Comparison of Oxford Nanopore Technologies and Illumina MiSeq sequencing with mock communities and agricultural soil

Bo Stevens (bo.stevens@usda.gov)
USDA ARS
Timothy Creed
USDA ARS https://orcid.org/0000-0003-4510-1596
Catherine Reardon
USDA ARS
Daniel Manter
USDA ARS

Article

Keywords:

Posted Date: June 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1731798/v1

Abstract

Illumina MiSeq is the current standard for characterizing microbial communities in soil. The Oxford Nanopore Technologies MinION sequencer is quickly gaining popularity because of the low initial cost and longer sequence reads. However, the accuracy of MinION, per base, is much lower than MiSeq (95% versus 99.9%). The effects of this difference in base-calling accuracy on taxonomic and diversity estimates remains unclear. We compared the output of short MiSeq to short and full-length MinION 16S rRNA amplicons from a mock community and agricultural soil samples. For all three methods, we found that taxonomic assignments of the mock community at both the genus and species level matched expectations with minimal deviation (genus: 80.9–90.5%; species: 70.9–85.2% Bray-Curtis similarity); however, the short MiSeq with error correction (DADA2) resulted in the correct estimate of species richness and much lower alpha diversity for soils. Several filtering strategies were tested to improve these estimates with varying results. The sequencing platform also had a significant influence on the relative abundances of taxa with MiSeq resulting in significantly higher abundances Actinobacteria, Chloroflexi, and Gemmatimonadetes and lower abundances of Acidobacteria, Bacteroides, Firmicutes, Proteobacteria, and Verrucomicrobia compared to the MinION platform. When comparing agricultural soils from two different sites (Fort Collins, CO and Pendleton, OR) methods varied in the taxa identified as significantly different between sites. At all taxonomic levels, the full-length MinION method had the highest similarity to the short MiSeq method with DADA2 correction with 73.2%, 69.3%, 74.1%, 79.3%, 79.4%, and 82.28% of the taxa at the phyla, class, order, family, genus, and species levels, respectively, showing similar patterns in differences between the sites. In summary, although both platforms appear suitable for 16S rRNA microbial community composition, biases for different taxa may make the comparison between studies problematic; and even with a single study (i.e., comparing sites or treatments), the sequencing platform can influence the differentially abundant taxa identified.

Introduction

The current standard for characterizing microbiomes is the Illumina MiSeq sequencing platform. MiSeq produces 16S rRNA reads up to 300bp, and around 550bp if forward and reverse reads are joined. Conversely, the Oxford Nanopore Technologies (ONT) MinION sequencer can potentially sequence more than 200,000 base pairs (Santos et al., 2020). The main concern for using MinION sequencing is the lower base-calling accuracy, which is currently estimated around 95% compared to 99.9% for MiSeq (Santos et al., 2020). However, continuous improvements are expected to improve accuracy substantially. The accuracy has increased to 96.5%, up from 65% when the MinION sequencer was first released (Kerkhof, 2021). Initial comparisons of these technologies indicate that MinION may be as good as, or better than MiSeq for taxonomic resolution at the genus and species level (Nygaard et al., 2020). The accuracy of taxonomic assignment at the species level is considered to be low for both technologies (Winand et al., 2020); however, the recent release of a new expectation-maximization algorithm-based classifier (EMU) may improve species level classification; particularly for full-length rRNA sequences (Curry et al., 2021).

Bias and errors are introduced at many steps throughout the data production and analysis pipeline. Different DNA extraction methods can have an effect on the relative abundance of microbes (Wesolowska-Andersen et al., 2014); high GC contents may reduce PCR efficiency (Laursen et al., 2017). PCR conditions such as annealing and denaturation time can have an impact on taxonomic output (Fujiyoshi et al., 2020). Furthermore, the reference database used for identification will also influence taxonomic assignments for both MiSeq and MinION (Nygaard et al., 2020).

Bioinformatic methods are still in development for the MinION sequencer. Throughout the history of MiSeq sequencing, continuous improvements were made to the bioinformatics pipeline, resulting in the removal of spurious sequences that artificially inflated estimates of alpha diversity (Ciuffreda et al., 2021; Edgar, 2017; Nearing et al., 2018; Straub et al., 2020). Previously, sequences were clustered into operational taxonomic units (OTUs) using a similarity threshold (e.g., 99%), to minimize sequencing artifacts. Filtering thresholds based on abundance data have been used to remove rare OTUs that are typically associated with PCR and sequencing errors (Bálint et al., 2016; Bokulich et al., 2013). Currently, denoising techniques provide the best methods to estimate richness of microbial communities. QIIME2 with DADA2 provides the best estimate of richness, based on sequencing of complex mock communities (Almeida et al., 2018; Straub et al., 2020). Unfortunately, DADA2 is not available for MinION output (Ciuffreda et al., 2021).

The purpose of this study was to compare the results of MiSeq output to the MinION sequencer. Specifically, we wanted to answer the following questions: (1) how does the taxonomic assignments of each sequencer and read length compare to internal standards, (2) do patterns in diversity for both mock and soil communities differ between methods, (3) what filtering thresholds may be necessary to match expected alpha diversity levels in mock communities.

Methods

Study sites

Soils were collected from two different sites (ARDEC: Colorado State University's Agricultural Research, Development and Education Center in Fort Collins, CO; and CPCRC: USDA Columbia Plateau Conservation Research Center in Pendleton, OR). At each site, four replicate plots of no-till corn (ARDEC) or no-till annual wheat (CPCRC) were sampled. At ARDEC, the soils are clay loam and CPCRC the soils are Walla Walla silt loams (fine-loamy, mesic Aridic Haplustalls). For each plot, six 1" diameter cores (15 cm deep) were sampled near plant crowns, composited in plastic bags and stored on ice in coolers until transfer to the laboratory. Once in the laboratory, the soils were homogenized by hand, sieved to 4 mm, placed in a plastic bag, and then stored in the freezer (-20°C) until DNA extraction. Prior to freezing, subsamples (~ 5 g) were removed from each sample to measure gravimetric soil water content.

DNA Extraction

DNA was extracted from three replicate 0.25 g soil samples from each plot using the Qiagen DNeasy Powersoil Pro Kit (Qiagen). The extraction process was carried out using a fully automated Qiagen QIAcube robot with a 10-min vortex lysis step. DNA quality was assess using a Nanodrop 1000 (Thermo Scientific) and quantified fluorometrically with the Invitrogen dsDNA HS Assay Kit on a Qubit 2.0 (Life Technologies)

Library Preparation

PCR amplifications were performed on each DNA sample using two different 16S rRNA gene primer pairs. The first primer pair, 341F/806R (Klindworth et al., 2013), targets the V3-V4 region of the 16S gene and was used for both platforms. The second primer pair, 27F/1492R (Lane 1991), targets the full-length 16S rRNA gene and was only used on the ONT MinION platform (Table 1).

Table 1
Summary of platforms and bioinformatics methods compared in this study.

Method	Platform	Adapter / Primer[1]	Target	Classifier	Error-correction
MinION V34	ONT MinION	341F: *TTTCTGTTGGTGCTGATATTGC* CCTACGGGNGGCWGCAG 806R: *ACTTGCCTGTCGCTCTATCTTC* GGACTACHVGGGTATCTAATCC	V3-V4	minimap2	EMU
MinION Full	ONT MinION	27F: *TTTCTGTTGGTGCTGATATTGC* AGRGTTYGATYMTGGCTCAG 1492R: *ACTTGCCTGTCGCTCTATCTTC* TACCTTGTTACGACTT	Full-length	minimap2	EMU
MiSeq V34	Illumina MiSeq	341F: *TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG* CCTACGGGNGGCWGCAG 806R: *GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG* GGACTACHVGGGTATCTAATCC	V3-V4	minimap2	EMU
MiSeq V34 DADA2	Illumina MiSeq	341F: *TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG* CCTACGGGNGGCWGCAG 806R: *GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG* GGACTACHVGGGTATCTAATCC	V3-V4	minimap2	DADA2

[1]Adapters are in italics; gene-specific primers are underlined

ONT MinION PCR Conditions and Library Preparation

Extracted DNA samples were amplified in 60 µL PCR reactions containing 30 µL Phusion HSII (Thermo Scientific, Waltham, MA) master mix, 0.6 µL of each forward and reverse primer (10µM concentration), 21.6 µL molecular grade H_2O, and 6 µL soil DNA diluted 1:20 with nuclease-free water. Reactions were held at 98 °C for 30 s, with amplification proceeding for 25 cycles at 98 °C for 15 s, 50 °C for 15 s, and 72 °C for 60 s with a final extension at 72 °C for 5 min. The PCR products (PCR1) were purified using AMPure XP beads (Beckman Coulter, Indianapolis, IN).

Unique barcodes (EXP-PBC096, ONT, Oxford, UK) were added to both ends of the DNA fragments by PCR. These were 50 µL PCR reactions containing 25 µL Phusion HSII master mix, 19 µL H_2O, 1 µL of each barcode, and 5 µL PCR1 product diluted 1:10 with nuclease-free water. Reactions were held at 98 °C for 30 s, with amplification proceeding for 15 cycles at 98 °C for 15 s, 62 °C for 15 s, and 72 °C for 60 s; a final extension at 72 °C for 5 min. The barcoded products of this PCR reaction were purified a second time using AMPure XP beads (Beckman Coulter, Brea, CA).

Barcoded amplicons from all samples were pooled and prepared for sequencing using the SQK-LSK109 Ligation Sequencing Kit (ONT, Oxford, UK). The library was loaded on a MinION flow cell FLO-MIN106D-R9 (ONT, Oxford,

UK) per manufacturers' protocol and sequencing was started with a runtime of 48 hours and voltage of -180 V. All libraries included no template (H_2O-only) negative controls and a mock community (ZymoBIOMICS Gut Microbiome Standard; Zymo Research, Irvine CA).

MiSeq PCR Conditions and Library Preparation

Extracted DNA was amplified in triplicate, in 20 µL quantitative PCR reactions containing 10 µL Maxima SYBR-green (Thermo Scientific, Waltham, MA), 2 µL of each forward and reverse primer (10uM concentration), 4 µL molecular grade H_2O, and 2 µL soil DNA diluted 1:20 with nuclease-free water. Reactions were held at 95°C for 5 min, with amplification proceeding for 28 cycles at 95°C for 40 s, 55°C for 120 s, and 72°C for 60 s; a final extension at 72°C for 7 min. Thermocycling was performed with a Roche 96 Lightcycler (Roche, Indianapolis, IN). Quantities were determined by comparison of the quantification cycle to a standard curve generated by serial dilution of purified *Pseudomonas putida* KT2440 gDNA. Copy numbers were normalized to the grams of dry soil extracted. The products of the triplicate PCR reactions were pooled and purified using AMPure XP beads.

Nextera XT barcode sequences (Illumina, San Diego, CA) were added to both ends of the DNA fragments by PCR using 50 µL PCR reactions containing 25 µL SYBR-green, 10 µL H_2O, 5 µL of each forward and reverse barcode (5µM concentration), and 5 µL of sample PCR1 product. Reactions were held at 95°C for 3 min, with amplification proceeding for 8 cycles at 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s; a final extension at 72°C for 5 min. The barcoded products of this PCR reaction were purified a second time using AMPure XP beads. Barcoded amplicons from all samples were pooled and sequenced on an Illumina MiSeq instrument at Colorado State University using an Illumina MiSeq v3 600-cycle Kit with 25% PhiX spike-in (Illumina).

Bioinformatics and sequence processing

EMU MinION and EMU MiSeq

Sequences generated on the MinION platform were base-called and demultiplexed using Guppy 6.0.1 (Oxford Nanopore Technologies, Oxford, UK). Sequences were filtered based on length (V34: 300–600 bp; Full: 1000–2000 bp) and a minimum q-score of 70 using Filtlong 0.2.1 (Wick, 2017) and Cutadapt 3.2 (Martin, 2011). Chimeras were filtered using vsearch (Rognes et al., 2016), and taxonomy was assigned with minimap2 2.22 (Li, 2018). Error-correcting was done with EMU (Curry et al., 2021) which applies an expectation minimization algorithm to adjust taxonomic assignments using up to 50 sequence alignments per sequence read.

Paired forward and reverse MiSeq reads were join using PEAR (Zhang et al., 2014). Sequences were then filtered based on length (V34: 300–600 bp) and a minimum quality score of 70 using Filtlong 0.2.1 (Wick, 2017) and Cutadapt 3.2 (Martin, 2011). Chimeras were filtered using vsearch UCHIME (Rognes et al., 2016), taxonomy was assigned with minimap2 and error-corrected with EMU (Curry et al., 2021).

DADA2 MiSeq

The MiSeq V34 library was analyzed using DADA2 bioinformatics pipeline (Callahan et al., 2016). Briefly, all primers were removed from demultiplexed raw fastq files using Cutadapt 3.2 (Martin, 2011) and amplicon sequence variants were inferred using the default pipeline in DADA2. Each sequence variant was classified to the EMU reference database using minimap2 2.22 (Li, 2018) and the primary alignment for each sequence was chosen with SAMtools 1.9 (Li et al., 2009) and used for taxonomic assignments.

Data analysis

To calculate the percent similarity of mock community sequences, we used three reference genomes available from Zymobionics for *Enterococcus faecalis*, *Lactobacillus fermentum*, and *Salmonella enterica* (https://www.zymoresearch.com/collections/zymobiomics-microbial-community-standards/products/zymobiomics-gut-microbiome-standard). Each sequence assigned to one of the three genera from all mock communities were BLASTed using the NcbiblastnCommandline command from the Biopython package in python (Cock et al., 2009). Results with no hit found were ignored.

Total library sizes are processing were as follows: MinION Full 1,695,436 total sequence reads with an average of 66,843 reads per sample; MinION V34 2,318,235 total sequence reads with an average of 96,730 reads per sample; and MiSeq V34 2,111,798 total sequence reads with an average of 83,345 reads per sample. Therefore, prior to diversity estimates all samples were rarefied to 50,000 reads. Principal Coordinates Analysis (PCoA) was performed using Bray-Curtis distances on Hellinger transformed data and significant differences between platforms and/or sites were tested using adonis in the vegan package for R (Oksanen et al., 2018). Differential abundances were tested using either the DESeq2 package or Wilcoxon test in the metacodeR package (Foster et al., 2017).

Results

Sequencing of the ZymoBIOMICS mock community standard for the MinION Full, MinION V34, and MiSeq V34 libraries yielded 55,747, 116,705, and 91,138 high-quality reads. Sequences from these libraries were classified to 18 90, 48, and 8 genera or 75, 284, 145, and 8 different species for the MinION Full, MinION V34, MiSeq V34, and MiSeq V34 DADA2 pipelines, respectively with all eight of the expected ZymoBIOMICS bacterial species present in each sample. Despite the extraneous taxa observed in the MinION and MiSeq EMU pipelines, the highest similarity to the mock community at the species level was observed for the MinION Full (0.852), followed by the MiSeq V34 DADA2 (0.809), MiSeq V34 (0.744), and MinION V34 (0.736) (Fig. 1).

For the soil samples from two different agricultural sites (ARDEC, Colorado and Pendleton, Oregon), all methods except MinION V34 detected significant differences in species richness estimates, however, methods varied greatly in their richness estimates (Fig. 2A). For example, the MinION Full method estimated species richness of 762 for ARDEC and 909 for Pendleton with a p-value = 0.002. For the MinION V34, species richness estimates were 1,276 for ARDEC and 1,228 for Pendleton with a p-value = 0.748. The MiSeq V34 method resulted in 887 species for ARDEC and 1,072 species for Pendleton (p-value = 0.001). Applying the DADA2 pipeline to this same library greatly reduced species richness; 248 species at ARDEC and 307 species at Pendleton (p-value = 0.047). Based on the inflated species richness in the non-DADA2 pipelines (> 8 species in the mock community and up to 5-fold greater richness in the soil samples), we tested the effect of three different filtering methods on alpha diversity estimates in the non-DADA2 pipelines. The first method was to (i) remove all species below a user-specified relative abundance threshold, the second method employed a permutation-based strategy (PERFect R package, Smirnova et al. 2018) using either (ii) all samples in each library (i.e., soils and mock communities) or (iii) only the soil samples in each library. Iterative testing showed that a relative abundance threshold of 0.07% and 1% was necessary to achieve richness estimates that were closest to the MiSeq DADA2 levels for the soil and mock communities, respectively (Figure S1). The threshold (0.07%) filtering approach resulted in similar species richness estimates between the filtered and non-filtered DADA2 pipelines with significant differences between

sites (Fig. 2B); MinION Full: $p < 0.001$ (mean = 277 ARDEC; 331 Pendleton), MinION V34: $p = 0.003$ (266 ARDEC; 292 Pendleton), MiSeq V34: $p < 0.001$ (244 ARDEC; 270 Pendleton), and MiSeq V34 DADA2: $p = 0.047$ (247 ARDEC; 306 Pendleton). Both permutation methods significantly reduced alpha diversity measurements but still resulted in greater richness estimates than the unfiltered MiSeq V34 DADA2 pipeline (Fig. 2C,D); furthermore, significant site differences ($p \leq 0.05$) were detected with the MiSeq but not the MinION pipelines.

Species-level community composition was significantly different between sequencing methods ($F = 26.1$, $p = 0.001$) and sites ($F = 26.1$, $p = 0.001$) based on a perMANOVA and visualized by PCoA (Fig. 3A) for the unfiltered data. The sites separated along Axis 1 (32.9%) and platforms (MinION vs MiSeq) separated along Axis 2 (28.5%). A biplot of phyla relative abundances showed that the MiSeq platform was enriched for Actinobacteria as compared to the MinION platform (Fig. 3A). Filtering had little impact on these patterns and perMANOVA showed that method and site differences ($p = 0.001$) were maintained (Fig. 3B-D). Biplots showed that phyla abundances showed similar patterns across all filtering methods with Actinobacteria positively correlated and Acidobacteria, Bacteroidetes, and Proteobacteria negative correlated with Axis 2 ($r^2 > 0.5$). These patterns were confirmed by differential abundance analysis (DESeq2) which showed that the MiSeq platform was enriched for Actinobactera, Chloroflexi, and Gemmatimonadetes; whereas, the MinION platform was enriched for Acidobacteria, Bacteroidetes, Firmicutes, Proteobacteria, and Verrucomicrobia regardless of the bioinformatics pipeline used (Fig. 4), filtering had little effect on the phyla differentially abundant between platforms (data not shown). Differential abundance between sites was also tested using the non-parametric Wilcox test at all taxonomic levels from Kingdom to Family (Figures S2-S4). Similar patterns were observed to the phyla level DESeq2 (e.g., Actinobacteia, Gemmatimonadetes, and Chloroflexi were all enriched with the MiSeq platform. However, at finer taxonomic levels this phyla-level bias was not always consistent. For example, when comparing the MiNION full to the MiSeq V34 DADA2 method, six sub-taxa of the Actinobacteria were lower with the MiSeq platform; consistent trends were also not seen with the taxa of the a, d, and g-Proteobacteria (Figure S2).

At the plot level, all methods detected significant site and plot differences (Minion Full: $F_{Site} = 134$ [$p = 0.001$], $F_{ARDEC} = 2.54$ [$p = 0.001$], $F_{Pend} = 5.14$ [$p = 0.001$], MinION V34: $F_{Site} = 109$ [$p = 0.001$], $F_{ARDEC} = 2.20$ [$p = 0.002$], $F_{Pend} = 3.67$ [$p = 0.001$], MiSeq V34: $F_{Site} = 151$ [$p = 0.001$], $F_{ARDEC} = 2.90$ [$p = 0.001$], $F_{Pend} = 6.88$ [$p = 0.001$], MiSeq V34 DADA2: $F_{Site} = 77.5$ [$p = 0.001$], $F_{ARDEC} = 1.78$ [0.002], $F_{Pend} = 4.50$ [$p = 0.001$]). Interestingly some of the patterns within a site differed; for example, at Pendleton, plot AW-2 is the most different for the MinION Full pipeline, whereas AW-3 is the most different for the MiSeq platform (Fig. 5). However, very little variation is explained by the second axis (2.9–4.7%).

Ideally, differential abundances of Phyla between the two soils should show the same magnitude and direction of change for all methods. However, this was not always the case within this study (Fig. 6). For instance, DESeq with unfiltered MinION Full data indicated that Acidobacteria was significantly more abundant in ARDEC, whereas MiSeq V34 DADA indicated Acidobacteria was higher in Pendleton, but both MinION V34 and MiSeq V34 showed no significant difference between sites (Fig. 6). After removing species with less than 0.07% relative abundance, the differential abundance of Acidobacteria was similar between MiSeq V34 and MinION Full, but MinION V34 still indicated no significant difference (Fig. 6B). With some exceptions (e.g., Acidobacteria, Firmicutes, Planctomycetes, and Proteobacteria), all methods agreed on the direction of significant differences of phyla between sites. A more detailed analysis of differential abundance at all taxonomic levels was perfomed using the non-parametric Wilcox test of log2 median fold-changes, and results for the Kingdom through Family taxonomic levels were visualized with the metacodeR package (Figures S5-S8). At all taxonomic levels, the full-length

MinION method had the highest similarity to the short MiSeq method with DADA2 correction; 73.2%, 69.3%, 74.1%, 79.3%, 79.4%, and 82.28% of the taxa at the phyla, class, order, family, genus, and species levels, respectively, showed a similar pattern in differences between the two sites (Fig. 7) with a small fraction of taxa exhibiting a mismatch between the two methods (i.e., significantly higher at opposite sites for each method).

Discussion

Illumina MiSeq and the Oxford Nanopore Technologies MinION sequencers have unique molecular methods for determining the sequence of DNA. Although MinION sequencing is a third-generation method with strong application in assessing microbial communities, it lacks the well-established bioinformatic methodology associated with second-generation MiSeq sequencing (Almeida et al., 2018; Straub et al., 2020). A previous comparison of dust microbial communities using these two sequencers suggested that the was generally good agreement between the two methods, with differences visible mainly at the genus and species taxonomic levels (Nygaard et al., 2020). Up until now, comparisons (Nygaard et al., 2020; Winand et al., 2020) of these two methods have not included mock communities along with complex agricultural soils, nor have they examined primer biases and different bioinformatics. Our study improves on these previous studies by including a mock community, using similar primers on both platforms (V34 primers), and similar bioinformatics pipelines. Since the MiSeq platform with DADA2 error-correction is considered to be the current gold standard for estimating microbial community diversity (Straub et al., 2020), we similarly used this method as our standard for comparison of MinION-generated data with full length and V3-V4 16S rDNA as well as MiSeq data analyzed with EMU.

The two platforms produce similar results with the low diversity, eight bacterial species mock community despite differences in the library preparation which were previously optimized for each platform. The MiSeq DADA2 pipeline resulted in no extraneous species; however, the MinION full-length method resulted in the closest similarity to the expected community composition. However, soil ecosystems have much a much higher complexity of bacterial community composition with more than 40 Phyla represented in a community versus the two Phyla in the mock community. This complexity was captured to different degrees by the various methods tested here and some biases between methods were observed. In general, regardless of the platform, sequencing error-corrections algorithms or filtering methods appear to be necessary to remove extraneous DNA sequences and correct for over-estimates of alpha-diversity; the MiSeq with DADA2 correction consistently results in the lowest estimates of diversity in soils.

Because sequencing methods produce experimental artifacts and inflate richness, we evaluated various filtering methods to remove potentially spurious taxa (Bokulich et al., 2013; Straub et al., 2020). We tested both user-defined (relative abundance threshold) and permutation filtering methods to remove sequencing errors and contaminants from soil samples. Assuming that the MiSeq DADA2 is the best estimate of soil bacteria richness, as it was for the mock community, only the relative abundance threshold method could result in similar species richness estimates for the other three methods. However, the required threshold appears to be dependent upon sample complexity (i.e., different thresholds for the soil and mock communities) and requires a user-defined threshold. In this study we were able to iteratively define the relative abundance threshold (0.07%) that resulted in similar estimates between methods; however, this will not always be possible because it is not feasible to sequence on both platforms for all future studies. Furthermore, this threshold will likely be dependent upon sequencing depth and in our case represents a much greater filtering threshold than just singletons or

doubletons. For example, the 0.07% threshold is more than 47, 68, and 58 sequence reads for the MinION full-length (66,843 average reads per sample), MinION V34 (96,730 average reads per sample), and MiSeq V34 (83,345 average reads per sample) methods. The impact of the permutation-based PERFect filtering method varied by method and complexity of samples in the sequencing library. For example, permutation filtering in the MinION full-length was closest to the MiSeq DADA2 pipeline only when the controls (ZymoBIOMICS mock community and H_2O controls) were included in the analysis. Diversity estimates for the MinION V34 and MiSeq V34 methods were more than three- and two-fold greater than the MiSeq V34 DADA2 pipeline regardless of the samples used in the permutation-based method

One of the main goals of microbial community sequencing efforts is to also determine estimates of the abundance of specific taxa in the community. These abundances are frequently used as a proxy of microbial processes and soil functions (Douglas et al., 2020). Due to the incomplete sampling of soil and the need to control for sequencing depths, comparisons are most frequently made using relative abundances or with normalized abundances. The relative abundance of phyla resulting from our various methods reveal that there are inherent biases in the sequencing platforms that can be seen at all taxonomic levels. For instance, MiSeq, regardless of bioinformatics method, tends to have a higher estimation of Actinobacteria and Bacteroidetes. (Browne et al., 2020) showed that both high and low GC contents can have a negative bias in the MiSeq platform; however, as they suggest this is usually a problem in metagenomic sequencing and not rRNA sequencing where rRNA GC contents tends to fall within the optimal range (~ 50%). A quick analysis of the GC contents in the full-length rRNA reference database (Curry et al., 2021) used here reveals an overall Phylum-specific mean of 55% with a minimum of 48% (Tenericutes) and maximum of 62% (Candidatus Bipolaricaulota). Early PCR termination is possible during amplification of GC-rich regions of the rRNA gene during library preparation (Laursen et al., 2017). However, all methods used here are reliant upon PCR amplification for library preparation and unless new biases are introduced during the MiSeq sequencing step (i.e., sequencing by synthesis) we suggest GC-biases are not likely to be the main driver of the platform differences. Furthermore, we did not see any systematic negative biases related to the GC content of ZymoBIOMICS mock community which was specifically designed to have a range of GC contents. For example, three of the species with > 50% GC content (*Salmonella enterica, Limosilactobacillus fermentum*, and *Pseudomonas aeruginosa*) all had higher, not lower, than expected frequencies with the MiSeq V34 DADA2 pipeline.

Another frequent use of microbial community data is to compare differences between locations and/or treatments for indicator taxa or changes in relative abundance. Ideally even if taxonomic abundances are biased, these biases would not interfere with the ability to identify relative differences between the treatments. Our results indicate that differential abundance trends were mostly consistent across sequencing methods with and without filtering, however exceptions were observed at all taxonomic levels examined. Site-level differences revealed that at least 70% all taxa showed similar differences between sites regardless of the method used. In previous studies it has been shown that differential abundance analyses are sensitive to sparsity (i.e., prevalence of samples with zero abundance) (Thorsen et al., 2016) and do not always limit the detection of false-positives (Hawinkel et al., 2019). Furthermore, relative abundance differences are dependent upon microbial load or the total population present in a sample (Morton et al., 2019). In this study, we compared sequencing analyses obtained from the same DNA extracts, so the different results should not arise due to different microbial loads.

Future Directions

Microbial references standards are severely lacking although a new fecal reference standard has recently been released commercially (ZymoBIOMICS Fecal Reference with TruMatrix™ Technology, Catalog # D6323). Such materials that have been highly characterized by multiple methodologies will greatly influence our ability to better estimate the biases between methodologies. A soil reference material, while greatly needed, is especially problematic due to the high microbial complexity, spatial and temporal heterogeneity, and differences in community structure associated with soil types, management, and location (Fierer 2017). However, these materials will be crucial in our ability to conduct meta-analyses in a field where technological changes are occurring at a rapid pace. The use of reference materials may also be helpful to use as standard reference bins (Morton et al. 2019) that allow for the comparison of taxonomic ratios that are comparable across studies and sequencing platforms.

Conclusion

The MiSeq and MinION sequencing platforms both appear adequate for the assessment of microbial community composition. However, there are trade-offs worth considering which platform to use for a study; MiSeq offers a more established bioinformatics pipeline, while the MinION is capable of producing longer reads which may offer better assessments for fungal communities. While the cost per sample to sequence with each platform in our study was not much different, the barrier-to-entry for new labs may be an incentive for procuring a MinION sequencer. Also, depending upon the diversity of the sample being studied, conflicting results for relative abundances and alpha- and beta-diversity may arise. Large differences in relative abundances of taxa between the sequencing and bioinformatics methods indicate we may need to be skeptical about relative abundance differences between studies, especially those with small trends. Overall, however, all methods were highly successful in identifying differences between sites and more than 70% of the taxa showed similar patterns in differential abundances. We suggest that additional studies are needed to identify if this variability is different than that would arise in multiple libraries generated within a single laboratory using consistent methods or between laboratories.

References

1. Almeida, A., Mitchell, A. L., Tarkowska, A., & Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience, *7*(5). https://doi.org/10.1093/gigascience/giy054
2. Bálint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., O'Hara, R. B., Öpik, M., Sogin, M. L., Unterseher, M., & Tedersoo, L. (2016). Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiology Reviews, *40*(5), 686–700.
3. Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature Methods, *10*(1), 57–59.
4. Browne, P. D., Nielsen, T. K., Kot, W., Aggerholm, A., Gilbert, M. T. P., Puetz, L., Rasmussen, M., Zervas, A., & Hansen, L. H. (2020). GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience, *9*(2). https://doi.org/10.1093/gigascience/giaa008
5. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, *13*(7), 581–583.

6. Ciuffreda, L., Rodríguez-Pérez, H., & Flores, C. (2021). Nanopore sequencing and its application to the study of microbial communities. Computational and Structural Biotechnology Journal, *19*, 1497–1511.
7. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, *25*(11), 1422–1423.
8. Curry, K. D., Wang, Q., Nute, M. G., Tyshaieva, A., Reeves, E., Soriano, S., Graeber, E., Finzer, P., Mendling, W., Wu, Q., Savidge, T., Villapol, S., Dilthey, A., & Treangen, T. J. (2021). *Emu: Species-Level Microbial Community Profiling for Full-Length Nanopore 16S Reads* (p. 2021.05.02.442339). bioRxiv. https://www.biorxiv.org/content/10.1101/2021.05.02.442339v1
9. Douglas, G. M., Maffei, V. J., Zaneveld, J., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., & Langille, M. G. I. (2020). PICRUSt2: An improved and customizable approach for metagenome inference. In *bioRxiv* (p. 672295). https://doi.org/10.1101/672295
10. Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. PeerJ, *5*, e3889.
11. Foster, Z. S. L., Sharpton, T. J., & Grünwald, N. J. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS Computational Biology, *13*(2), e1005404.
12. Fujiyoshi, S., Muto-Fujita, A., & Maruyama, F. (2020). Evaluation of PCR conditions for characterizing bacterial communities with full-length 16S rRNA genes using a portable nanopore sequencer. Scientific Reports, *10*(1), 12580.
13. Hawinkel, S., Mattiello, F., Bijnens, L., & Thas, O. (2019). A broken promise: microbiome differential abundance methods do not control the false discovery rate. Briefings in Bioinformatics, *20*(1), 210–221.
14. Kerkhof, L. J. (2021). Is Oxford Nanopore sequencing ready for analyzing complex microbiomes? FEMS Microbiology Ecology, *97*(3). https://doi.org/10.1093/femsec/fiab001
15. Laursen, M. F., Dalgaard, M. D., & Bahl, M. I. (2017). Genomic GC-Content Affects the Accuracy of 16S rRNA Gene Sequencing Based Microbial Profiling due to PCR Bias. *Frontiers in Microbiology*, *8*, 1934.
16. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, *34*(18), 3094–3100.
17. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, *25*(16), 2078–2079.
18. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, *17*(1), 10–12.
19. Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K., & Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. Nature Communications, *10*(1), 2719.
20. Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ, *6*, e5364.
21. Nygaard, A. B., Tunsjø, H. S., 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. Scientific Reports, *10*(1), 3209.

22. Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., & Stevens, M. H. H. (2018). *Package 'vegan'. Community ecology package, R package version 2.5-4.* https://cran.r-project.org/package=vegan.
23. Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, *4*, e2584.
24. Santos, A., van Aerle, R., Barrientos, L., & Martinez-Urtaza, J. (2020). Computational methods for 16S metabarcoding studies using Nanopore sequencing data. Computational and Structural Biotechnology Journal, *18*, 296–305.
25. Straub, D., Blackwell, N., Langarica-Fuentes, A., Peltzer, A., Nahnsen, S., & Kleindienst, S. (2020). Interpretations of Environmental Microbial Community Studies Are Biased by the Selected 16S rRNA (Gene) Amplicon Sequencing Pipeline. *Frontiers in Microbiology*, *0*. https://doi.org/10.3389/fmicb.2020.550420
26. Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., Sørensen, S., Bisgaard, H., & Waage, J. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome, *4*(1), 62.
27. Wesolowska-Andersen, A., Bahl, M. I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R., & Licht, T. R. (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome, *2*(1), 19.
28. Wick, R. (2017). *Filtlong*. https://github.com/rrwick/Filtlong
29. Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Van Braekel, J., Fu, Q., Roosens, N. H. C., De Keersmaecker, S. C. J., & Vanneste, K. (2020). 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. International Journal of Molecular Sciences, *21*(1), 298.
30. Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, *30*(5), 614–620.

Figures

Figure 1

Taxonomic classification at the genus (A) and species (B) level for mock community sequencing output from the four platform and bioinformatics pipelines (MinION Full, MinION v34, MiSeq v34, and MiSeq V34 DADA2). Taxa not contained in the ZymoBIOMICS standard were grouped into the 'Other' category. Similarity (Bray-Curtis) between the expected mock community output and the sample is displayed above each sample.

Figure 2

Species richness at two sites for all four sequencing methods and bioinformatics methods (MinION Full, MinION V34, MiSeq V34, and MiSeq V34 DADA2) for unfiltered (A), filtered at 0.07% relative abundance threshold (B)

filtered by permutation using all samples (C), and filtered by permutation using only soil samples (D). Data were rarefied to 50,000 reads per sample and asterisks indicate a significant difference between sites ($p \leq 0.05$). In all panels, the MinION V34 DADA2 method was calculated using unfiltered data only.

Figure 3

PCoA biplot of all four sequencing and bioinformatics methods (MinION Full, MinION V34, MiSeq V34, and MiSeq V34 DADA2) for unfiltered (A), filtered at 0.07% relative abundance threshold (B) filtered by permutation using all samples (C), and filtered by permutation using only soil samples (D). In all panels, the MinION V34 DADA2 method used unfiltered data only. Vectors indicate a significant correlation ($r^2 > 0.5$, $p < 0.01$) between phyla relative abundance and ordination axes.

Figure 4

Relative abundance of the 10 most abundant phyla for all four sequencing bioinformatics methods (MinION Full, MinION V34, MiSeq V34, and MiSeq V34 DADA2). Bars with different letters are significant at an adjusted $p \leq 0.05$ based on DESeq2 analysis with Benjamini-Hochberg correction. All data was unfiltered.

Figure 5

PCoA ordinations for each of the four sequencing and bioinformatics methods (MinION Full, MinION V34, MiSeq v34, and MiSeq DADA2). Shapes indicate different sites (ARDEC, Colorado and Pendleton, Oregon), while colors indicate separate plots within each site. Vectors indicate a significant correlation ($r^2 > 0.8$, $p < 0.01$) between phyla relative abundance and ordination axes.

Figure 6

Log2 fold-change results from differential abundance (DESeq2) for each of the four sequencing and bioinformatics methods (MinION Full, MinION V34, MiSeq V34, and MiSeq DADA2). Each panel is a different filtering method: unfiltered (A), filtered at 0.07% relative abundance threshold (B), filtered by permutation using all samples (C), and filtered by permutation using only soil samples (D). Teal colors indicate significantly higher relative abundances in Pendleton, Oregon, while red indicates significantly higher abundances in ARDEC, Colorado. The lack of a point indicates that the test was not significant (false-discovery rate > 0.05).

Figure 7

Percent of taxa significantly different between the two sites based on a Wilcox test or relative abundances. All methods are compared to the MiSeq V34 DADA2 pipeline where dark green bars (Both Sig.) are significantly different (FDR < 0.05) for both methods, light green bars (Both N.S.) are both not significantly different (FDR ≥ 0.05), grey bars (Mis-match) are significantly different (FDR < 0.05) but tests differ in enriched site, light red bars (Test Sig. only) are significantly different (FDR < 0.05) for only the test (i.e., listed) method, and dark red bars (DADA2 Sig. only) are significantly different (FDR < 0.05) for only the MiSeq V34 DADA2 pipeline.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementalmaterials.docx