AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Biotechnology | Full-Length Text

The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling

Tianyuan Zhang,[1,2,3] Hanzhou Li,[3] Silin Ma,[1] Jian Cao,[3] Hao Liao,[1] Qiaoyun Huang,[1,4] Wenli Chen[1,2]

AUTHOR AFFILIATIONS See affiliation list on p. 17.

ABSTRACT The long-read amplicon provides a species-level solution for the community. With the improvement of nanopore flowcells, the accuracy of Oxford Nanopore Technologies (ONT) R10.4.1 has been substantially enhanced, with an average of approximately 99%. To evaluate its effectiveness on amplicons, three types of microbiomes were analyzed by 16S ribosomal RNA (hereinafter referred to as "16S") amplicon sequencing using Novaseq, Pacbio sequel II, and Nanopore PromethION platforms (R9.4.1 and R10.4.1) in the current study. We showed the error rate, recall, precision, and bias index in the mock sample. The error rate of ONT R10.4.1 was greatly reduced, with a better recall in the case of the synthetic community. Meanwhile, in different types of environmental samples, ONT R10.4.1 analysis resulted in a composition similar to Pacbio data. We found that classification tools and databases influence ONT data. Based on these results, we conclude that the ONT R10.4.1 16S amplicon can also be used for application in environmental samples.

IMPORTANCE The long-read amplicon supplies the community with a species-level solution. Due to the high error rate of nanopore sequencing early on, it has not been frequently used in 16S studies. Oxford Nanopore Technologies (ONT) introduced the R10.4.1 flowcell with Q20+ reagent to achieve more than 99% accuracy as sequencing technology advanced. However, there has been no published study on the performance of commercial PromethION sequencers with R10.4.1 flowcells on 16S sequencing or on the impact of accuracy improvement on taxonomy (R9.4.1 to R10.4.1) using 16S ONT data. In this study, three types of microbiomes were investigated by 16S ribosomal RNA (rRNA) amplicon sequencing using Novaseq, Pacbio sequel II, and Nanopore PromethION platforms (R9.4.1 and R10.4.1). In the mock sample, we displayed the error rate, recall, precision, and bias index. We observed that the error rate in ONT R10.4.1 is significantly lower, especially when deletions are involved. First and foremost, R10.4.1 and Pacific Bioscience platforms reveal a similar microbiome in environmental samples. This study shows that the R10.4.1 full-length 16S rRNA sequences allow for species identification of environmental microbiota.

KEYWORDS nanopore sequencing, mock communities, full-length 16S rRNA sequencing, ONT R10.4.1, environmental sample

Microorganisms are closely related to environmental ecology and biological health. DNA sequencing is the core technology of the life sciences and biological sciences. Current innovations in DNA sequencing technology have had a progressive influence on clinical microbiology. Among sequence-established investigations, amplicon sequencing of the 16S ribosomal RNA (rRNA) gene has developed to be a trustworthy and effective option for the taxonomic classifications of the microbial communities and

Editor Nicole R. Buan, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

Address correspondence to Wenli Chen, wlchen@mail.hzau.edu.cn.

The authors declare no conflict of interest.

See the funding table on p. 17.

Received 11 April 2023
Accepted 4 August 2023
Published 6 October 2023

their respective characteristics (1). The second-generation sequencing platforms typically provide a large number of short reads (PE250–PE300) and high accuracy (>99%) for 16S studies (2–4). The bacterial 16S rRNA gene comprises nine variable regions "V1 to V9" divided by highly preserved sequences within distinct taxa. For bacterial detection, the first step is the amplification of the 16S rRNA gene via PCR by using primers annealing to conserved regions, followed by sequencing (5). The gathered data after sequencing are processed via various bioinformatic analyses, where varying regions are utilized to differentiate among bacterial taxa. For the short-read sequencers, the V3-V4 and V4-V5 variable regions showed the highest classification accuracy in bacterial diversity studies (6). Thus, primers in the V3-V4 region showed good coverage and specificity (7). Nowadays, short-read 16S sequencing platforms, including Illumina Miseq PE300, Illumina Hiseq PE250, Illumina-Novaseq, and BGI-seq PE250, dominate the commercial markets. However, short reads cannot cover the full-length 16S sequences, thus limiting the resolution for species-level identification (8). The progress in throughput has, nevertheless, been at the expense of read length, and this balance has certainly caused less accurate categorization of limited 16S sequences, especially at the genus or species level. Recent developments in long-read sequencing techniques, for example, the Pacific Biosciences (PacBio, PB) platform based on the circular consensus sequencing (CCS) model, can completely cover the full-length 16S sequences to achieve high resolution for species-level identification with an average length of 1.5 kb and high accuracy (>99.9%) (5, 9). As a consequence, the PB platform is increasingly being used for 16S studies.

Although available since ≈2014, Oxford Nanopore Technologies (ONT) produces more than 2 Mb of long reads, which shows promise in microbial genomics research (10). ONT offers two benchtop sequencers (Flongle and MinION) and two commercial sequencers (GridION and PromethION). Due to the high error rate (5%–38.5%) (11), which is higher than most taxonomist's requirements for percentage identity scores of ≥97% (genus levels) and ≥99% (species levels), ONT sequencers have not been widely used. Currently, most studies using ONT sequencing are based on MiniON sequencers and R9.4 (R9.4.1) flow cells (12–15). R9.4 can achieve an accuracy of 85%–94% for ONT reads, which limits this technology for 16S studies. Due to its portability, ONT 16S has been applied in environmental investigations (16) and clinical diagnosis (17–19). As flowcells and sequencing reagents are being improved, the latest R10.4.1 flowcell can generate whole genome sequencing data with a model read accuracy of ~99% using Q20+ chemistry and the SQK-LSK114 kit. The accuracy can be even further improved with homopolymers (20, 21).

Given the low species-level accuracy caused by sequencing errors, the classic DADA2 software is unsuitable for ONT 16S data. To improve the accuracy of ONT data, researchers have explored many bioinformatics tools for nanopore 16S sequencing, such as NanoClust (22), Emu (23), and BugSeq 16S (24). Besides, they developed experimental technologies to improve 16S sequence accuracy, such as INC-seq (25) and unique molecular identifier (UMI) technology (26). To date, many commercial mock communities have been used to assess the accuracy of different sequencing platforms and different bioinformatics pipelines. However, there is neither a published study on the performance of commercial PromethION sequencers with R10.4.1 flowcells nor the influence of accuracy improvement on taxonomic classification (R9.4.1 to R10.4.1) by 16S ONT data.

This study aimed to assess the performance of the ONT R10.4.1 flowcell in community profiles. Four types of the microbiome (standard mock community, synthetic agricultural community, soils, and water samples) were analyzed by 16S rRNA amplicon sequencing using Illumina Novaseq PE250, Pacbio Sequel II, and Nanopore PromethION platforms (R9.4.1 and R10.4.1). We focused on the comparison of taxonomic assignments and accuracy evaluation metrics among these platforms.

RESULTS

Experiment design and workflows

In the current research work, using a custom workflow, Novaseq, Pacbio, ONT R9.4.1, and ONT R10.4.1 nanopore full-length (V1-V9) 16S ribosomal RNA-sequencing were performed on the Zymo and Synthesis communities (Fig. 1A). To evaluate the taxonomic assignments in R10.4.1, LAST and Emu for long-reads and DADA2 for Novaseq classification tools were compared using the NCBI and SILVA databases. For mock samples, we focused on error profiles, the proportion of correctly classified reads, L1 distance, true positive (TP) taxa, false positive (FP) taxa, precision, recall, and F1 scores.

For the environmental samples, we focused on the performances of R10.4.1 at the species level and the comparison of ONT and PB, as well as correlations and differences from different diversity indices. We want to know what advantages R10.4.1 has over R9. Then can R10.4.1 reach a reliable taxonomic assignment at the species level compared to PB? Globally, we showed that Oxford Nanopore R10.4.1 is capable and can be used for profiling the microbiota.

The quantification of genomic DNA and reconstruction of synthetic community

Currently, 12 species have been selected to construct a synthetic agriculture community. Considering the variation in copy numbers of the 16S rRNA gene across different bacterial genomes, we employed quantitative real-time PCR (qRT-PCR) for its quantification to obtain the abundance of copy numbers. The genomic DNA of the strains was measured in triplicate by using an SYRB Green qPCR, and the efficiency was found to be

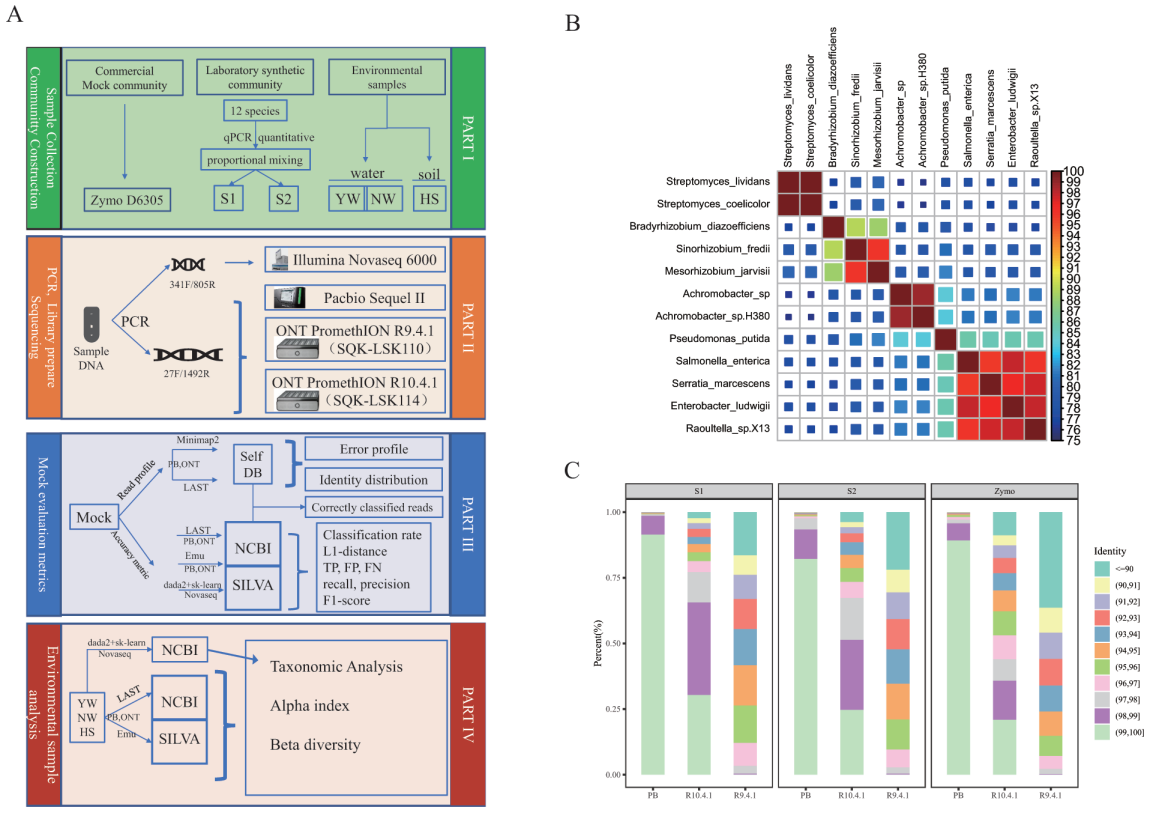

FIG 1 Project design, strain 16S identity, and alignment identity. (A) The current workflow for the evaluation of R10.4.1 full-length 16S sequences. (B) The 16S identity among the 12 species of synthetic community. Blue represents the lowest similarity, whereas red represents the highest similarity. The square size is proportional to the identity. (C) The identity distribution among sequences that were obtained from different sequencing platforms as compared to the reference.

97.57%. Each gDNA was quantified using a Qubit fluorometer. The average copy number of *Serratia marcescens*, *Enterobacter ludwigii*, *Achromobacter* sp., *Streptomyces coelicolor*, *Streptomyces lividans*, *Sinorhizobium fredii*, *Bradyrhizobium diazoefficiens*, *Mesorhizobium jarvisii*, *Pseudomonas putida*, *Achromobacter* sp. H380, *Salmonella enterica*, and *Raoultella* sp. X13 are 9.07×10^7, 9.74×10^7, 2.77×10^7, 1.41×10^7, 1.48×10^7, 2.14×10^7, 5.45×10^6, 1.30×10^7, 7.29×10^7, 1.63×10^7, 6.79×10^7, and 9.61×10^7 per ng, respectively (Table S1).

The ratio of each extracted genomic DNA in a synthetic community was determined by different proportions of mass to obtain the copy number of 16S rRNA genes (Table S1). The largest DNA value in the S1 sample was 1,000 times the smallest value, and the mass value was dispersed in a gradient. The S1 sample consists of *S. marcescens* (39.77%), *E. ludwigii* (38.44%), *Achromobacter* sp. (9.716%), *S. coelicolor* (2.467%), *S. lividans* (1.946%), *S. fredii* (1.880%), *B. diazoefficiens* (0.2391%), *M. jarvisii* (0.05683%), *P. putida* (1.279%), *Achromobacter* sp. H380 (3.574%), *S. enterica* (0.5955%), and *Raoultella* sp. X13 (0.04217%), with a 943-fold difference in 16S copy number. In the S2 sample, a 20-fold difference in genomic DNA was maintained, including *S. marcescens* (1.564%), *E. ludwigii* (1.679%), *Achromobacter* sp. (0.4776%), *S. coelicolor* (0.2426%), *S. lividans* (0.2551%), *S. fredii* (7.395%), *B. diazoefficiens* (1.881%), *M. jarvisii* (4.470%), *P. putida* (25.16%), *Achromobacter* sp. H380 (0.2811%), *S. enterica* (23.42%), and *Raoultella* sp. X13 (33.17%), with a 136-fold difference in 16S copy number. Some species have a higher similarity in 16S sequences with the same genus, such as *S. lividans* and *S. coelicolor*, which share 99.93% identity, and *Achromobacter* sp. and *Achromobacter* sp. H380, which share 98.92% similarity (Fig. 1B).

Statistics of reads generated via different platforms

The Novaseq platform generated an average of 46,932 (SD ±5,879) high-quality pair-end reads per sample, with Q20 >96% and Q30 >90% for all the samples, respectively (Table S2). For the PB platform, 32,858 (mean) ± 2,543 (SD) clean CCS tags for environmental samples and 13,575 (mean) ± 743 (SD) for mock samples were generated (Table S3).

The ONT R9.4.1 platform displayed 42,840 (mean) ± 1,066 (SD) clean reads (Q score ≥7) per sample, with an average Q score of 12.7 (Table S4). The ONT R10.4.1 platform obtained 57,252 (mean) ± 1,309 (SD) clean reads (Q score ≥10) per sample. The average Q score for raw data for the ONT R10.4.1 platform was "18.77" (raw accuracy: 98.67%), while for data without primer, it was "22.41" (accuracy: 99.42%) (Table 1).

Characterization of the mock community using individual sequencing platforms

To evaluate the accuracy of the R10.4.1 flowcell on the amplicon after improvements, error profiling was performed on the long-read sequencing platforms (Table 2). We calculated the mismatch error rates and indel error rates of samples, meanwhile obtaining a mean error rate among all the samples.

Error profiling

CCS reads exhibited a mean error rate of 0.4658%, including a 0.2656% mismatch error rate, a 0.136% insertion error rate, and a 0.0636% deletion error rate, respectively. R9.4.1 reads have the highest error rate of 7.16% with 94.62% accuracy compared to raw reads. Compared to R9.4.1, the total error rate of R10.4.1 reads was only 3.16%. Moreover, R10.4.1 had a significantly lower error rate, with the mismatch error rate reduced to less than half and the insertion error rate reduced to one-third. In the case of deletion error rates, R10.4.1 performed better than R9.4.1. A deletion error rate of 0.6188% was recorded in the case of R10.4.1, whereas the deletion error rate for R9.4.1 was recorded at 2.6635%. As the synthetic community has greater sequence diversity, the error rates of ONT reads in the synthetic community are lower than those in Zymo. To demonstrate the existence of errors, we chose sequences of *B. diazoefficiens* from different platforms with a single copy of the 16S gene. The Integrative Genomics Viewer (IGV) results showed that

TABLE 1 Statistics of reads generated via the ONT R10.4.1 platform[a]

Sample	Raw reads	Raw_meanQ	Clean reads	Mean_len (bp)	Mean_Q
S1-1	65,289	18.93	59,206	1,458	22.38
S1-2	64,886	18.87	57,631	1,458	22.39
S1-3	65,326	18.89	58,209	1,458	22.42
S2-1	64,953	18.88	57,948	1,457	22.52
S2-2	65,236	18.86	57,412	1,455	22.47
S2-3	65,002	18.85	58,994	1,455	22.42
Zymo-1	64,950	19.03	58,991	1,470	22.67
Zymo-2	64,969	18.92	57,770	1,468	22.61
Zymo-3	64,700	18.92	57,327	1,470	22.66
YW-1	64,843	18.58	55,588	1,440	22.36
YW-2	65,004	18.56	56,032	1,443	22.33
YW-3	64,914	18.5	54,658	1,442	22.3
NW-1	65,159	18.71	55,660	1,433	22.37
NW-2	65,090	18.54	56,214	1,433	22.43
NW-3	65,150	18.72	56,055	1,433	22.37
HS-1	65,420	18.73	57,610	1,450	22.25
HS-2	65,175	18.71	58,783	1,450	22.22
HS-3	64,566	18.69	56,464	1,450	22.25

[a]Clean reads remove the barcode, adapter, and primer.

R10.4.1 performed well and had lower error rates as compared to other platforms (Fig. S1).

Alignment of obtained reads against the reference

The similarity of 16S impacts the accuracy of species identification. At least 82.18% of the reads on the PB platform had a similarity of more than 99% (sample S1, Fig. 1C). However, the similarity of reads obtained via R9.4.1 is mostly between 92% and 95%. Moreover, for R10.4.1, more than 97% of identity was accounted for by Zymo (43.89%), S1 (77.10%), and S2 (67.26%), respectively.

Evaluation of sequencers, approaches, and databases for classification

To quantify the results obtained from R10.4.1 in comparison to other platforms, the following matrices were used to evaluate the performance at both the genus and species levels. For the Zymo sample, PB- and ONT-based mapping methods accurately detected the eight TP species; however, Novaseq 16S was not suitable for species-level detection, which only detected three TP species (Fig. S2; Table S5). For PB data, LAST_Silva was used to obtain the least FPs with the lowest L1 distance, while Emu_NCBI possessed

TABLE 2 Mean error rates of raw reads generated via different platforms among all the samples

Platform	Sample	Mismatch error rate	Insertion error rate	Deletion error rate	All error rate
PB	S1	0.2290%	0.1216%	0.0525%	0.4031%
PB	S2	0.2907%	0.1726%	0.0809%	0.5443%
PB	Zymo	0.2783%	0.1149%	0.0573%	0.4505%
PB	All	0.2656%	0.1366%	0.0636%	0.4658%
R9.4.1	S1	3.1497%	1.4629%	2.5933%	7.2059%
R9.4.1	S2	3.4947%	1.5117%	2.6360%	7.6424%
R9.4.1	Zymo	4.4546%	1.6059%	2.7653%	8.8259%
R9.4.1	All	3.6390%	1.52573%	2.6635%	7.8789%
R10.4.1	S1	1.5049%	0.3828%	0.5542%	2.4419%
R10.4.1	S2	1.8361%	0.4252%	0.6005%	2.8618%
R10.4.1	Zymo	2.9071%	0.5746%	0.7022%	4.1838%
R10.4.1	All	2.0815%	0.4607%	0.6188%	3.1610%

the second lowest FPs. For ONT data, R10.4.1 with "Emu_Silva" produced the lowest FPs. R9.4.1 was slightly better(Lowest FPs) than R10.4.1 when using the NCBI database across different methods. Using LAST, the R10.4.1 platform gets a smaller L1 distance than the R9.4.1 platform across different databases. The comparison of R10.4.1's FP is as follows: Emu_SILVA < LAST_NCBI < Emu_NCBI < LAST_SILVA. Overall, PB outperformed ONT for the Zymo sample.

The S1 sample was an intentionally challenging community containing several microbes of the same genus but different species with similar 16S sequences (Fig. 2; Table S6). At the species level, using the LAST_NCBI method, the ONT data obtained the complete recall (12/12). However, one replicate of PB data detected 11 TPs (not detected *Raoultella* sp. X13, with an abundance of 0.0004). Novseq data cannot achieve species-level identification. Software, and databases contribute to the major differences in community abundance using long-reads platforms. When compared to LAST, the Emu method gets less "Other" results. Also, the number of Other found by the NCBI database is less than what the SILVA database reports. The R10.4.1 reads produced the lowest L1 distances (Emu_NCBI: 0.348), followed by R9.4.1 (Emu_NCBI: 0.364). Notably, PB produced the largest L1 distance (Emu_NCBI: 0.435). But the Emu method yields fewer FPs as well as lesser TPs (Table S6; Emu_silva for PB: lost 5 TPs; Emu_Silva for R10.4.1 and R9.4.1: lost 3 TPs; Emu_Self cannot get the full TPs). In addition, among the methods that achieve complete recall, R10.4.1 obtains the lowest L1 distance (0.525). At the species level, using the NCBI database results in a smaller L1 distance than using the Silva database. However, at the genus level, both databases exhibit similar L1 distances. When using the NCBI database, Emu achieves a smaller L1 distance at the species level, while LAST exhibits the smallest L1 distance at the genus level.

Among the expected species in sample S2, all of the long-read platforms obtained complete recall at both the genus and species levels by using LAST_NCBI. Using the above methods, PB performs better (L1 distance: 0.198 at the genus level and 0.211 at the species level) than ONT R10.4.1 (0.552 at the genus level and 0.624 at the species level). The L1 distance obtained from the NCBI database is smaller than that from the Silva database. When using the Emu method, none of the long-read platforms can achieve complete recall. In addition, PB gets far fewer FPs than ONT data (Table S7). The

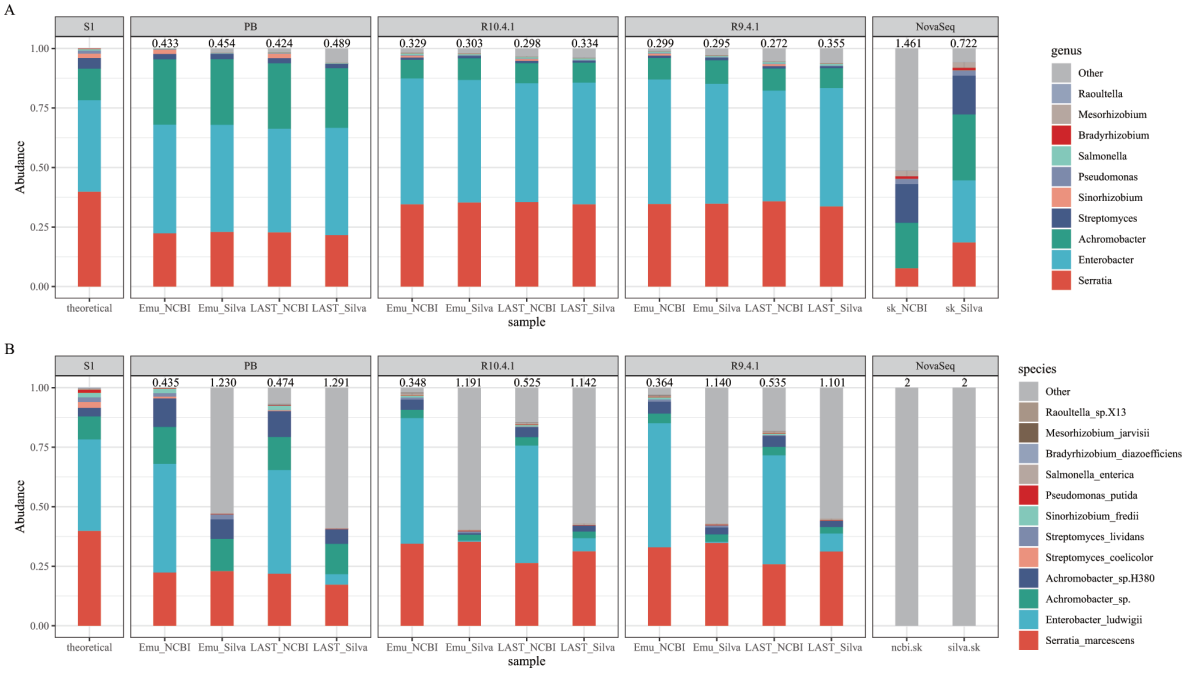

FIG 2 Classification results on genus (A) and species (B) levels for the S1 sample. Theoretical abundance and validation are given for comparison. The L1 distance is marked above the stack. Other includes unclassified.

L1 distances of the three long-read methods analyzed using LAST_NCBI are as follows: PB < R10.4.1 < R9.4.1 (Fig. S3; Table S7). At both the species and genus levels, using the NCBI database results in a smaller L1 distance than using the Silva database. Similar to sample S1, the Emu method lost TPs more than LAST.

Correctly assigned rates for obtained reads

Correctly assigned rates from the NCBI database were higher than those from the SILVA database in the synthetic communities, but in Zymo, an opposite trend was observed (Fig. 3A and B). While using the LAST_NCBI method, the correctly assigned rates increase with increased accuracy. Interestingly, while using Emu_NCBI at species level, R9.4.1 performs better than R10.4.1 (Fig. 3B).

Classification of variations by software and databases

Using the NCBI and SILVA databases, we aligned the generated sequencing data using the Emu and LAST aligners (Fig. 3C and D). Emu successfully assigns all generated data to species and genus levels and determines the target species based on the best hit according to the algorithm. However, while using LAST, the sequences with identity less than 90% utilized the lowest common ancestor (LCA) method to determine species, and the assigned ratio of the NCBI is greater than that of the SILVA database, which grows as the accuracy improves. Particularly for synthetic communities, the assigned rate of PB at the species level was greatly decreased. These results indicate that the ability to classify sequences to the genus or species level depends on sample type, method, and database.

Comparison of complex microbial community analyses and taxonomic classifications among different sequencing platforms

Mock communities can be useful for evaluating the performance of different sequencing platforms, but they lack richness. Therefore, we analyzed the microbial compositions produced by different platforms using real samples (water and soil). All samples were rarified based on the minimal number of reads (26,000 reads). As with mock, Emu successfully assigned all reads to species and genus levels. Although the rate of assigned reads by Emu was greater than that of LAST, the number of annotated species and genus

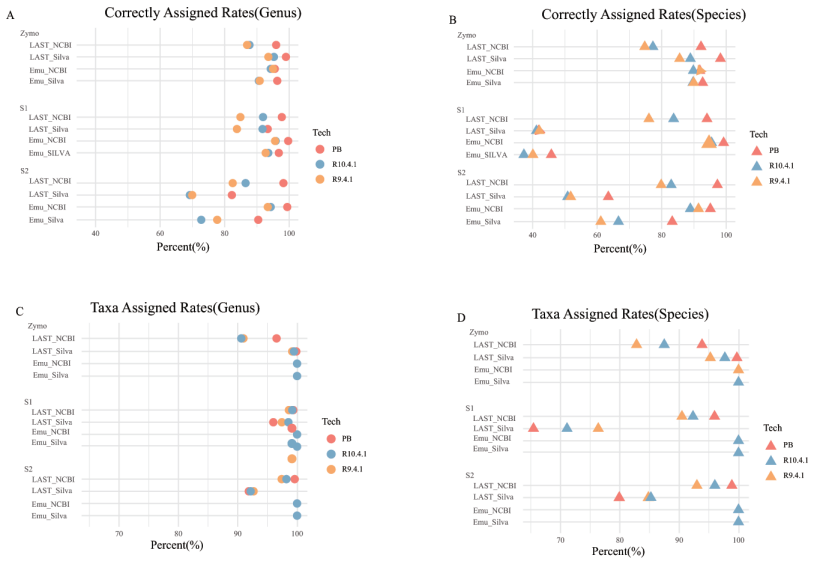

FIG 3 Taxonomic profiling results and statistics of correctly classified reads using the NCBI and SILVA databases. Generated sequencing data were aligned using the Emu and LAST aligners. Percentage of reads that could be correctly classified into genus (A) and species (B). Percentage of reads that could be classified into genus (C) and species (D). The data were curated from the SILVA and NCBI databases across different platforms and aligners.

was fewer. Furthermore, the number of species and genus identified in ONT R10.4.1 was higher than those identified via PB (Table 3). Next, we used the LAST aligner for further comparisons. For YW samples, 699 genera and 1,052 species were identified by PB, while 946 genera and 1,562 species were identified by R10.4.1, and 973 species and 816 genera were identified by R9.4.1. The results of read classifications were in the following order: PB > R10.4.1 > R9.4.1 for species and R10.4.1 > PB > R9.4.1 for the genus.

For NW samples, PB identified 281 genera and 405 species, while for ONT, R10.4.1 identified 566 genera and 836 species, and R9.4.1 identified the least number of species and genus. The number of species identified by R10.4.1 was much higher than that of R9.4.1. Regardless of species or genus, the rate of PB read classifications was greater than that of R10.4.1. For HS samples, R10.4.1 reads identified more species and genus than PB, although the ratio of read classifications was lower than PB. In short, the rate of R10.4.1 read classifications and the number of species and genus identified were better than those identified by R9.4.1 (Fig. S4). However, R10.4.1 obtained more species than PB.

Evaluation of ONT platform to profile the complex microbial community microbiota at species and genus levels

For overall species composition results, we first compared the correlations between LAST and Emu aligners. The correlations between the two aligners were poor (Table S8). We analyzed the average abundance of the top 15 species across different platforms, aggregated them, and then calculated the correlations.

For YW samples, the average abundance of the top 15 species shared by PB and R10.4.1 was higher than that of PB and R9.4.1 at both species and genus levels (Fig. S5). Comparing the number of top 15 common species in PB and R10.4.1, the LAST results were more accurate than Emu, and the correlation between LAST and Emu was not significant. Next, we focused on the results of LAST aligners. When the LAST aligner was used, the R10.4.1 and PB had a Pearson correlation coefficient (PCC) of 0.51 ($P = 0.013$) for species level and 0.83 ($P = 3.1e^{-6}$) for genus level. Moreover, for NW samples, the average abundance of the common top 15 species shared by PB and R10.4.1 was no less than that of PB and R9.4.1 at both species and genus levels (Fig. S6). PB and R10.4.1 have a PCC of 0.796 ($P = 9.2e^{-6}$) at species levels and 0.811 ($P = 1.4e^{-5}$) at genus levels, respectively.

TABLE 3 Statistics of the number of taxa at different phylogenetic levels identified via different platforms

Sample	Method	Platforms	Phylum	Class	Order	Family	Genus	Species
HS	LAST	PB	20	51	116	229	549	1,088
HS	LAST	R10.4.1	19	51	122	243	702	1,436
HS	LAST	R9.4.1	16	51	118	238	597	870
HS	Emu	PB	20	53	101	168	324	495
HS	Emu	R10.4.1	15	43	88	158	374	669
HS	Emu	R9.4.1	15	43	85	146	330	571
HS	dada2_sk	Novaseq	16	42	91	163	259	199
NW	LAST	PB	16	35	68	123	281	405
NW	LAST	R10.4.1	18	44	103	216	566	836
NW	LAST	R9.4.1	17	38	92	199	496	544
NW	Emu	PB	10	19	37	66	146	219
NW	Emu	R10.4.1	16	31	62	109	281	489
NW	Emu	R9.4.1	13	25	55	102	270	472
NW	dada2_sk	Novaseq	18	38	72	109	161	127
YW	LAST	PB	27	63	136	280	699	1,052
YW	LAST	R10.4.1	29	67	152	326	946	1,562
YW	LAST	R9.4.1	26	63	146	310	816	973
YW	Emu	PB	27	55	111	202	412	583
YW	Emu	R10.4.1	24	55	113	207	526	897
YW	Emu	R9.4.1	24	53	106	196	487	793
NW	dada2_sk	Novaseq	18	38	72	109	161	127

Furthermore, for HS samples, there were 10 common species and 9 genus detected by R10.4.1 and PB, but only 6 common species and 8 genus were detected by R9.4.1 and PB. PCC at the species level was 0.99 (P = 1.3e^{-19}) and 0.985 (P = 6.9e^{-17}) at the genus level (Fig. 4). These results suggest that the major species of R10.4.1 and PB have a similar distribution, especially in HS samples.

Boxplots of alpha diversity across three platforms are shown in Fig. 5. R10.4.1 can obtain a higher PB for Chao1 and feature. In addition, PB and R10.4.1 lie closer on the Shannon and Simpson indices, and their significance is lower than PB and R9.4.1 (Fig. 5A through C). Especially, PB and R10.4.1 do not show statistical significance (Shannon: P = 0.76; Simpon: P = 0.076) for YW samples, but there were significant differences among PB and R10.4.1.

Rarefaction analysis of the obtained reads

The results of rarefaction curves showed that the Emu aligner was easier to saturate because of the identification of fewer species (Fig. 5D). By using LAST aligners, R10.4.1 is often harder to saturate. For NW samples, species saturation can be reached at 26,000 reads, but for HS and YW samples, more reads (~40,000) are required.

Principal Coordinate Analysis(PCoA) and permutational multivariate analysis of variance(PERMANOVA) analysis for environmental samples

Finally, we compared different platforms based on the Bray–Curtis matrix. Current results suggest that species-level community composition was significantly diverse across different platforms (HS: F = 13.32, P = 0.001; NW: F = 9.80, P = 0.001; YW: F = 4.93, P = 0.007) and aligners (HS: F = 70.35, P = 0.001; NW: F = 31.16, P = 0.001; YW: F = 43.72, P = 0.001) based on a permutational multivariate analysis of variance (PERMANOVA) and visualized by Principal Coordinates Analysis (PCoA) for the unfiltered data (Fig. 6A through C). The aligners separated along the x-axis (HS: 72.97%; NW: 55.13%; YW: 66.65%), and the platforms separated along the y-axis (HS: 16.53%; NW: 24.7%; YW: 15.44%). The effect of aligners on species composition is remarkable.

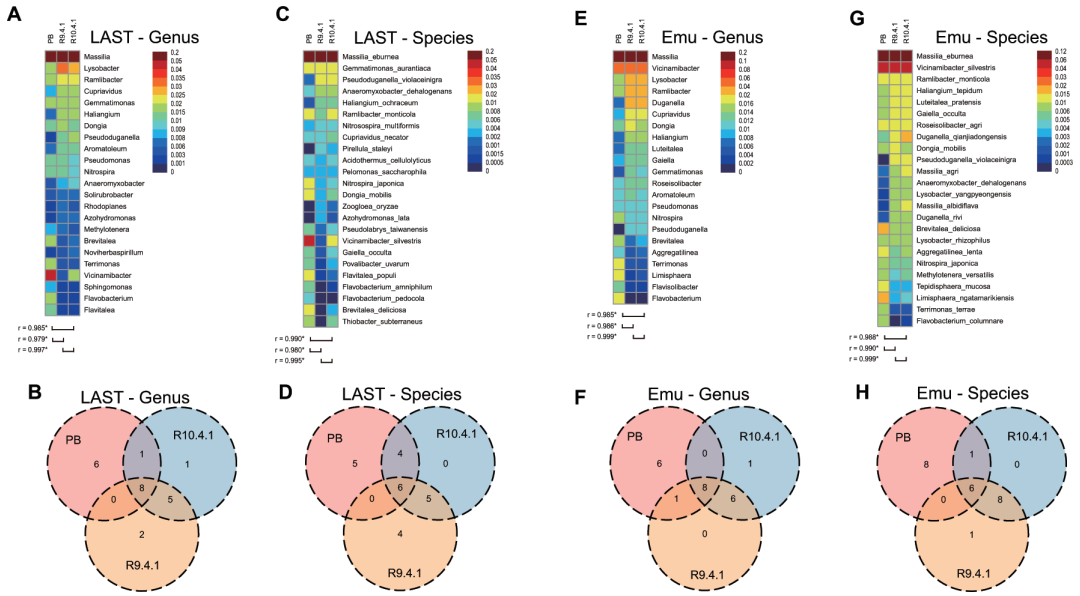

FIG 4 Distribution of the major taxa determined using PB, R9.4.1, and R10.4.1 platforms. Only the taxa with the highest relative abundance in at least one platform in the HS samples were included in the heat map (A, C, E, G) and Venn diagram (B, D, F, H). In the heat map, the scale colors indicate the relative abundances. Cells with asterisks indicate a significant (P < 0.05) difference in means compared to PB. (A and B) represent the LAST genus; (C and D) represent the LAST species; (E and F) represent the Emu genus; and (G and H) represent the Emu species.

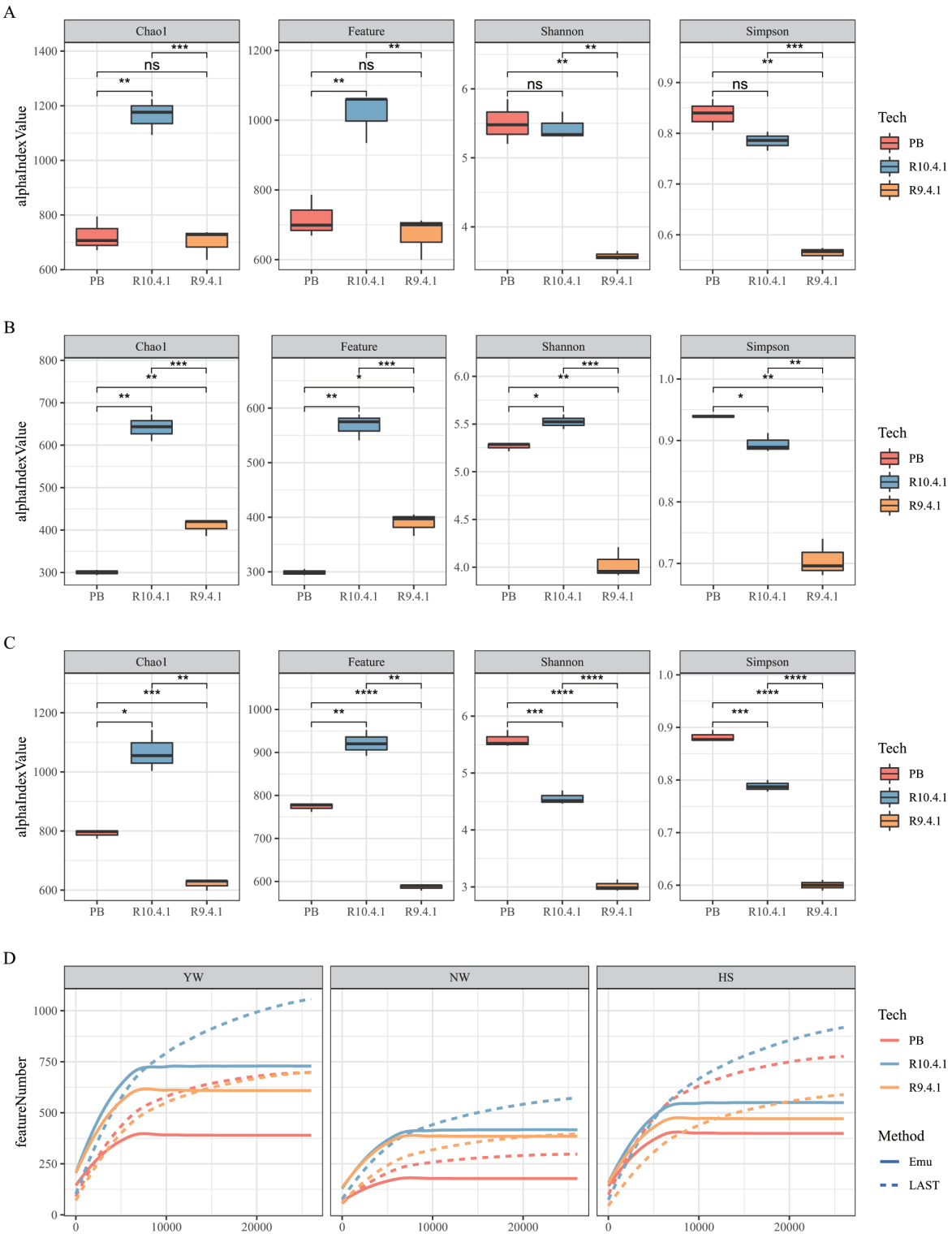

FIG 5 Alpha diversity plot for environmental samples for the evaluation of the ONT platform to profile the complex microbial community microbiota at species and genus levels. (A) Water sample YW; (B) water sample NW; (C) soil sample HS. (D). Rarefaction curves based on reads obtained from R10.4.1 (blue) compared with PB (red) and R9.4.1 (orange). Rarefaction curves separate the reads obtained from three different platforms. The LAST and Emu aligners were used.

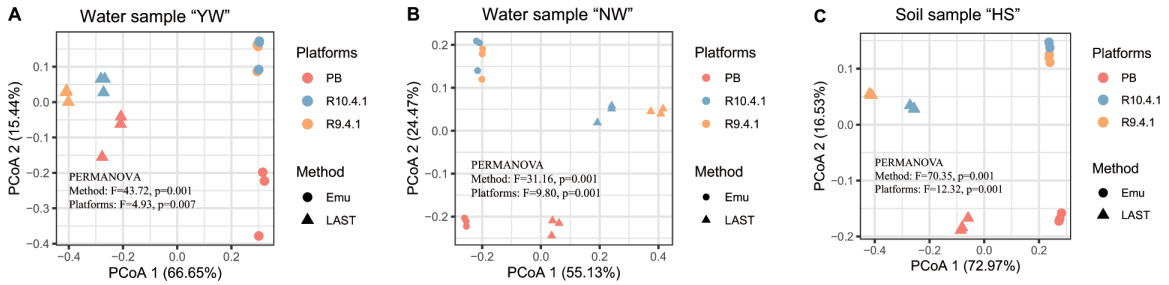

FIG 6 The PCOA and PERMANOVA analysis of environmental samples. (A) Water sample YW; (B) water sample NW; (C) soil sample HS.

DISCUSSION

PB and ONT long-read amplicon sequencing are getting popular. Accuracy is essential for amplicon sequencing. Previous research relied on the R9.4.1 (R9) flowcell. A previous comparison of Rumen microbiota reported MinION and PB platforms applied successfully at the species level (15). ONT has improved its technique over the years. Recently, the raw read accuracy of R10.4.1 data has been improved to ~99%, which implies that the bioinformatics tool should enhance its methodology and establish parameters for R10.4.1. There are currently few published studies on the performance of R10.4.1/Kit14 for 16S analyses. Its performance in amplicon analysis needs to be evaluated. In this study, we described the performance of various sequencing platforms' toolkits.

PB's accuracy can be improved by acquiring CCS sequences, which are comparable to those for the Illumina 454 platforms (27). Early on, Nanopore R7 was only ~64% raw accurate (28). R9 increased sequencing yield and accuracy by ~87% (29). The R9.4.1 achieved raw accuracy of 85%–94% with an average of 92%, which is being used commercially (20). In our study, R10.4.1 has a lower error rate than R9.4.1. We re-analyze the error profile using public long-read 16S data from the Zymo community (5, 8, 30–32). The error rate of R10.4.1 is about a quarter of that of R9, especially since insertion and deletion were reduced by ~90%; meanwhile, the error rate for PB was ~0.02% (Table S9). The synthetic community's higher 16S diversity may explain its lower error rate than Zymo's. Although not as accurate as PB CCS reads, R10.4.1 also achieved unprecedented accuracy. After improving the chemicals and algorithms, it is thought that the accuracy of ONT R10.4.1 amplification can reach the accuracy of individual genome sequencing (Table 1).

As previously described, no pipeline called all anticipated sequences for any mock community at 100% identity (33). The preparation steps show a large bias toward specific 16S rRNA gene variants within an organism. Most of the time, the Operational Taxonomic Unit (OTU)/amplicon sequence variants (ASV) method is used for the traditional short-read 16S (33, 34). ASV can analyze more than one set of data at once, and it can only use the reference genome when the taxonomy has been annotated. However, the ASV method may eliminate some species with extremely low abundance in the sample as a result of a sequencing error. Therefore, ASV needs an error model that is accurate enough to find the wrong sequences (35, 36). PB sequences are highly accurate, and clustering or mapping-based methods can be used for amplicons. Some researchers used 85% identity OTU with qiime2 to classify ONT data (https://github.com/DeniRibicic/q2ONT). However, 85% of the identities cannot suit the analysis at the species level by QIIME2 (37).

Although R10.4.1 data cannot achieve the same level of fidelity as PB data, it has been greatly enhanced in comparison to R9.4.1 data. In particular, an average of 77.11% of reads with similarity greater than 97% were found in the synthetic community (Fig. 1C), which is suitable for taxonomic classification with requiring at least a 97% sequence identity threshold (38).

In the Zymo sample, R10.4.1 data at both the genus and species levels are better, although there are more false-positive species and a larger L1 distance than PB. Yet,

in the synthetic sample, when using LAST_NCBI, R10.4.1 obtained the minimum L1 distance. Sometimes PB may miss some TP species with low abundance, especially in different species of the same genus. In addition, for R10.4.1 data, the Emu technique has not been tested before. Herein, Emu has poor performance when used for the detection of different species in the same genus. Although Emu methods yield fewer FPs, they also have fewer TPs. This leads to the detection of fewer species in the original samples. A previous study reported that more than 99% of R9 reads were correctly categorized at the genus level and up to 40% were incorrectly classified at the species level (13). In contrast, the incorrectly classified reads of R10.4.1 are greatly reduced. Interestingly, PB had the largest L1 distance in S1 but the lowest L1 distance in S2. These two synthetic communities, having the same members but different compositions, could have very different performances for PacBio. The results of the recall also prove that the performance of PB in complex samples is relatively poor. The proportion of chimeras may lead to bias in community composition, and the average chimera of the S1 sample was 10.78%, and that of the S2 sample was 3.21%. Meanwhile, sequencing bias, amplification preference, chimeras, and other reasons can lead to such errors.

We also evaluated the two commonly used 16S databases: the SILVA and NCBI 16S databases. Our findings indicate that the database adds to the variation in outcomes (Fig. S2; Table S5). In the synthetic community, the Silva database typically has a larger L1 distance and lower TPs. Moreover, we pinpoint the underlying reasons. Firstly, the 16S sequences of the SILVA database are highly similar to those in the synthetic community, such as *Enterobacter cloacae*, which is classified as *Enteractor ludwigii*. Similar results have been reported in previous studies. Winand et al. confirmed that the NCBI 16S database often yields greater benefits for bacterial identification (13). *S. enterica* was correctly recognized at the genus level in two and six samples by using the SILVA and NCBI 16S databases, respectively (39). *Azomonas* was identified as *Pseudomonas* in the SILVA database, but it is not present in the NCBI database (40). The same 16S sequence within the genus is common, such as among *Rickettsia japonicum*, *Mesorhizobium loti*, and *Mesorhizobium amorphae*, which leads to misclassification. In our study, *Raoultella* sp. X13 was incorrectly identified as *Raoultella ornithinolytica* and *Klebsiella aerogenes*. *Raoultella* sp. X13, which had highly similar 16S rRNA gene sequences to *R. ornithinolytica* NBRC 105727, was mistakenly assigned to both *R. ornithinolytica* and *K. aerogenes* (41). However, the misclassification could be avoided by using the whole genome rather than the 16S rRNA gene sequence. Ma suggested the reunification of the genus *Raoultella* with the genus *Klebsiella* (42). Besides, different databases have different taxonomic rules and synonyms that can interfere with classification; for instance, *Sinorhizobium fredii* is considered a senior synonym of *Ensifer fredii*. *Streptomyces coelicolor* and *Streptomyces albidoflavus* belong to the *Streptomyces albidoflavus* group, but it is a synonym in Prokaryotic Nomenclature Up-to-Date (43).

The use of the NCBI database helps lessen the likelihood of the mentioned misclassifications. As the NCBI 16S database has fewer sequences, bootstrapping is less likely to identify other bacteria, thus increasing support values. Furthermore, the microbiome of female bladders revealed that the NCBI database had the highest recall accuracy (44). Meanwhile, many published ONT-16S studies are based on the NCBI 16S database (12, 45). In summary, the NCBI database is prepared for analyses of ONT data at the species level (13).

Current research has a few limitations, such as the mock communities' lack of complexity. There are a lot more microbes in the real environment. For instance, more than 1,000 species of microbes have been found in the human gut (46), while 1 g of soil could have more than 109 different kinds of microbes (47). Therefore, evaluating more complex communities and environmental samples is necessary for future applications of ONT R10.4.1 amplicons. In addition, instead of researching comprehensive methods for classified tools, we only adopted two methods: LAST and Emu. LAST is a classical aligner for long reads. Emu utilizes the minimap2 program for mapping, which is a faster long-read aligner, and even uses the EM algorithm to correct the error rate and provide

optimized results. Yet, different tools display a bias for data with various error rates. Furthermore, the database is another effective factor in species classification.

Thus, in the analysis of the environmental samples, the classical LAST aligner against the NCBI database is more recommended. Even though R10.4.1 is more accurate than R9.4.1 but less accurate than PB, it has identified more species in environmental data than other methods (Table 3), suggesting that sequencing errors do not lead to excessive identification of species and may be related to sequencing sensitivity or bias. The official ONT white paper claims that ONT detects more species, especially uncommon and low-abundance species, and several studies have validated this. Species classification is significantly influenced by tools and algorithms. The LAST or Emu directly yields virtually full annotations, although reads with low identity will lead to misclassification (15). Low-abundance species can affect subsequent studies such as diversity and community assembly. However, the use of the LCA algorithm will lead to the classification of poor-quality data as unclassified. But it will reduce the identification of species.

Most of the previous studies focused on the comparison of Illumina and Nanopore (48–51). The results showed that the classification was generally consistent at and above the genus level. Wei et al. compared the correlations of taxa at the genus level and species level by using Miseq and Minion R9.4.1, showing that ONT and Miseq have the consistency of the 20 most abundant OTUs at the genus level but not at the species level (52). Besides, Epi2me and CLC genome workflows did not correlate. Similarly, in our study, ONT data from different methodologies differed greatly. In the current study, the 15 most abundant taxa on different platforms were compared (Fig. 4; Fig. S5 and S6). There are many intersections between R10.4.1 and PB. Meanwhile, different types of samples showed different correlations, with species-level correlations ranging from 0.51 (YW) to 0.99 (HS), indicating consistency in the abundance of major species between PB and R10.4.1 (Fig. 4). A study on the rumen microbiome compared PB and R9.4 at the species level, which showed the consistency of PB and R9 at the species level (15). Furthermore, our results indicate that R10.4.1 and PB exhibit similar levels of diversity, while R9.4.1 shows a more pronounced difference in diversity compared to PB (Fig. 5A through C).

Typically, 20,000 to 30,000 PB reads and 40,000 to 50,000 ONT reads are necessary for the samples to reach the saturation phase. The cost of 50,000 ONT sequences is nearly that of 30,000 PB sequences, according to a recent promotion. The current study projected per 10K reads is CNY ¥ 50 (ONT R10.4.1), CNY ¥ 45 (ONT R9.4.1), CNY ¥ 200 (PB), and CNY ¥ 30 (Novaseq). Considering the advantages of high throughput and portability of ONT sequencing, we believe that more researchers will employ ONT sequencing to explore environmental microbes.

ONT data will not yield the accuracy of PB data in the short term. To achieve proper species-level resolution, it is necessary to construct specialized databases for the various types of samples (53, 54), for example, MiDAS4 (55). More studies show that species-level identification of near-full-length ribosomal operons is possible (26, 56, 57). After database and analysis methods are developed with improved ONT accuracy, they can be widely employed (58, 59). Besides, high error rates and biased bioinformatics tools are limited to ONT amplicons. Although a variety of software for ONT 16S data has been developed, directly supervised by each of the tools, all are dealing with the high error rate (22–24, 60). Moreover, improvements in PCR conditions and experimental design methods are required.

Our findings show that the ONT R10.4.1 flowcell for full-length 16S enabled species-level taxonomic identification for environmental samples. We hope that in the future, with fewer sequencing errors and better bioinformatics tools specifically for Nanopore sequencers, the large-read amplicons will replace the short-read amplicons that are currently being used and revolutionize the field.

Conclusions

In the current research work, we have for the first time compared the ONT R10.4.1 amplicon with the PB amplicon in 16S rRNA. First, we evaluate the performance of R10.4.1 on mock samples. In the synthetic community, 77.10% of the R10.4.1 reads were more than 97% identical to the reference (Fig. 1C). ONT R10.4.1 achieved a good recall at species levels (Tables S5 to S7). In addition, for environmental samples, R10.4.1 can obtain more species than PB (Table 3); meanwhile, PB and R10.4.1 are closer on the Shannon and Simpson indices (Fig. 5). Among the major species, R10.4.1 and PB platforms reveal a similar microbiome. In the current research, we tried to show that both classifiers and databases affect species composition (Fig. 6). Finally, we demonstrated that the ONT R10.4.1 flowcell for full-length 16S enabled species-level taxonomic identification for environmental samples.

MATERIALS AND METHODS

Mock communities, synthetic mock agricultural communities, sample collection, and genomic DNA extraction

The "mock sample" includes two types of communities: "commercial community" and "laboratory synthetic community." The ZymoBIOMICSTM Microbial Community DNA Standard (https://www.zymoresearch.com/zymobiomics-community-standard, D6305) was employed as the reference sample. The mock community contains the genomic DNA of eight bacterial species and two fungi. The synthetic community contains 12 species commonly isolated in the agricultural environment, obtained from the State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University. *Streptomyces coelicolor* A (3)2 and *S. lividans* ZX1 were cultured at 28°C in Murashig and Skoog medium. *Pseudomonas putida* TK2440, *Achromobacter* sp. H380, *Serratia marcescens*, *Raoultella* sp. X13, and *Enterobacter ludwigii* were cultured in Luria-Bertani (LB) liquid medium at 28°C. *Salmonella enterica* (ATCC14028) was cultured overnight at 37°C in LB broth. *Mesorhizobium huakuii* 7653R and *Sinorhizobium fredii* HH103 were cultured for 3 days at 28°C in a Trypticase-Yeast extract medium. *Bradyrhizobium diazoefficiens* USDA110 was cultured in Yeast Mannitol Agar medium.

Genomic DNA quantification, 16S qRT-PCR, and construction of synthetic community

The concentration of purified genomic DNA was initially determined using Qubit 3.0 (Thermo Scientific) according to the manufacturer's instructions. The 341 F/805 R primers were used to amplify the 16S rRNA gene for qRT-PCR (61). Reactions were run in volumes of 20 µL, containing 10 µL of MonAmp Fast SYBR Green qPCR Mix [MonAmp SYBRR Green qPCR Mix (Low ROX), MQ10201S], 8.2 µL of nuclease-free water, 0.4 µL of forward primers (10 µM), 0.4 µL of reverse primers (10 µM), and 1 µL of DNA template. PCR amplification was carried out with the following protocol: 30 s at 95°C and 40 cycles of 15 s at 95°C, 15 s at 56°C, and 45 s at 72°C. The Bio-Rad CFX Manager software was used to perform data analysis.

The synthetic communities were mixed according to the Max/Min DNA content ratio (1:20 and 1:1,000) to construct two mock communities (named S1 and S2). The theoretical bacterial compositions are provided in Table S1.

Environmental sample collection and pretreatment

Water samples were collected from Lake Nanhu (114.37°E, 30.48°N, Wuhan, China) "NW" and the bank of the Xinwuli Yangtze River Green Shipping Comprehensive Service area "YW" (114.25°E, 30.51°N, Wuhan, China). Three replicates are spaced at intervals of 5 m. Soil samples "HS" were collected in three replicates from the foot of Shizi Mountain (114.37°E, 30.48°N, Huazhong Agricultural University, Wuhan, China). Samples

were collected in July 2022. The soils were stored in the freezer (−20°C). Water samples were filtered using sterile 0.22-µm MCE membranes and a Pall vacuum manifold (Wuhan Century Trusty Technology, HX-PV91).

DNA extraction

DNA was extracted from 0.3 g samples using the Qiagen DNeasy Powersoil Pro Kit No. 47014 (Qiagen). Nanodrop ONE (Thermo Scientific) was used for DNA quality control, and Qubit 4.0 (Thermo Scientific) was used for quantification with the Invitrogen dsDNA HS Assay Kit. Water samples "100 mL" were mixed thoroughly and filtered using sterile 0.22-µm MCE filter membranes (Shanghai Xingya Purification Materials Factory) and a Pall vacuum manifold (Wuhan Century Trusty Technology, HX-PV91). The 16S sequences of *B. diazoefficiens* were verified by sanger sequencing.

16S rRNA gene primer pairs, amplicon, library generation, and sequencing

For the Novaseq platform, the 16S universal primer 341 F/805 R (7) targets the V3-V4 region of the 16S gene. The PCR program was performed under the following conditions: denaturation for 3 min at 95°C; 25 cycles of denaturation at 95°C for 30 s, at 55°C for 30 s, and at 72°C for 15 s; and final extension at 72°C for 5 min. The purified DNA samples were quantified using the Qubit 4.0 fluorometer (Invitrogen, Thermo Fisher Scientific, Oregon, USA). Then, the sequencing libraries were built using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, MA, USA), following the manufacturer's instructions. Finally, the library was sequenced on an Illumina NovaSeq 6000 platform at Wuhan Benagen Technology Company Limited (Wuhan, China). The 27 F/1492 R(62)targets the full-length 16S rRNA gene and was only used on the ONT and PB platforms. For the PB platform, the KOD One PCR Master Mix (TOYOBO Life Science) was used to perform 25 cycles of PCR amplification, with initial denaturation at 95°C for 2 min, followed by 25 cycles of denaturation at 98°C for 10 s, annealing at 55°C for 30 s, extension at 72°C for 1 min 30 s, and a final step at 72°C for 2 min. The PCR amplicons were purified with Agencourt AMPure XP Beads (Beckman Coulter, Indianapolis, IN, USA) and quantified using the Qubit dsDNA HS Assay Kit and Qubit 4.0 Fluorometer (Invitrogen, Thermo Fisher Scientific, Oregon, USA). SMRTbell libraries were prepared from the amplified DNA using the SMRTbell Express Template Prep Kit 2.0 according to the manufacturer's instructions (Pacific Biosciences). Purified SMRTbell libraries from the pooled and barcoded samples were sequenced on a single PacBio Sequel II 8 M cell using the Sequel II Sequencing Kit 2.0. For the ONT platform, the first PCR protocol was 30 s at 94°C, 12 cycles of 30 s at 94°C, 20 s at 60°C, 65°C for 2 min, and 10 min at 65°C. The second PCR barcoded primer in 10 cycles. The barcoded amplicons were purified using the AMPure XP beads (Beckman Coulter, Brea, CA, USA) as per Nanopore's instructions. Samples were then quantified using the Qubit 4.0 fluorometer (Life Technologies, Carlsbad, USA) and pooled in equimolar concentrations. A pooled DNA mixture was constructed using 1 µg of purified PCR product from each sample for library construction. The library was built using the SQK-LSK110 (for R9.4.1 flow cell) and SQK-LSK114 (for R10.4.1 flow cell) ligation kits (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's instructions. The purified library was loaded onto R9.4.1 and R10.4.1 flowcells and then sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) with 48-h runs at Wuhan Benagen Technology Company Limited (Wuhan, China).

Data analyses

For Novaseq data, an average of 50,000–60,000 reads were generated for each of the samples in the present study. Cutadapt (v.3.5) software was used to identify and remove primer sequences and perform length filtering to obtain clean sequences.

For ONT data, the raw fast5 files were base-called using Guppy version 6.2.7 to generate fastq files. Raw reads were demultiplexed and adapters trimmed using

Porechop (version 0.2.4, https://github.com/rrwick/porechop). Next, all reads shorter than 1.2 kbp and longer than 1.8 kbp were removed with Nanofilt (version 2.5.0, https://github.com/wdecoster/nanofilt).

For PB data, the raw reads generated from sequencing were filtered and demultiplexed using the SMRT Link software (version 8.0) with minPasses ≥5 and minPredictedAccuracy ≥0.9, to obtain the CCS reads. Subsequently, Lima (version 1.7.0) was employed to assign the CCS sequences to the corresponding samples based on their barcodes. CCS reads containing no primers and those reads beyond the length range (1,200–1,800 bp) were discarded through the recognition of forward and reverse primers and quality filtering using the Cutadapt quality control process (version 2.7).

Database information and classification methods

The 16S rRNA gene of eight bacterial species was constructed in a database for "Zymo-specific." Thus, 16S from the 12 species in our synthetic community created a database for "Synthesis-specific." These two databases are named "Self-database." Then, the NCBI 16S database and the SILVA 138 database were used. Then, the percentage identity matrix of 16S sequences was calculated using Clustal2.1 (63).

In this study, three methods were used for classification. Firstly, for the identification of feature sequences (ASVs), DADA2 (36) was used to filter raw data, splice paired-end reads, and remove chimeric sequences with QIIME2 software. Subsequently, ASVs were assigned to taxonomic groups for species annotation by using an SK-learn algorithm. This method was used for Novaseq and PB data. Secondly, the LAST program (parameters: -s 2 T 0 -a 1 -b 1 -q 1 -e 45 -m 500) was used to map clean reads to the database. For sequences whose identity exceeds 90%, the taxonomy was established based on the best matches; otherwise, the LCA of the best matching taxonomy estimation was used. Thirdly, Emu was used for taxonomy assignment with minimap2 by default. The latter two methods were used for PB data and ONT data, respectively.

Reads evaluation metrics of the Zymo and Synthesis community

Reads were used in minimap2 against the reference sample. The error profile was determined by count-errors.py (https://github.com/arq5x/nanopore-scripts/blob/master/count-errors.py), calculating mismatches, insertions, and deletions by MN and CIGAR tags (64). IGV was used for visualization (65). The identity distribution of reads was described based on the results of LAST against the Self-reference sequence. Reads were mapped to the Self database for accurate classification of each read by individual methods. Then, reads in either the SILVA or the NCBI results with Self-reference were considered to be correctly classified reads. The proportion of correctly classified reads from different classifiers and different databases was counted, respectively.

Accuracy evaluation matrices

L1 distance is essentially the linear error and is calculated using the equation $s \in \mathrm{S}|\mathrm{Es} - \mathrm{Is}|$. where set S is the union of all species in the database and the threshold values, and Es and Is are the estimated and observed relative abundances for species "s." "TP" represents true positives, "FP" represents false positives, and "FN" represents false negatives. Precision is defined as the percentage of TP that exists in the sample: $(TP)/(TP + FP)$. Recall shows how many true positives were picked up: $(TP)/(TP + FN)$. The *F*-score is calculated by taking the harmonic mean of the two values $2 \cdot (\mathrm{precision} \times \mathrm{recall})/(\mathrm{precision} + \mathrm{recall})$. All mock results are trimmed to include only taxa with an abundance of ≥0.01%, before the calculation of performance metrics.

Statistical analysis for environmental samples

PCC (r) and *P*-value of significance between microbial compositions obtained by different methods were calculated using the Python package "scipy.stats.pearsonr." Heat maps

were generated using the R package "pheatmap." An α-diversity index was calculated with a vegan package (66) based on rarefied abundance counts. A *t*-test was used for calculating whether these samples are significantly different with a *P*-value of <0.05 ($^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$). The Bray-Curtis dissimilarity matrix was also calculated. PCoA and PERMANOVA were used to analyze statistical differences between platforms and aligners based on the Bray-Curtis matrix. The statistical analysis and graphical visualizations were performed in R version 4.0.3.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (42020104003).

conceptualization, T.Z., Q.H., and W.C.; methodology, T.Z. and HZ.L.; software, T.Z. and HZ.L.; validation, HZ.L., S.M., and W.C.; formal analysis, T.Z., HZ.L., and S.M.; investigation, J.C. and S.M.; resources, H.L., Q.H., and W.C.; data curation, HZ.L.; writing-original draft preparation, T.Z.; writing-review and editing, H.L., Q.H., and W.C.; visualization, T.Z. and HZ.L.; supervision, W.C.; project administration, Q.H. and W.C.; funding acquisition, Q.H. and W.C. All the authors have read and agree to publish the manuscript.

The authors declare no conflict of interest.

AUTHOR AFFILIATIONS

[1]National Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China
[2]College of Life Science and Technology, Huazhong Agricultural University, Wuhan, China
[3]Wuhan Benagen Technology Co., Ltd., Wuhan, China
[4]Hubei Key Laboratory of Soil Environment and Pollution Remediation, Huazhong Agricultural University, Wuhan, China

AUTHOR ORCIDs

Tianyuan Zhang http://orcid.org/0000-0001-8968-563X
Hanzhou Li http://orcid.org/0000-0001-6861-404X
Silin Ma http://orcid.org/0000-0003-1828-4032
Jian Cao http://orcid.org/0000-0003-4797-397X
Hao Liao http://orcid.org/0000-0002-5561-9295
Qiaoyun Huang http://orcid.org/0000-0002-2733-8066
Wenli Chen http://orcid.org/0000-0003-1717-1263

FUNDING

Funder	Grant(s)	Author(s)
MOST | National Natural Science Foundation of China (NSFC)	42020104003	Qiaoyun Huang

AUTHOR CONTRIBUTIONS

Tianyuan Zhang, Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – original draft | Hanzhou Li, Data curation, Formal analysis, Methodology, Software, Visualization | Silin Ma, Formal analysis, Investigation, Validation | Jian Cao, Investigation | Hao Liao, Resources, Writing – review and editing | Qiaoyun Huang, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review and editing | Wenli Chen, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review and editing

DATA AVAILABILITY

The data collected were from the BioProject accession number PRJNA925180.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental file 1 (AEM00605-23-s0001.docx). Legends to Tables S1 to S9 and legends to Fig. S1 to S6.
Supplemental file 2 (AEM00605-23-s0002.xlsx). Tables S1 to S9.
Supplemental file 3 (AEM00605-23-s0003.eps). Fig. S1.
Supplemental file 4 (AEM00605-23-s0004.eps). Fig. S2.
Supplemental file 5 (AEM00605-23-s0005.eps). Fig. S3.
Supplemental file 6 (AEM00605-23-s0006.eps). Fig. S4.
Supplemental file 7 (AEM00605-23-s0007v2.eps). Fig. S5.
Supplemental file 8 (AEM00605-23-s0008.eps). Fig. S6.

REFERENCES

1. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635–645. https://doi.org/10.1038/nrmicro3330
2. Slatko BE, Gardner AF, Ausubel FM. 2018. Overview of next-generation sequencing technologies. Curr Protoc Mol Biol 122:e59. https://doi.org/10.1002/cpmb.59
3. Na HS, Yu Y, Kim SY, Lee J-H, Chung J. 2020. Comparison of the performance of MiSeq and HiSeq 2500 in a microbiome study. J Microbiol Biotechnol 48:574–581. https://doi.org/10.48022/mbl.2008.08003
4. Jia Y, Zhao S, Guo W, Peng L, Zhao F, Wang L, Fan G, Zhu Y, Xu D, Liu G, Wang R, Fang X, Zhang H, Kristiansen K, Zhang W, Chen J. 2022. Sequencing introduced false positive rare taxa lead to biased microbial community diversity, assembly, and interaction interpretation in amplicon studies. Environ Microbiome 17:43. https://doi.org/10.1186/s40793-022-00436-y
5. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 10:5029. https://doi.org/10.1038/s41467-019-13036-1
6. Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW. 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res 38:e200–e200. https://doi.org/10.1093/nar/gkq873
7. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 41:e1. https://doi.org/10.1093/nar/gks808
8. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res 47:e103–e103. https://doi.org/10.1093/nar/gkz569
9. Wang S, Su X, Cui H, Wang M, Hu X, Ding W, Zhang W. 2022. Microbial richness of marine biofilms revealed by sequencing full-length 16S rRNA genes. Genes 13:1050. https://doi.org/10.3390/genes13061050
10. Ammer-Herrmenau C, Ellenrieder V, Neesse A. 2022. Chronic pancreatitis: update diagnostic and therapeutic concepts. Z Gastroenterol 60:1131–1138. https://doi.org/10.1055/a-1659-4636
11. Heikema AP, Horst-Kreft D, Boers SA, Jansen R, Hiltemann SD, de Koning W, Kraaij R, de Ridder MAJ, van Houten CB, Bont LJ, Stubbs AP, Hays JP. 2020. Comparison of illumina versus nanopore 16S rRNA gene sequencing of the human nasal microbiota. Genes (Basel) 11:1105. https://doi.org/10.3390/genes11091105
12. Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, Iwai T, Imanishi T, Hirota K. 2019. Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION™ nanopore Sequencer . FEBS Open Bio 9:548–557. https://doi.org/10.1002/2211-5463.12590
13. Winand R, Bogaerts B, Hoffman S, Lefevre L, Delvoye M, Braekel JV, Fu Q, Roosens NH, Keersmaecker SCD, Vanneste K. 2019. Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (illumina) and third (Oxford Nanopore Technologies) generation sequencing technologies. Int J Mol Sci 21:298. https://doi.org/10.3390/ijms21010298
14. Mann BC, Bezuidenhout JJ, Swanevelder ZH, Grobler AF. 2021. MinION 16S datasets of a commercially available microbial community enables the evaluation of DNA extractions and data analyses. Data Brief 36:107036. https://doi.org/10.1016/j.dib.2021.107036
15. Miura H, Takeda M, Yamaguchi M, Ohtani Y, Endo G, Masuda Y, Ito K, Nagura Y, Iwashita K, Mitani T, Suzuki Y, Kobayashi Y, Koike S. 2022. Application of MinION amplicon sequencing to buccal SWAB samples for improving resolution and throughput of rumen microbiota analysis. Front Microbiol 13:783058. https://doi.org/10.3389/fmicb.2022.783058
16. Wiryawan A, Eginarta W, Hermanto F, Ustiatik R, Dinira L, Mustafa I. 2022. Changes in essential soil nutrients and soil disturbance directly affected soil microbial community structure – a metagenomic approach. J Ecol Eng 23:238–245. https://doi.org/10.12911/22998993/149972
17. Hong NTT, Nghia HDT, Thanh TT, Lan NPH, Ny NTH, Ngoc NM, Hang VTT, Chau LTM, Quynh VX, Diem LT, Hanh BTB, Hanh NHH, Duc DT, Man DNH, Campbell J, Oanh PKN, Day J, Phu NH, Chau NVV, Thwaites G, Tan LV. 2020. Cerebrospinal fluid MinION sequencing of 16S rRNA gene for rapid and accurate diagnosis of bacterial meningitis. Journal of Infection 80:469–496. https://doi.org/10.1016/j.jinf.2019.12.011
18. Jang Y, Kim S, Kim N, Son H, Ha EJ, Koh EJ, Phi JH, Park CK, Kim JE, Kim SK, Lee SK, Cho WS, Moon J, Chu K. 2022. Nanopore 16S sequencing enhances the detection of bacterial meningitis after neurosurgery . Ann Clin Transl Neurol 9:312–325. https://doi.org/10.1002/acn3.51517
19. Omi M, Matsuo Y, Araki-Sasaki K, Oba S, Yamada H, Hirota K, Takahashi K. 2022. 16S rRNA nanopore sequencing for the diagnosis of ocular infection: a feasibility study. BMJ Open Ophthalmol 7:e000910. https://doi.org/10.1136/bmjophth-2021-000910
20. Huang Y-T, Liu P-Y, Shih P-W. 2021. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome Biol. 22:95. https://doi.org/10.1186/s13059-021-02282-6

21. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods 19:823–826. https://doi.org/10.1038/s41592-022-01539-7
22. Rodríguez-Pérez H, Ciuffreda L, Flores C. 2021. Nanoclust: a species-level analysis of 16S rRNA nanopore sequencing data. Bioinformatics 37:1600–1601. https://doi.org/10.1093/bioinformatics/btaa900
23. Curry KD, Wang Q, Nute MG, Tyshaieva A, Reeves E, Soriano S, Wu Q, Graeber E, Finzer P, Mendling W, Savidge T, Villapol S, Dilthey A, Treangen TJ. 2022. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. Nat Methods 19:845–853. https://doi.org/10.1038/s41592-022-01520-4
24. Jung A, Chorlton SD. 2021. BugSeq 16S: nanoCLUST with Improved Consensus Sequence Classification. Bioinformatics. https://doi.org/10.1101/2021.03.16.434153
25. Li C, Chng KR, Boey EJH, Ng AHQ, Wilm A, Nagarajan N. 2016. INC-Seq: accurate single molecule reads using nanopore sequencing. Gigascience 5:34. https://doi.org/10.1186/s13742-016-0140-7
26. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M. 2021. High-accuracy long-read amplicon sequences using unique molecular identifiers with nanopore or PacBio sequencing. Nat Methods 18:165–169. https://doi.org/10.1038/s41592-020-01041-y
27. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ 4:e1869. https://doi.org/10.7717/peerj.1869
28. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance Island. Nat Biotechnol 33:296–300. https://doi.org/10.1038/nbt.3103
29. Minei R, Hoshina R, Ogura A. 2018. De novo assembly of middle-sized genome using MinION and illumina sequencers. BMC Genomics 19:700. https://doi.org/10.1186/s12864-018-5067-1
30. Burton AS, Stahl SE, John KK, Jain M, Juul S, Turner DJ, Harrington ED, Stoddart D, Paten B, Akeson M, Castro-Wallace SL. 2020. Off earth identification of bacterial populations using 16S rDNA nanopore sequencing. Genes (Basel) 11:76. https://doi.org/10.3390/-genes11010076
31. Park C, Kim SB, Choi SH, Kim S. 2021. Comparison of 16S rRNA gene based microbial profiling using five next-generation sequencers and various primers. Front Microbiol 12:715500. https://doi.org/10.3389/fmicb.2021.715500
32. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. 2018. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. F1000Res 7:1755. https://doi.org/10.12688/f1000research.16817.2
33. Nearing JT, Douglas GM, Comeau AM, Langille MGI. 2018. Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ 6:e5364. https://doi.org/10.7717/peerj.5364
34. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R, Gilbert JA. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:mSystems https://doi.org/10.1128/mSystems.00191-16
35. Joos L, Beirinckx S, Haegeman A, Debode J, Vandecasteele B, Baeyen S, Goormachtig S, Clement L, De Tender C. 2020. Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. BMC Genomics 21:733. https://doi.org/10.1186/s12864-020-07126-4
36. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869
37. Santos A, van Aerle R, Barrientos L, Martinez-Urtaza J. 2020. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. Comput Struct Biotechnol J 18:296–305. https://doi.org/10.1016/j.csbj.2020.01.005
38. Benítez-Páez A, Portune KJ, Sanz Y. 2016. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore Sequencer . Gigascience 5:4. https://doi.org/10.1186/s13742-016-0111-z
39. Anzai Y, Kim H, Park JY, Wakabayashi H, Oyaizu H. 2000. Phylogenetic affiliation of the pseudomonads based on 16S rRNA sequence. Int J Syst Evol Microbiol 50 Pt 4:1563–1589. https://doi.org/10.1099/00207713-50-4-1563
40. Özen AI, Ussery DW. 2012. Defining the Pseudomonas genus: where do we draw the line with Azotobacter Microb Ecol 63:239–248. https://doi.org/10.1007/s00248-011-9914-8
41. Drancourt M, Bollet C, Carta A, Rousselier P. 2001. Phylogenetic analyses of Klebsiella species delineate Klebsiella and Raoultella gen nov., with description of Raoultella ornithinolytica comb. nov., Raoultella terrigena comb. nov. and Raoultella planticola comb. Int J Syst Evol Microbiol 51:925–932. https://doi.org/10.1099/00207713-51-3-925
42. Ma Y, Wu X, Li S, Tang L, Chen M, An Q. 2021. Proposal for reunification of the genus Raoultella with the genus Klebsiella and reclassification of Raoultella electrica as Klebsiella electrica comb. nov. Res Microbiol 172:103851. https://doi.org/10.1016/j.resmic.2021.103851
43. Parte AC. 2018. LPSN - list of Prokaryotic names with standing in nomenclature (bacterio.net), 20 years on. Int J Syst Evol Microbiol 68:1825–1829. https://doi.org/10.1099/ijsem.0.002786
44. Hoffman C, Siddiqui NY, Fields I, Gregory WT, Simon HM, Mooney MA, Wolfe AJ, Karstens L, Chia N. 2021. Species-level resolution of female bladder microbiota from 16S rRNA amplicon sequencing. mSystems 6. https://doi.org/10.1128/mSystems.00518-21
45. Acharya K, Khanal S, Pantha K, Amatya N, Davenport RJ, Werner D. 2019. A comparative assessment of conventional and molecular methods, including MinION nanopore sequencing, for surveying water quality. Sci Rep 9:15726. https://doi.org/10.1038/s41598-019-51997-x
46. Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, Wan D, Jiang R, Su L, Feng Q, Jie Z, Guo T, Xia Z, Liu C, Yu J, Lin Y, Tang S, Huo G, Xu X, Hou Y, Liu X, Wang J, Yang H, Kristiansen K, Li J, Jia H, Xiao L. 2019. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol 37:179–185. https://doi.org/10.1038/s41587-018-0008-8
47. Sokol NW, Slessarev E, Marschmann GL, Nicolas A, Blazewicz SJ, Brodie EL, Firestone MK, Foley MM, Hestrin R, Hungate BA, Koch BJ, Stone BW, Sullivan MB, Zablocki O, LLNL Soil Microbiome Consortium, Pett-Ridge J. 2022. Life and death in the soil microbiome: how ecological processes influence biogeochemistry. Nat Rev Microbiol 20:415–430. https://doi.org/10.1038/s41579-022-00695-z
48. Shin J, Lee S, Go M-J, Lee SY, Kim SC, Lee C-H, Cho B-K. 2016. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Sci Rep 6:29681. https://doi.org/10.1038/srep29681
49. Quan L, Dong R, Yang W, Chen L, Lang J, Liu J, Song Y, Ma S, Yang J, Wang W, Meng B, Tian G. 2019. Simultaneous detection and comprehensive analysis of HPV and microbiome status of a cervical liquid-based cytology sample using nanopore MinION sequencing. Sci Rep 9:19337. https://doi.org/10.1038/s41598-019-55843-y
50. Wongsurawat T, Nakagawa M, Atiq O, Coleman HN, Jenjaroenpun P, Allred JI, Trammel A, Puengrang P, Ussery DW, Nookaew I. 2019. An assessment of Oxford Nanopore sequencing for human gut metagenome profiling: a pilot study of head and neck cancer patients. J Microbiol Methods 166:105739. https://doi.org/10.1016/j.mimet.2019.105739
51. Nygaard AB, Tunsjø HS, Meisal R, Charnock C. 2020. A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. Sci Rep 10:3209. https://doi.org/10.1038/s41598-020-59771-0
52. Wei P-L, Hung C-S, Kao Y-W, Lin Y-C, Lee C-Y, Chang T-H, Shia B-C, Lin J-C. 2020. Characterization of fecal microbiota with clinical specimen using long-read and short-read sequencing platform. Int J Mol Sci 21:7110. https://doi.org/10.3390/ijms21197110
53. F. Escapa I, Huang Y, Chen T, Lin M, Kokaras A, Dewhirst FE, Lemon KP. 2020. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome 8:65. https://doi.org/10.1186/s40168-020-00841-w
54. Dueholm MS, Andersen KS, McIlroy SJ, Kristensen JM, Yashiro E, Karst SM, Albertsen M, Nielsen PH. 2020. Generation of comprehensive ecosystem-specific reference databases with species-level resolution by

high-throughput full-length 16S rRNA gene sequencing. mBio 11:e01557-20. https://doi.org/10.1128/mBio.01557-20
55. Dueholm MKD, Nierychlo M, Andersen KS, Rudkjøbing V, Knutsson S, MiDAS Global Consortium, Albertsen M, Nielsen PH. 2022. MIDAS 4: a global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. Nat Commun 13:4017. https://doi.org/10.1038/s41467-022-31423-z
56. Kerkhof LJ, Roth PA, Deshpande SV, Bernhards RC, Liem AT, Hill JM, Häggblom MM, Webster NS, Ibironke O, Mirzoyan S, Polashock JJ, Sullivan RF. 2022. A ribosomal operon database and MegaBLAST settings for strain-level resolution of microbiomes. FEMS Microbes 3:xtac002. https://doi.org/10.1093/femsmc/xtac002
57. Rozas M, Brillet F, Callewaert C, Paetzold B. 2021. MinION™ nanopore sequencing of skin microbiome 16S and 16S-23S rRNA gene amplicons . Front Cell Infect Microbiol 11:806476. https://doi.org/10.3389/fcimb.2021.806476
58. Kinoshita Y, Niwa H, Uchida-Fujii E, Nukada T. 2021. Establishment and assessment of an amplicon sequencing method targeting the 16S-ITS-23S rRNA operon for analysis of the equine gut microbiome. Sci Rep 11:11884. https://doi.org/10.1038/s41598-021-91425-7
59. Seol D, Lim JS, Sung S, Lee YH, Jeong M, Cho S, Kwak W, Kim H. 2022. Microbial identification using rRNA operon region: database and tool for metataxonomics with long-read sequence. Microbiol Spectr 10:e0201721. https://doi.org/10.1128/spectrum.02017-21
60. Vierstraete AR, Braeckman BP. 2022. Amplicon_sorter: a tool for reference-free amplicon sorting based on sequence similarity and for building consensus sequences. Ecol Evol 12:e8603. https://doi.org/10.1002/ece3.8603
61. Lee E, Park S, Um S, Kim S, Lee J, Jang J, Jeong H, Shin J, Kang J, Lee S, Jeong T. 2021. Microbiome of saliva and plaque in children according to age and dental caries experience. Diagnostics 11:1324. https://doi.org/10.3390/diagnostics11081324
62. Lane DJ. 1991. 16S/23S rRNA Sequencing. Nucleic Acid Techniques in Bacterial Systematics 10.4135/9781446279281.n7.
63. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. https://doi.org/10.1093/bioinformatics/btm404
64. Nicholls SM, Quick JC, Tang S, Loman NJ. 2019. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience 8:giz043. https://doi.org/10.1093/gigascience/giz043
65. Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. https://doi.org/10.1093/bib/bbs017
66. Dixon P. 2003. VEGAN, a package of R functions for community ecology. J Veg Sci 14:927–930. https://doi.org/10.1111/j.1654-1103.2003.tb02228.x