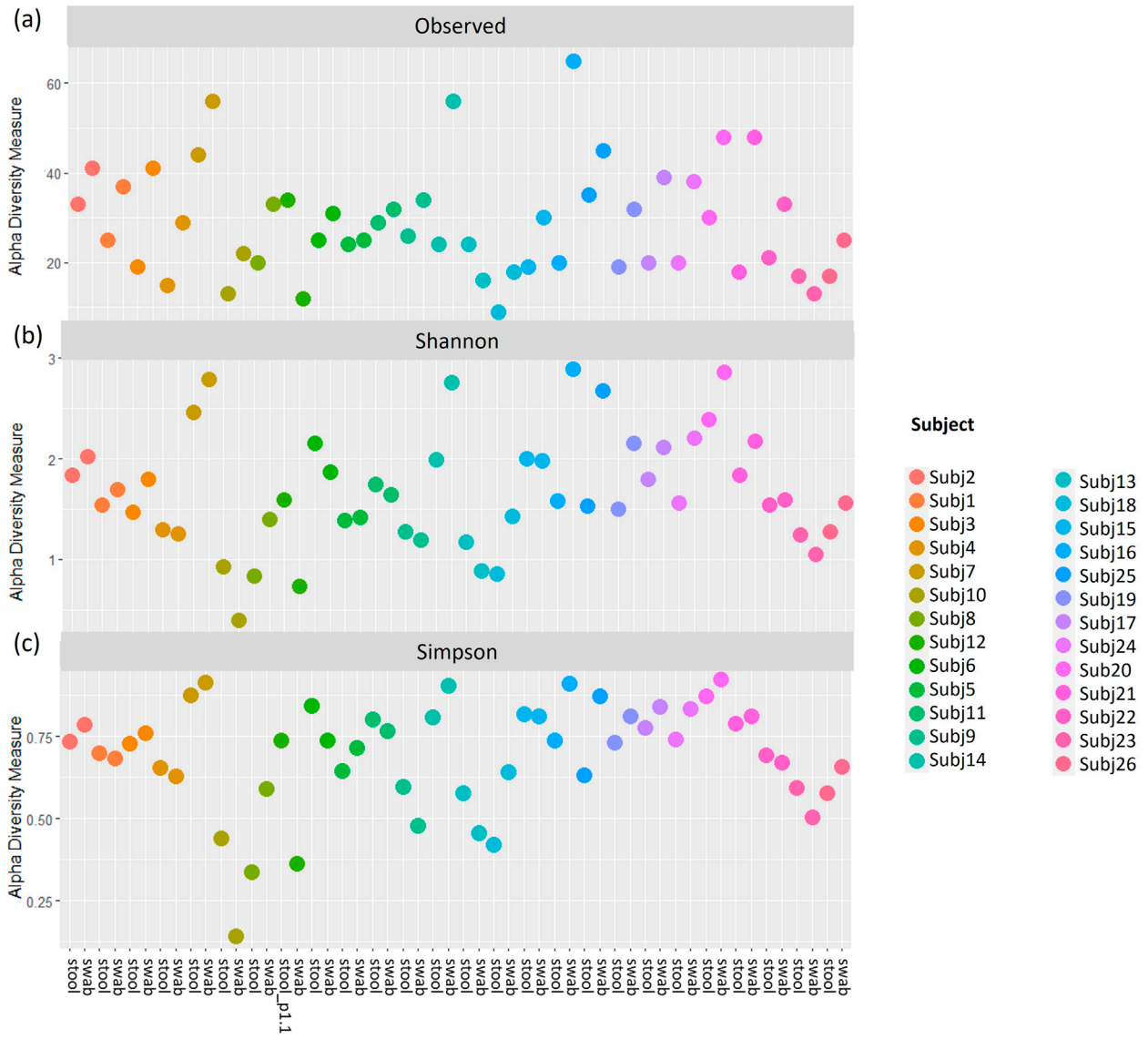

FIGURE 9
(A) Observed richness, **(B)** Shannon's index and **(C)** Simpson's index for pairs of stool and rectal swab samples collected from the same infants using data classified to a species level. p# = run and plate number.

carrying out multiple variable region analysis. In this analysis we have worked with compositional data, looking at the relative abundances of taxa. However, in future, it would be useful to explore absolute abundances for samples sequenced using this protocol. Moreover, we could explore whether other databases – such as databases specific to the human gut microbiome – might enable more effective and accurate classification of consensus sequences from the SNAPP-py3 pipeline.

Researchers intending to utilise this sequencing kit and pipeline should ideally strive (where possible) to use data from samples obtained through the same collection approach. In cases where a combination of sample collection approaches is used, analysis should be done to assess whether collection approaches influence microbial composition. If they do, sample collection approach should be controlled for in any analysis in which this data is used.

The goal of this paper was to assess the use of xGen kits and the SNAPP-py3 pipeline, to set the stage for an observational clinical study. Our assessment of this methodology is limited by the fact that we did not have access to a wider variety of datasets sequenced using the xGen kits. Future research should explore the use of this methodology in a greater selection of clinical settings and in the study of environmental microbiome samples. This might result in the approach being of greater relevance and interest to a broader audience.

4.5 Conclusion

The results of our 16S rRNA multiple variable region analysis, using short-read Illumina sequencing data from all 9 variable regions, show that there is promise for using this protocol for species-level analysis. We have identified areas for improvement and future work should assess whether this approach is comparable to other bioinformatics pipelines and tools that have been established

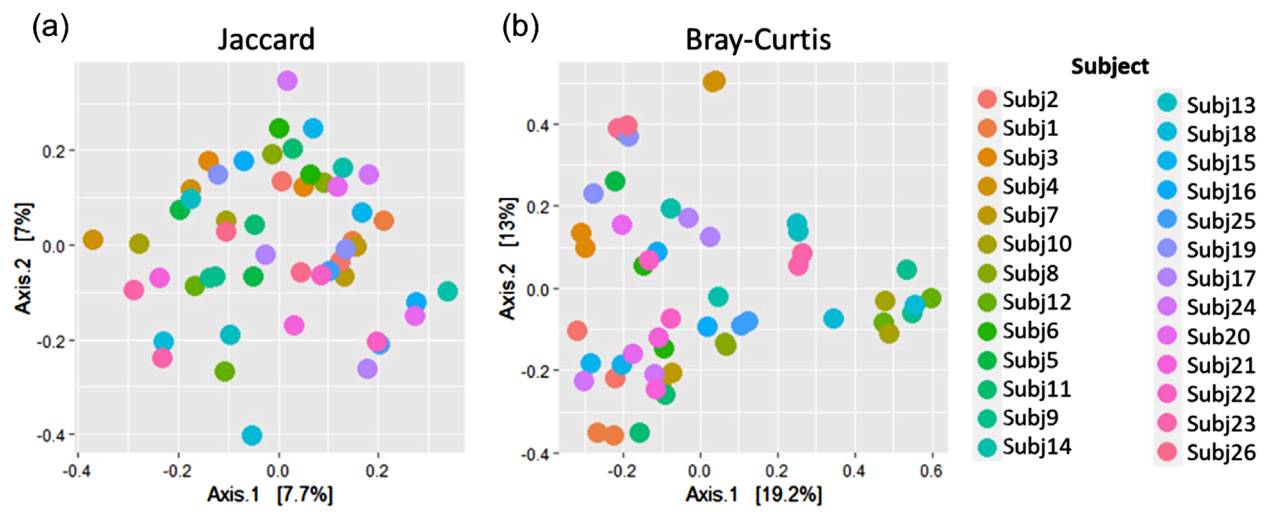

FIGURE 10
Principal coordinates analysis plots showing **(A)** Jaccard and **(B)** Bray Curtis beta diversity measures for pairs of stool and rectal swab samples collected from the same infants at species-level classification.

for analysing multiple variable region short-read data. As we found differences in diversity between stool and swab sample pairs, future analysis of data which has been generated using this protocol should take this into account. Furthermore, our findings indicate that when using new sequencing kits and protocols to study both stool and swab samples, it would be advisable to do analysis to compare pairs of stool and rectal swabs from the same individual to confirm whether these yield comparable results.

Data availability statement

The data presented in the study are deposited in the Sequence Read Archive (SRA) repository, BioProject accession number PRJNA1195741.

Ethics statement

This study involving humans was approved by the Human Research Ethics Committees at the University of Cape Town (801/2016 and 557/2020) and Stellenbosch University (M16/10/041). The study was conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

Author contributions

AG: Writing–original draft, Formal Analysis, Visualization. FP: Writing–review and editing, Methodology. FL: Writing–review and editing, Supervision. AK: Funding acquisition, Writing–review and editing. MK: Funding acquisition, Supervision, Writing–review and editing, Conceptualization. MH: Funding acquisition, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work has been supported in part by the National Research Foundation of South Africa (Grant Number: MND200610529926, NRF Postgraduate Scholarships). The National Institutes of Health (NIH) (Fogarty International Center (FIC) and National Institute of Child Health and Human Development (NICHD) R01HD093578 and R01HD085813) provided funding for this research.

Acknowledgments

We would like to acknowledge and thank Dr Veronica Allen for assisting Dr Fadheela Patel with DNA extraction and the 16S rRNA sequencing. We thank Dr Samantha Fry and Thandiwe Hamana for organizing visits and the sometimes difficult task of sample collection. We thank Dr Farai Mberi and Caylin Mc Farlane for their assistance with sample transportation and storage. Our sincere thanks to Slindile Mbhele for all she did to arrange the storage and transportation of samples as project manager. Thank you also to Dr Benli Chai, the developer of the SNAPP-py3 pipeline, for providing assistance and advice. The original preprint for this manuscript can be accessed at https://www.biorxiv.org/content/10.1101/2024.05.13.591068v1. (Graham et al., 2024).

Conflict of interest

The 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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2025.1484113/full#supplementary-material

References

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., and Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J. Bacteriol.* 186, 2629–2635. doi:10.1128/jb.186.9.2629-2635.2004

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21 (30), 1–16. doi:10.1186/s13059-020-1935-5

Amir, A., Zeisel, A., Zuk, O., Elgart, M., Stern, S., Shamir, O., et al. (2013). High-resolution microbial community reconstruction by integrating short reads from multiple 16S rRNA regions. *Nucleic acids Res.* 41, e205. doi:10.1093/nar/gkt1070

Andrews, S. (2010). Fastqc A quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Balle, C., Konstantinus, I. N., Jaumdally, S. Z., Havyarimana, E., Lennard, K., Esra, R., et al. (2020). Hormonal contraception alters vaginal microbiota and cytokines in South African adolescents in a randomized trial. *Nat. Commun.* 11, 5578. doi:10.1038/s41467-020-19382-9

Bassis, C. M., Moore, N. M., Lolans, K., Seekatz, A. M., Weinstein, R. A., Young, V. B., et al. (2017). Comparison of stool versus rectal swab samples and storage conditions on bacterial community profiles. *BMC Microbiol.* 17 (78), 1–7. doi:10.1186/s12866-017-0983-9

Bennato, F., Martino, C., Di Domenico, M., Ianni, A., Chai, B., Di Marcantonio, L., et al. (2022). Metagenomic characterization and volatile compounds determination in rumen from saanen goat kids fed olive leaves. *Veterinary Sci.* 9, 452. doi:10.3390/vetsci9090452

Bharti, R., and Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings Bioinforma.* 22, 178–193. doi:10.1093/bib/bbz155

Bokulich, N. A., Maldonado, J., Kang, D.-W., Krajmalnik-Brown, R., and Caporaso, J. G. (2019). Rapidly processed stool swabs approximate stool microbiota profiles. *Msphere* 4. doi:10.1128/msphere.00208-19

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi:10.1038/s41587-019-0209-9

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325–349. doi:10.2307/1942268

Brooks, J. P., Edwards, D. J., Harwich, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., et al. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* 15 (66), 1–14. doi:10.1186/s12866-015-0351-6

Bukin, Y. S., Galachyants, Y. P., Morozov, I., Bukin, S., Zakharenko, A., and Zemskaya, T. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci. Data* 6, 190007–190014. doi:10.1038/sdata.2019.7

Callahan, B. J., Grinevich, D., Thakur, S., Balamotis, M. A., and Yehezkel, T. B. (2021). Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome* 9 (130), 1–13. doi:10.1186/s40168-021-01072-3

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci.* 108, 4516–4522. doi:10.1073/pnas.1000080107

CDC (2015). *Guidelines for specimen collection: Instructions for collecting stool specimens*. Atlanta, GA: Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/specimen-collection.html (Accessed March, 2023).

Chai, B. (2021). 16S-SNAPP-py3. Available at: https://github.com/swiftbiosciences/16S-SNAPP-py3 (Accessed February, 2022).

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. methods* 69, 330–339. doi:10.1016/j.mimet.2007.02.005

Chanderraj, R., Brown, C. A., Hinkle, K., Falkowski, N., Woods, R. J., and Dickson, R. P. (2022). The bacterial density of clinical rectal swabs is highly variable, correlates with sequencing contamination, and predicts patient risk of extraintestinal infection. *Microbiome* 10, 2. doi:10.1186/s40168-021-01190-y

Chen, J., and Zhang, X. (2022). dICC: distance-based intraclass correlation coefficient for metagenomic reproducibility studies. *Bioinformatics* 38, 4969–4971. doi:10.1093/bioinformatics/btac618

Claassen-Weitz, S., Gardner-Lubbe, S., Mwaikono, K. S., Du Toit, E., Zar, H. J., and Nicol, M. P. (2020). Optimizing 16S rRNA gene profile analysis from low biomass nasopharyngeal and induced sputum specimens. *BMC Microbiol.* 20 (113), 1–26. doi:10.1186/s12866-020-01795-7

Claassen-Weitz, S., Gardner-Lubbe, S., Nicol, P., Botha, G., Mounaud, S., Shankar, J., et al. (2018). HIV-exposure, early life feeding practices and delivery mode impacts on faecal bacterial profiles in a South African birth cohort. *Sci. Rep.* 8 (5078), 1–15. doi:10.1038/s41598-018-22244-6

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6 (226), 1–14. doi:10.1186/s40168-018-0605-2

Drengenes, C., Eagan, T. M., Haaland, I., Wiker, H. G., and Nielsen, R. (2021). Exploring protocol bias in airway microbiome studies: one versus two PCR steps and 16S rRNA gene region V3 V4 versus V4. *BMC genomics* 22 (3), 1–15. doi:10.1186/s12864-020-07252-z

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecol.* 12, 42–58. doi:10.2307/1411

Flygel, T. T., Sovershaeva, E., Claassen-Weitz, S., Hjerde, E., Mwaikono, K. S., Odland, J. Ø., et al. (2020). Composition of gut microbiota of children and adolescents with perinatal human immunodeficiency virus infection taking antiretroviral therapy in Zimbabwe. *J. Infect. Dis.* 221, 483–492. doi:10.1093/infdis/jiz473

Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J., and Cotter, P. D. (2016). 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol.* 16 (123), 1–13. doi:10.1186/s12866-016-0738-z

Freedman, S. B., Xie, J., Nettel-Aguirre, A., Lee, B., Chui, L., Pang, X.-L., et al. (2017). Enteropathogen detection in children with diarrhoea, or vomiting, or both, comparing rectal flocked swabs with stool specimens: an outpatient cohort study. *lancet Gastroenterology and hepatology* 2, 662–669. doi:10.1016/s2468-1253(17)30160-7

Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., et al. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* 6 (17), 1–13. doi:10.1186/s40168-017-0396-x

Gao, X., Jia, R., Xie, L., Kuang, L., Feng, L., and Wan, C. (2018). A study of the correlation between obesity and intestinal flora in school-age children. *Sci. Rep.* 8, 14511. doi:10.1038/s41598-018-32730-6

Goldfarb, D. M., Steenhoff, A. P., Pernica, J. M., Chong, S., Luinstra, K., Mokomane, M., et al. (2014). Evaluation of anatomically designed flocked rectal swabs for molecular detection of enteric pathogens in children admitted to hospital with severe gastroenteritis in Botswana. *J. Clin. Microbiol.* 52, 3922–3927. doi:10.1128/jcm.01894-14

Gorzelak, M. A., Gill, S. K., Tasnim, N., Ahmadi-Vand, Z., Jay, M., and Gibson, D. L. (2015). Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PloS one* 10, e0134802. doi:10.1371/journal.pone.0134802

Graham, A. S., Patel, F., Little, F., van der Kouwe, A., Kaba, M., and Holmes, M. J. (2024). Using short-read 16S rRNA sequencing of multiple variable regions to generate high-quality results to a species level. bioRxiv.

Graspeuntner, S., Lupatsii, M., Dashdorj, L., Rody, A., Rupp, J., Bossung, V., et al. (2023). First-Day-of-Life rectal swabs fail to represent meconial microbiota composition and underestimate the presence of antibiotic resistance genes. *Microbiol. Spectr.* 11. doi:10.1128/spectrum.05254-22

Gratton, J., Phetcharaburanin, J., Mullish, B. H., Williams, H. R., Thursz, M., Nicholson, J. K., et al. (2016). Optimized sample handling strategy for metabolic profiling of human feces. *Anal. Chem.* 88, 4661–4668. doi:10.1021/acs.analchem.5b04159

Guo, F., Ju, F., Cai, L., and Zhang, T. (2013). Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment. *PloS one* 8, e76185. doi:10.1371/journal.pone.0076185

Hosgood III, H. D., Sapkota, A. R., Rothman, N., Rohan, T., Hu, W., Xu, J., et al. (2014). The potential role of lung microbiota in lung cancer attributed to household coal burning exposures. *Environ. Mol. Mutagen.* 55, 643–651. doi:10.1002/em.21878

Huson, D. H., Steel, M., El-Hadidi, M., Mitra, S., Peter, S., and Willmann, M. (2017). A simple statistical test of taxonomic or functional homogeneity using replicated microbiome sequencing samples. *J. Biotechnol.* 250, 45–50. doi:10.1016/j.jbiotec.2016.10.020

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: an overview. *Hum. Immunol.* 82, 801–811. doi:10.1016/j.humimm.2021.02.012

ILLUMINA. 16S Metagenomic Sequencing Library Preparation (2024). *Prep. 16S ribosomal RNA gene amplicons Illumina MiSeq Syst*. Available at: https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf (Accessed December, 2024).

Ji, B., and Nielsen, J. (2015). From next-generation sequencing to systematic modeling of the gut microbiome. *Front. Genet.* 6, 219. doi:10.3389/fgene.2015.00219

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* 10 (5029), 1–11. doi:10.1038/s41467-019-13036-1

Jones, R. B., Zhu, X., Moan, E., Murff, H. J., Ness, R. M., Seidner, D. L., et al. (2018). Inter-niche and inter-individual variation in gut microbial community assessment using stool, rectal swab, and mucosal samples. *Sci. Rep.* 8, 4139. doi:10.1038/s41598-018-22408-4

Kers, J. G., and Saccenti, E. (2021). The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report analysis the results. *Front. Microbiol.* 12:796025. doi:10.3389/fmicb.2021.796025

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids Res.* 41, e1. doi:10.1093/nar/gks808

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi:10.1016/j.jcm.2016.02.012

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci.* 82, 6955–6959. doi:10.1073/pnas.82.20.6955

Laursen, M. F., Dalgaard, M. D., and Bahl, M. I. (2017). Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. *Front. Microbiol.* 8, 1934. doi:10.3389/fmicb.2017.01934

Lee, C. K., Herbold, C. W., Polson, S. W., Wommack, K. E., Williamson, S. J., Mcdonald, I. R., et al. (2012). Groundtruthing next-gen sequencing for microbial ecology–biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One* 7, e44224. doi:10.1371/journal.pone.0044224

Liang, Y., Dong, T., Chen, M., He, L., Wang, T., Liu, X., et al. (2020). Systematic analysis of impact of sampling regions and storage methods on fecal gut microbiome and metabolome profiles. *Msphere* 5. doi:10.1128/msphere.00763-19

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi:10.1101/gr.097261.109

Ludwig, J. A., and Reynolds, J. F. (1988). *Statistical ecology: a primer in methods and computing*. John Wiley and Sons.

Maki, K. A., Wolff, B., Varuzza, L., Green, S. J., and Barb, J. J. (2023). Multi-amplicon microbiome data analysis pipelines for mixed orientation sequences using QIIME2: assessing reference database, variable region and pre-processing bias in classification of mock bacterial community samples. *Plos one* 18, e0280293. doi:10.1371/journal.pone.0280293

Ma, S. (2022). _MMUPHin: meta-analysis methods with uniform pipeline for heterogeneity in microbiome studies_. *Genome Biol.* 23 (208), 1–31. doi:10.18129/B9.bioc.MMUPHin

Ma, S., Shungin, D., Mallick, H., Schirmer, M., Nguyen, L. H., Kolde, R., et al. (2022). Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. *Genome Biol.* 23, 208–231. doi:10.1186/s13059-022-02753-4

Mcmurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one* 8, e61217. doi:10.1371/journal.pone.0061217

Mcmurdie, P. J., and Holmes, S. (2015). Shiny-phyloseq: web application for interactive microbiome analysis with provenance tracking. *Bioinformatics* 31, 282–283. doi:10.1093/bioinformatics/btu616

Midha, M. K., Wu, M., and Chiu, K.-P. (2019). Long-read sequencing in deciphering human genetics to a greater depth. *Hum. Genet.* 138, 1201–1215. doi:10.1007/s00439-019-02064-y

Nearing, J. T., Comeau, A. M., and Langille, M. G. (2021). Identifying biases and their potential solutions in human microbiome studies. *Microbiome* 9 (113), 1–22. doi:10.1186/s40168-021-01059-0

Nuccio, D. A., Normann, M. C., Zhou, H., Grippo, A. J., and Singh, P. (2023). Microbiome and metabolome variation as indicator of social stress in female prairie voles. *Int. J. Mol. Sci.* 24, 1677. doi:10.3390/ijms24021677

Özkurt, E., Fritscher, J., Soranzo, N., Ng, D. Y., Davey, R. P., Bahram, M., et al. (2022). LotuS2: an ultrafast and highly accurate tool for amplicon sequencing analysis. *Microbiome* 10 (176), 1–14. doi:10.1186/s40168-022-01365-1

Patel, J. B. (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Mol. Diagn.* 6, 313–321. doi:10.2165/00066982-200106040-00012

Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., Van Doorn, L.-J., et al. (2019). Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front. Microbiol.* 10, 1277. doi:10.3389/fmicb.2019.01277

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* 13 (341), 1–13. doi:10.1186/1471-2164-13-341

Radhakrishnan, S. T., Gallagher, K. I., Mullish, B. H., Serrano-Contreras, J. I., Alexander, J. L., Miguens Blanco, J., et al. (2023). Rectal swabs as a viable alternative to faecal sampling for the analysis of gut microbiota functionality and composition. *Sci. Rep.* 13, 493. doi:10.1038/s41598-022-27131-9

R CORE TEAM (2022). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reyman, M., Van Houten, M. A., Arp, K., Sanders, E. A., and Bogaert, D. (2019). Rectal swabs are a reliable proxy for faecal samples in infant gut microbiota research based on 16S-rRNA sequencing. *Sci. Rep.* 9, 16072. doi:10.1038/s41598-019-52549-z

Rogers, G. B., and Bruce, K. D. (2010). Next-generation sequencing in the analysis of human microbiota: essential considerations for clinical application. *Mol. diagnosis and Ther.* 14, 343–350. doi:10.1007/bf03256391

Schriefer, A. E., Cliften, P. F., Hibberd, M. C., Sawyer, C., Brown-Kennerly, V., Burcea, L., et al. (2018). A multi-amplicon 16S rRNA sequencing and analysis method for improved taxonomic profiling of bacterial communities. *J. Microbiol. methods* 154, 6–13. doi:10.1016/j.mimet.2018.09.019

Schroeder, P. J., and Jenkins, D. G. (2018). How robust are popular beta diversity indices to sampling error? *Ecosphere* 9, e02100. doi:10.1002/ecs2.2100

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., VON Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. methods* 15, 461–468. doi:10.1038/s41592-018-0001-7

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Short, M. I., Hudson, R., Besasie, B. D., Reveles, K. R., Shah, D. P., Nicholson, S., et al. (2021). Comparison of rectal swab, glove tip, and participant-collected stool techniques for gut microbiome sampling. *BMC Microbiol.* 21 (26), 1–9. doi:10.1186/s12866-020-02080-3

Simpson, E. H. (1949). Measurement of diversity. *nature* 163, 688. doi:10.1038/163688a0

Sun, S., Zhu, X., Huang, X., Murff, H. J., Ness, R. M., Seidner, D. L., et al. (2021). On the robustness of inference of association with the gut microbiota in stool, rectal swab and mucosal tissue samples. *Sci. Rep.* 11, 14828. doi:10.1038/s41598-021-94205-5

Szoboszlay, M., Schramm, L., Pinzauti, D., Scerri, J., Sandionigi, A., and Biazzo, M. (2023). Nanopore is preferable over Illumina for 16S amplicon sequencing of the gut microbiota when species-level taxonomic classification, accurate estimation of richness, or focus on rare taxa is required. *Microorganisms* 11, 804. doi:10.3390/microorganisms11030804

THE JACKSON LABORATORY. (2019). *Analysing 16S data: Part 2* Available at: https://thejacksonlaboratory.github.io/microbiome-workshop-2019/jekyll/update/2019/10/30/Analysing-16S-data-part-2.html.

Tucker, T., Marra, M., and Friedman, J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* 85, 142–154. doi:10.1016/j.ajhg.2009.06.022

Urban, L., Holzer, A., Baronas, J. J., Hall, M. B., Braeuninger-Weimer, P., Scherm, M. J., et al. (2021). Freshwater monitoring by nanopore sequencing. *Elife* 10, e61504. doi:10.7554/elife.61504

Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.* 34, 666–681. doi:10.1016/j.tig.2018.05.008

Videnska, P., Smerkova, K., Zwinsova, B., Popovici, V., Micenkova, L., Sedlar, K., et al. (2019). Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform. *Sci. Rep.* 9, 13837. doi:10.1038/s41598-019-49520-3

Wang, S., Sun, B., Tu, J., and Lu, Z. (2016). Improving the microbial community reconstruction at the genus level by multiple 16S rRNA regions. *J. Theor. Biol.* 398, 1–8. doi:10.1016/j.jtbi.2016.03.016

Wang, Y., and Qian, P.-Y. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PloS one* 4, e7401. doi:10.1371/journal.pone.0007401

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., and Forney, L. J. (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PloS one* 7, e33865. doi:10.1371/journal.pone.0033865

Yu, G., Phillips, S., Gail, M. H., Goedert, J. J., Humphrys, M. S., Ravel, J., et al. (2017). The effect of cigarette smoking on the oral and nasal microbiota. *Microbiome* 5 (3), 1–16. doi:10.1186/s40168-016-0226-6

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi:10.1101/gr.074492.107

Zhou, H.-W., Li, D.-F., Tam, N. F.-Y., Jiang, X.-T., Zhang, H., Sheng, H.-F., et al. (2011). BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME J.* 5, 741–749. doi:10.1038/ismej.2010.160

Zreloff, Z. J., Lange, D., Vernon, S. D., Carlin, M. R., and Cano, R. J. (2023). Accelerating gut microbiome research with robust sample collection. *Res. and Rev. J. Microbiol. Biotechnol.* 12, 33–47.

ZYMOBIOMICS (2022). ZymoBIOMICS™ microbial community DNA standard; catalog nos. D6305 (200ng) and D6306 (2000ng). Available at: https://files.zymoresearch.com/protocols/_d6305_d6306_zymobiomics_microbial_community_dna_standard.pdf (Accessed October, 2022).

ZYMOBIOMICS (2024). ZymoBIOMICS™ DNA miniprep kit. Available at: https://files.zymoresearch.com/protocols/_d4300t_d4300_d4304_zymobiomics_dna_miniprep_kit.pdf (Accessed December, 2024).