Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets

Daniel M. Portik[1*], C. Titus Brown[2], and N. Tessa Pierce-Ward[2]

1. Pacific Biosciences, 1305 O'Brien Dr, Menlo Park, California 93025 USA

2. Department of Population Health and Reproduction, University of California Davis, Davis, California USA

*Correspondence: dportik@pacificbiosciences.com

ABSTRACT

Long-read shotgun metagenomic sequencing is gaining in popularity and offers many advantages over short-read sequencing. The higher information content in long reads is useful for a variety of metagenomics analyses, including taxonomic profiling. The development of long-read specific tools for taxonomic profiling is accelerating, yet there is a lack of consensus regarding their relative performance. Here, we perform a critical benchmarking study using five long-read methods and four popular short-read methods. We applied these tools to several mock community datasets generated using Pacific Biosciences (PacBio) HiFi or Oxford Nanopore Technology (ONT) sequencing, and evaluated their performance based on read utilization, detection metrics, and relative abundance estimates. Our results show that long-read methods generally outperformed short-read methods. Short-read methods (including Kraken2, Bracken, Centrifuge, and MetaPhlAn3) produced many false positives (particularly at lower abundances), required heavy filtering to achieve acceptable precision (at the cost of reduced recall), and produced inaccurate abundance estimates. By contrast, several long-read methods displayed very high precision and acceptable recall without any filtering required, including BugSeq, MEGAN-LR using translation alignments (DIAMOND to NCBI nr) or nucleotide alignments (minimap2 to NCBI nt). Furthermore, in the PacBio HiFi datasets these long-read methods detected all species down to the 0.1% abundance level with high precision. Other long-read methods, such as MetaMaps and MMseqs2, required moderate filtering to reduce false positives to achieve a suitable balance between precision and recall. We found read quality affected performance for methods relying on protein prediction or exact kmer matching, and these methods performed better with PacBio HiFi datasets. We also found that long-read datasets with a large proportion of shorter reads (<2kb length) resulted in lower precision and worse abundance estimates, relative to length-filtered datasets. Finally, for a given mock community we found that the long-read datasets produced significantly better results than short-read datasets, demonstrating clear advantages for long-read metagenomic sequencing. Our critical assessment of available methods provides recommendations for current research using long reads and establishes a baseline for future benchmarking studies.

INTRODUCTION

The identification of microbial species in environmental communities is an essential task in microbiology. Shotgun metagenomic sequencing (or metagenomics) can provide relatively unbiased sampling of the species in such communities, which can include bacteria, archaea, viruses, and eukaryotes. Whereas selective amplification (e.g., 16S, ITS) targets specific gene regions, the goal of metagenomics is to sequence complete genomic DNA for all species in a sample. Consequently, the set of tools used to predict the identities and relative abundances of microbial species differs greatly between these approaches. In particular, the difficulty of performing this task for complex shotgun sequencing data has led to the development of many taxonomic profiling methods, particularly for second-generation/short-read technologies (reviewed in Breitwieser et al. 2019). The rapid expansion of short-read profiling tools led to recognition of the importance of methods comparisons, benchmarking, and standardized test datasets (Lindgreen et al. 2016; McIntyre et al. 2017; Sczyrba et al. 2017; Escobar-Zepeda et al. 2018; Breitwieser et al. 2019; Meyer et al. 2019; Tamames et al. 2019; Ye et al. 2019; Parks et al. 2021). These benchmarking studies have been critical for understanding the relative performance of taxonomic profiling methods for different use-cases, which can vary greatly among microbiologists.

Though much of metagenomics has focused on short-read sequencing, there is rising awareness of the new opportunities offered by third-generation sequencing technologies which produce longer sequencing reads. Whereas short reads typically contain a single gene fragment, long reads often span multiple genes and intergenic regions which can be used for alignment algorithms and sequence matching. Among the most popular long-read sequencing platforms are those produced by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). While long reads have historically been accompanied by higher error rates, continual improvements in library preparation, sequencing chemistries and post-processing have dramatically reduced the error rates associated with longer reads. For example, the most recent combination of ONT "Q20" chemistry and the Bonito basecaller (v0.3.5+) is reported to produce modal read accuracies of 99% (~Q20), and the development of PacBio HiFi sequencing allows for highly accurate consensus reads (>Q20, median Q30) that are 10–20 kb in length (Wenger et al. 2019). As a result of these improvements, both PacBio HiFi and ONT long reads offer new

potential for metagenomic analyses, including metagenome assembly, functional annotation, and taxonomic profiling.

Until recently, few studies have evaluated the performance of taxonomic profiling methods for long reads, in part because few tailored methods were available. However, the rate of development for long-read taxonomic profiling methods appears to be increasing. For example, MetaMaps (Dilthey et al. 2019) and MEGAN-LR (Huson et al. 2018) were among the first long-read profiling methods, and they became available over the course of several years. By contrast, multiple methods have appeared in the beginning of 2021, including MMseqs2 taxonomy (Mirdita et al. 2021) and BugSeq (Fan et al. 2021). Prior long-read benchmarking studies applied short-read profiling methods to long reads (McIntyre et al. 2017; Leidenfrost et al. al 2020) or compared the potential of long and short-read profiling (Pearman et al. 2020), yet only one study has included a comparison of long-read methods (Marić et al. 2020). Given the dramatic decreases in long read error rates and the proliferation of long read profiling methods, there is a pressing need to assess the performance of taxonomic profiling using long reads.

Here, we perform a critical benchmarking study to evaluate the performance of taxonomic profiling methods for long-read datasets. We include five long-read methods and four popular short-read methods. The ideal taxonomic profiler should assign all assignable reads (e.g., those with database representation), display a low number of false positives, have high detection power, and accurately estimate the relative abundances of taxa (McIntyre et al. 2017; Sczyrba et al. 2017; Breitwieser et al. 2019; Tamames et al. 2019; Ye et al. 2019; Parks et al. 2021). We evaluate the relative performance of methods based on these criteria, using publicly available datasets. These datasets consist of mock communities of known compositions, which were generated using PacBio HiFi or ONT sequencing. Mock communities are considered simplistic relative to environmental samples, but they allow a clear assessment of detection metrics (such as precision, recall, and F-scores) and are therefore highly informative for benchmarking. In order to tease apart the impacts of error profile and read length on taxonomic profiler performance, we also include comparisons using Illumina short-read datasets for two of the mock communities. Our main goals are to 1) identify which methods perform best for long-read datasets, 2) understand if long reads provide better taxonomic profiling or abundance estimation results relative to short reads, and 3) identify if differences in long read quality have any effects on performance. Overall, we provide a baseline assessment of available methods using

reproducible analyses, which can inform current research and establish a foundation for future benchmarking studies.

MATERIALS AND METHODS

Mock Community Datasets

We obtained two PacBio HiFi datasets and two ONT datasets from publicly available sources. We chose empirical datasets versus simulated datasets because simulations do not capture true variation in error profiles, read length heterogeneity, and the effects of DNA extraction, library preparation, and sequencing. Furthermore, pseudo-mock communities (e.g., those created from multiple isolate sequencing datasets) may combine older and newer sequencing chemistries/platforms for a given technology, creating additional confounding effects.

The two PacBio datasets are available on NCBI (Table 1). The first PacBio HiFi dataset is for the ATCC MSA-1003 mock community (PRJNA546278: SRX6095783, released June 2019). The ATCC MSA-1003 mock community contains 20 bacteria species in staggered abundances (5 species at 18%, 1.8%, 0.18% and 0.02% abundance levels, respectively). The PacBio ATCC dataset was generated using the Sequel II System and contains 2.4 million HiFi reads with a median length of 8.3 kb, for a total of 20.54 Gb of data (Fig. 1, Table 1). We refer to this dataset as HiFi ATCC MSA-1003. The second PacBio HiFi dataset is for the ZymoBIOMICS Gut Microbiome Standard D6331 (PRJNA680590: SRX9569057, released November 2020). The Zymo D6331 mock community contains 17 species (including 14 bacteria, 1 archaea, and 2 yeasts) in staggered abundances. Five species occur at 14% abundance, four at 6%, four at 1.5%, and one species per 0.1%, 0.01%, 0.001%, and 0.0001% abundance level. There are five strains of *E. coli* contained in this community (each at 2.8% abundance), which we treat here as one species at 14% abundance. The PacBio Zymo D6331 dataset was generated using the Sequel II System and contains 1.9 million HiFi reads with a median length of 8.1 kb, for a total of 17.99 Gb of data (Fig. 1, Table 1). We refer to this dataset as HiFi Zymo D6331.

We obtained two ONT datasets for the ZymoBIOMICS D6300 microbial community standard. The Zymo D6300 standard is simpler in design and contains 10 species in even abundances, including 8 bacteria at 12% abundance and two yeasts at 2% abundance. The two ONT datasets contained a broader distribution of read lengths which included a large tail of

shorter reads (<2kb in length). Our initial work indicated these shorter reads may have an adverse effect on taxonomic profiling, a result also supported by Govender and Eyre (2022). We therefore included two variations of each ONT dataset. The primary datasets are the focus of our methods comparison and resulted from length filtering to remove all short reads (<2kb) and ultra-long reads (>50kb). We found ultra-long reads caused compatibility issues with some taxonomic profiling programs (particularly the short-read methods). To investigate the potential effects of shorter reads, we created secondary datasets which contained a large proportion of shorter long reads. The first ONT dataset comes from a continually updated resource produced by Nicholls et al. (2019). We downloaded the R10.3 chemistry data release (February 2020) which was produced from two flowcells on an ONT GridION, resulting in 1.16 million reads (4.64 Gb data). We used NanoFilt (De Coster et al. 2018) to remove all short (<2kb) and ultra-long reads (>50kb). Length-filtering resulted in the removal of 873,079 short reads and 12,129 ultra-long reads (1.33 Gb total; 75% and 0.01% of total reads, respectively), and the retention of 275,318 ONT reads (23% of total reads). The resulting length filtered ONT reads have a median length of 6.6 kb, for a total of 3.31 Gb of data (Fig. 1, Table 1). We refer to this primary dataset as ONT R10 Zymo D6300. The secondary version of this dataset uses all reads <50kb in length. It contains 3.86 Gb data (1,148,397 reads) with a median read length of 660 bp and mean read length of 3.3 kb, and is referred to as ONT R10 Short (Supplementary Figure S1). The second ONT dataset was obtained from the European Nucleotide Archive (PRJEB43406: ERR5396170, released March 2021) and represents the 'Q20 chemistry' release for the Zymo D6300 standard (described at: https://github.com/Kirk3gaard/2020-05-20_ZymoMock_Q20EA). It was generated using a PromethION, resulting in 5.4 million reads (17.95 Gb data). We again used NanoFilt to remove short reads (<2 kb) and ultra-long reads (>50 kb), which resulted in the elimination of 2.13 million (39%) and 819 (<0.001%) of the total reads, respectively. From the remaining ~3.2 million reads, we subsampled to obtain 2 million reads (a number comparable to the HiFi datasets). This produced a length filtered ONT dataset of 2 million reads with a median length of 4.2 kb, for a total of 9.6 Gb of data (Fig. 1, Table 1). We refer to this primary dataset as ONT Q20 Zymo D6300. The secondary version of this dataset contains a comparable number of shorter long reads. We used NanoFilt to remove all reads >3kb in length and subsampled the remaining reads to obtain 2 million reads. We refer to this as ONT Q20 Short, and this dataset

contains 2.72 Gb data with a median read length of 1.2 kb and mean read length of 1.3 kb (Supplementary Figure S1).

As a final comparison to the long-read datasets, we included short-read sequence data for two of the mock communities (Table 1). We downloaded Illumina sequence data for ATCC MSA-1003 (PRJNA510527: SRX5169925, released December 2018), which included a total of ~10 million 150 bp paired-end reads produced by a HiSeq2500 (but available pre-trimmed to 125 bp). We also obtained Illumina sequence data for the Zymo D6300 community (PRJNA648136: SRX8824472, released July 2020). These data were produced using a NovaSeq 6000 and include ~100 million 150bp PE reads. Given the large difference in read numbers between these datasets, we subsampled the Zymo Illumina data to obtain 20 million total reads. We refer to these datasets as Illumina ATCC MSA-1003 and Illumina Zymo D6300, respectively. A variety of factors, including different DNA extraction methods, can affect the final composition of DNA sequenced for metagenomic samples and potentially bias relative abundance estimates (Sui et al. 2020). Additionally, variation in error profiles across sequencing technologies could also cause potential differences in results. To control for these potential confounding effects in the Illumina datasets, we also "simulated" short-read data from our long-read datasets. Each long read was divided into 150 bp non-overlapping segments, and 10 segments were randomly selected to create a "simulated" short-read dataset. We chose this subsampling strategy (versus retaining all available segments) to create a consistent number of short reads per long read, which varied in length. This strategy generated ~21 million 150 bp "reads" from the HiFi ATCC MSA-1003 dataset, and 20,000,000 150 bp "reads" from the ONT Q20 Zymo D6300 dataset. We refer to these datasets as SR-Sim ATCC MSA-1003 and SR-Sim ZymoD6300, respectively.

Taxonomic Profiling Methods

We evaluated the performance of nine taxonomic profiling methods on the long-read mock community datasets. We included five profiling methods developed specifically for long reads, as well as four popular short-read methods (Table 2), which we summarize here. We ran all methods for the primary long-read datasets and secondary ONT datasets, and used only short-read methods for the short-read datasets.

The short-read methods include Kraken2 (Wood & Salzberg 2014; Wood et al. 2019), Bracken (Lu et al. 2017), Centrifuge (Kim et al. 2016), and MetaPhlAn3 (Beghini et al. 2021). Kraken2 is a kmer-based approach, which is often paired with Bracken. Following Kraken2 analyses, Bracken can be used for Bayesian re-estimation and refinement of the results. Centrifuge uses a Burrows-Wheeler transform and Ferragina-Manzini index for storing and mapping sequences. We include two variations of Centrifuge analyses, one using the default settings suitable for short reads (referred to as Centrifuge-h22), and another with settings for long reads (referred to as Centrifuge-h500; see details below). MetaPhlAn3 uses coverage scores to calculate the relative abundances of taxa, based on read mapping to a clade-specific marker database.

The long-read methods include MetaMaps (Dilthey et al. 2019), MEGAN-LR (Huson et al. 2016, 2018), MMseqs2 (Mirdita et al. 2021), and BugSeq (Fan et al. 2021). MetaMaps was among the first methods designed specifically for long reads, and it uses approximate mapping with probabilistic scoring to estimate sample composition. MEGAN-LR was developed from MEGAN6 and was designed to interpret translation alignments of long nucleotide sequences to a protein reference database. These alignments can be made using any program capable of translation alignment (e.g., blastx mode), but here we specifically use DIAMOND (Buchfink et al. 2015) due to its favorable long-read options (e.g., range-culling and frameshift-aware alignment; Arumugam et al. 2019). MEGAN-LR assigns reads to taxa using a lowest common ancestor (LCA) algorithm, in combination with other relevant features (e.g., lcaCoveragePercent, minSupportPercent, minPercentReadCover). MEGAN-LR can likewise interpret alignments to nucleotide databases using similar options, such as those created with minimap2 (Li 2018). For this experiment, we created alignments based on protein references (using DIAMOND) and nucleotide references (using minimap2), and subsequently used MEGAN-LR for taxonomic classification. To distinguish between these methods, we refer to them as MEGAN-LR-prot and MEGAN-LR-nuc. Furthermore, we tested settings in minimap2 that were specific to HiFi or ONT data (see below) and ran both settings on all mock communities. We refer to these analyses as MEGAN-LR-nuc-HiFi and MEGAN-LR-nuc-ONT. Thus, we include three analyses that involve MEGAN-LR: MEGAN-LR-prot, MEGAN-LR-nuc-HiFi, and MEGAN-LR-nuc-ONT. The MMseqs2 taxonomy tool extracts all possible protein fragments in six frames from the long reads, pre-filters the protein sequences, aligns the retained protein sequences to the reference

protein database, and ultimately assigns reads to taxa using a novel LCA algorithm ("approximate 2bLCA"). The published BugSeq algorithm (V1) performs minimap2 alignments using a nucleotide database, followed by Bayesian reassignment and LCA identification (Fan et al. 2021). Following initial development, a BugSeq V2 method was developed which includes minimap2 alignment of sequences to a nucleotide database followed by LCA identification and abundance calculation (S. Chorlton, personal communication). BugSeq V2 performs better for longer reads (>1kb), higher sequencing depth, and shotgun metagenomics (vs. cDNA sequencing experiments). An auto-detect feature selects the V1 or V2 version based on the dataset uploaded to the online platform, and in our experiment BugSeq V2 was selected for all long-read datasets.

A standardized output format was required to facilitate comparisons of the results across methods. We selected kraken-report (kreport) format because it contains cumulative counts and level counts across the complete hierarchical taxonomy for each taxon assigned. The level count is the number of reads specifically assigned to a taxon, whereas the cumulative count is the sum of the level counts for a taxon plus its descendants. For example, the cumulative count of a genus is the level count for that genus plus the level counts of all species and strains contained in that genus. This output format is readily available for Kraken2, Bracken, MMseqs2, and BugSeq. We created conversion tools for all other methods (MetaPhlAn3, MetaMaps, MEGAN-LR), which are available on github: https://github.com/PacificBiosciences/pb-metagenomics-tools.

Comparative Analyses

We evaluated method performance using several criteria. We assessed read utilization, detection metrics at the species and genus level, and relative abundance estimates. We provide details for each of these categories below.

Read Utilization. We evaluated read utilization for each profiling method in two ways. First, we simply calculated the total percent of reads that received a taxonomic assignment. Second, we calculated the percentage of reads that were assigned to specific taxonomic levels. We performed this for the following ranks: class, order, family, genus, species, and subspecies/strain. Values were obtained by summing the level counts of all taxa within a given rank. In general, we expected methods that utilize LCA algorithms to display read assignments across multiple taxonomic levels, relative to methods that do not.

Detection Metrics. The species compositions of the mock communities are known, allowing for a complete evaluation of detection metrics. For each profiling method, we scored the presence/absence of a taxon based on whether or not the cumulative read count for that taxon exceeded a minimum percent threshold of the total reads. We used a minimum percent threshold (versus a fixed number of reads) because our datasets contained different numbers of total reads. We recognize that setting a minimum detection threshold in this way penalizes methods that assign a smaller proportion of the total reads available. However, setting a threshold based on the number of reads assigned in a given analysis could produce misleading results (for example, a method could assign only 10% of total reads but achieve perfect precision). We evaluated three minimum read thresholds, including 0.001% (mild filtering, mainly for removing singleton count taxa for short-read methods), 0.1% (moderate filtering), and 1% (heavy filtering) of the total number of reads per dataset (Table 3). The threshold filtering was mainly used to explore the effects on precision (particularly the impact on false positives) across the four primary datasets. However, we also used filtering to investigate the effects on the staggered abundance communities (ATCC MSA-1003 and Zymo D6331). These two mock communities contained several taxa in low abundances, and we explored how filtering might cause detection dropout for different abundance levels. We performed our evaluations at the species level and the genus level. We expected detection to be more difficult at the species level and easier at the genus level. This is because assignments to multiple non-target species within a genus would be considered incorrect at the species level, but correct at the genus level.

We calculated several detection metrics (precision, recall, F-scores) which are based on the number of true positives, false positives, and false negatives. In this context, we define a true positive as the detection of a mock community taxon (based on a read count exceeding the minimum read threshold). We define a false positive as the detection of taxon that is not present in the mock community. We define a false negative as the failure to detect a taxon in the mock community (based on a zero count or count below the minimum threshold). The formulas for precision, recall and F-scores are as follows:

Precision = true positives / (true positives + false positives)

Recall = true positives / (true positives + false negatives)

$\mathbf{F_1}$ = (2 * precision * recall) / (precision + recall)

$$\mathbf{F_{0.5}} = ((1 + 0.5^2) * \text{precision} * \text{recall}) / ((0.5^2 * \text{precision}) + \text{recall})$$

The values for the above metrics each range from 0 to 1. For precision, a score of 1 indicates only mock community taxa were detected, whereas lower scores indicate detection of additional taxa (e.g., false positives). For recall, a score of 1 indicates all taxa in the mock community were detected, whereas a lower score indicates some taxa were not detected. The F-scores provide a useful way to summarize the information from precision and recall. The F_1 score is the harmonic mean of precision and recall (both measures are weighted equally), whereas the $F_{0.5}$ score gives more weight to precision (placing more importance on minimizing false positives). A value of 1 for either F-score indicates perfect precision and recall.

We controlled for two issues that can negatively impact these metrics. First, we observed and accounted for differences in taxonomy, particularly as it relates to synonymies. In the case of a species synonomy, we used the sum of cumulative counts for the species and all synonyms as the read count for the taxon. This included two species in ATCC MSA-1003 (*Luteovulum sphaeroides = Rhodobacter sphaeroides, Cereibacter sphaeroides; Phocaeicola vulgatus = Bacteroides vulgatus*), one species in Zymo D6300 (*Limosilactobacillus fermentum = Lactobacillus fermentum*), and two species in Zymo D6331 (*Limosilactobacillus fermentum = Lactobacillus fermentum; Bacillus subtilis = Bacillus spizizenii*). Second, we observed that sequences and/or taxonomy information was lacking for two species (Zymo D6331: *Veillonella rogosae, Prevotella corporis*) in multiple databases ("PlusPF", Refseq ABVF, MiniSeq+H, NCBI nt). To remedy this issue, we excluded the two species from the set of taxa used to calculate detection metrics at the species-level for all methods. However, we observed that many reads were assigned to alternate species in the same genus, so we included the two genera in the genus-level analysis.

Relative abundance estimates. Relative abundances were estimated for each profiling method at the species and genus level. We obtained cumulative counts for the mock community species or genera and the sum of cumulative counts for all false positives at the species or genus level (classified as "Other"). These data were normalized to obtain the percent abundance of each taxon. We corrected for the absence of two species from multiple databases (*Veillonella rogosae, Prevotella corporis*) in HiFi Zymo D633. For methods affected by these databases, we observed many reads were assigned to other species in these two genera. Rather than scoring these as

"Other", we allowed all species-level assignments within these genera to contribute to the read counts for these two species. To be consistent, we allowed this for all methods for HiFi Zymo D633. In other words, genus-level counts for *Veillonella rogosae* and *Prevotella corporis* were used for the species abundances, rather than exclude these two taxa. For each method, we performed a chi-squared goodness of fit test to determine if the estimated abundances were significantly different from the theoretical abundances. We compared the chi-squared statistic to the critical value obtained at the 95% significance level and obtained a corresponding P-value. For this test, larger chi-squared statistic values indicate greater differences between the observed and expected values. We applied a Bonferroni correction for multiple testing (n = 11) per dataset, for which $\alpha_{altered} = 0.05/11 = 0.0045$. A P-value < 0.0045 allows rejection of the null hypothesis, and indicates the observed distribution is significantly different from the theoretical distribution.

Reference Databases

The choice of reference database directly affects the outcome of taxonomic profiling. For example, the use of a complete reference database versus a subset of that database can result in drastically different assignments if the same profiling method is run with otherwise identical settings. Under ideal conditions, all profiling methods would use an identical reference database. This would control for differences in information content and taxonomy, allowing observed differences in assignment results to be attributed to the profiling methods. However, differences in method design and matching algorithms required the use of multiple reference databases. We therefore provide a brief description and comparison of these databases below.

The databases used for Kraken2, Bracken, and Centrifuge are highly similar. For Kraken2 and Bracken, we used a pre-built database that includes all RefSeq sequences for archaea, bacteria, viruses, plasmid, human, protozoa, and fungi ("PlusPF", released 1/27/2021, available from: https://benlangmead.github.io/aws-indexes/k2). The Centrifuge database was built from RefSeq sequences for archaea, bacteria, viruses, and fungi (downloaded 4/2021). The Centrifuge database used can be considered a subset of the PlusPF database, but with complete overlap for several target groups (archaea, bacteria, fungi).

MetaPhlAn3 uses a highly distinct reference database which is composed of ~1.1 million unique clade-specific markers from ~99,500 bacteria/archaea reference genomes and ~500 eukaryotic reference genomes. We used the mpa_v30_CHOCOPhlAn_201901 database release.

MetaMaps provides a pre-built database composed of 12,058 complete RefSeq genomes (215 archaeal, 5774 bacterial, 6059 viral/viroidal, 7 fungi, 1 human), which is referred to as MiniSeq+H. The option to create a custom database (such as NCBI nt) was initially developed for MetaMaps, but this feature is currently not functional. The MiniSeq+H database was therefore the only option available for running MetaMaps in our experiment, and it represents the smallest and most incomplete database across the methods used.

We used the NCBI non-redundant protein database (NCBI nr) for MMseqs2 and MEGAN-LR-prot, and the NCBI nucleotide database (NCBI nt) for MEGAN-LR-nuc and BugSeq v2 (both databases downloaded April 2021). These represent the most complete reference databases across the methods. We note that the RefSeq databases for Kraken2, Bracken, and Centrifuge are contained in NCBI nt.

Profiling Method Commands

To facilitate reproducible results, we provide the general commands or instructions to run each method.

***Kraken2*.** We ran Kraken version 2.1.1 for each sample. We used the pre-built PlusPF database described above, and used the following command:

```
kraken2 --db PlusPF --threads 24 –report SAMPLE.kreport.txt
SAMPLE.fasta > SAMPLE.kraken
```

***Bracken*.** We ran Bracken version 2.6.0 for each sample, using the kreport outputs from Kraken2. We used the pre-built PlusPF database described above, and the following command to obtain abundances at the species level (-l S):

```
bracken -d PlusPF -i SAMPLE.kreport.txt -o SAMPLE.bracken -r 50
-l S -t 10
```

***Centrifuge*.** We ran Centrifuge version 1.0.4. We were unable to use centrifuge-download to obtain the RefSeq sequences required to build the database. We instead used kraken2-build to obtain the relevant RefSeq sequences and taxonomy files. The kraken headers were removed from the fasta sequences, and the database was built using the following command:

```
centrifuge-build -p 24 --conversion-table centrifuge-seqid2taxid.map --taxonomy-tree /taxonomy/nodes.dmp --name-table /taxonomy/names.dmp arc-bac-vir-fungi.fna abvf
```

Centrifuge offers the option to specify the minimum length of partial hits required for classification (--min-hitlen). We used two values for this option. We used the default value of 22, which is suitable for short read analysis, and used a value of 500 which is suitable for long reads (labeled as Centrifuge-h22 and Centrifuge-h500, respectively).

We ran Centrifuge-h22 for each sample using the following command:

```
centrifuge -f --min-hitlen 22 -k 20 -t -p 24 -x abvf -U SAMPLE.fasta -S SAMPLE-h22.txt --report-file SAMPLE-h22.centrifuge_report.tsv
```

We ran Centrifuge-h500 for each sample using the following command:

```
centrifuge -f --min-hitlen 500 -k 20 -t -p 24 -x abvf -U SAMPLE.fasta -S SAMPLE-h500.txt --report-file SAMPLE-h500.centrifuge_report.tsv
```

Outputs were converted to kreport format using the centrifuge-kreport module.

MetaPhlAn3. Analyses were run using MetaPhlAn v3.0.7. The settings used in MetaPhlAn3 to run Bowtie2 will fail for long reads, so we first created alignments externally using Bowtie2:

```
bowtie2 -p 12 -f --local --no-head --no-sq --no-unal -S SAMPLE.sam -x /metaphlan/mpa_v30_CHOCOPhlAn_201901 -U SAMPLE.fasta
```

After alignments were created, we ran MetaPhlAn3 with the following settings (adjusting the number of reads per dataset, --neads):

```
metaphlan SAMPLE.sam --nproc 24 --input_type sam --nreads
2419037 -o SAMPLE.profiled_metagenome.txt --index
mpa_v30_CHOCOPhlAn_201901 --bowtie2db /metaphlan
```

MetaMaps. We used MetaMaps v0.1 to run analyses with the following set of commands:

```
metamaps mapDirectly --all -r /databases/miniSeq-H/DB.fa -q
SAMPLE.fasta --maxmemory 35 -t 24 -o SAMPLE_results

metamaps classify -t 12 --mappings SAMPLE_results --DB
/databases/miniSeq-H
```

The conversion from MetaMaps output format to kreport format was performed at the species level, but we note that MetaMaps can produce a large number of strain assignments that are not represented in our results.

MMseqs2. We used MMseqs2 v13.45111 to run all analyses. We first built the database for NCBI nr using the following command:

```
mmseqs databases NR /mmseqs-database/NR_db /scratch --threads 24
```

We then used the easy-taxonomy module to run analyses for each sample, using the following general command:

```
mmseqs easy-taxonomy SAMPLE.fasta /mmseqs-database/NR_db SAMPLE
/scratch --threads 48 --split-memory-limit 120G
```

MEGAN-LR-prot. A streamlined workflow for MEGAN-LR-prot is available as two pipelines (Taxonomic-Functional-Profiling-Protein, MEGAN-RMA-Summary) at: https://github.com/PacificBiosciences/pb-metagenomics-tools. These pipelines are provided as configurable snakemake workflows. To use the workflows, we first downloaded the NCBI nr database and created a DIAMOND index using the following command:

```
diamond makedb --in nr.gz --db diamond_nr_db --threads 24
```

We downloaded MEGAN6 community edition to obtain the executable tools required for these workflows (sam2rma, rma2info), as well as the required MEGAN protein mapping file (megan-map-Jan2021.db). We then ran the Taxonomic-Functional-Profiling-Protein pipeline. The locations of the nr index, sam2rma, and the mapping file were specified in the main configuration file for the analysis (config.yaml), and we used all other default settings (see documentation). The information for the sample fasta files was added to the sample configuration file (Sample-Config.yaml), and the snakemake (Snakefile-taxprot) was executed. Details for the usage of each program are provided in the online documentation.

Analyses resulted in RMA output files, which were used as inputs for the MEGAN-RMA-Summary pipeline. The location of rma2info was specified in the main configuration file for the analysis (config.yaml), information for the sample fasta files was added to the sample configuration file (Sample-Config-protein.yaml), and we created the required sample-read-counts file. This snakemake (Snakefile-summarizeProteinRMA) was run using all other default settings, and kreport files were included in the outputs.

MEGAN-LR-nuc. A streamlined workflow for MEGAN-LR-nuc is available as two pipelines (Taxonomic-Profiling-Nucleotide, MEGAN-RMA-Summary) at: https://github.com/PacificBiosciences/pb-metagenomics-tools. These pipelines are provided as configurable snakemake workflows. To use the workflow, we first downloaded the NCBI nt database and indexed it with minimap2 using the following command:

```
minimap2 -k 19 -w 10 -I 10G -d mm_nt_db.mmi nt.gz
```

We downloaded MEGAN6 community edition to obtain the executable tools required for these workflows (sam2rma, rma2info), as well as the required MEGAN nucleotide mapping file (megan-nucl-Jan201.db). We then ran the Taxonomic-Profiling-Nucleotide pipeline. The locations of the minimap2 nt index, sam2rma, and the mapping file were specified in the main configuration file for the analysis (config.yaml), and we also changed the maximum number of secondary alignments from 20 to 5. The information for the sample fasta files was added to the

sample configuration file (Sample-Config.yaml), and the snakemake (Snakefile-taxnuc) was executed. Details for the usage of each program are provided in the online documentation.

Analyses resulted in RMA output files, which were used as inputs for the MEGAN-RMA-Summary pipeline. The location of rma2info was specified in the main configuration file for the analysis (config.yaml), information for the sample fasta files was added to the sample configuration file (Sample-Config-nucleotide.yaml), and we created the required sample-read-counts file. This snakemake (Snakefile-summarizeNucleotideRMA) was run using all other default settings, and kreport files were included in the outputs.

The above instructions are for the MEGAN-LR-nuc-HiFi analysis. Running the MEGAN-LR-nuc-ONT analysis required some changes. Specifically, we indexed the database with minimap2 using the following command:

```
minimap2 -k 15 -w 10 -I 10G -d mm_nt_db_ONT.mmi nt.gz
```

We then edited the minimap2 command in the snakemake file to include the ONT recommended settings:

```
minimap2 -ax map-ont
```

BugSeq. We uploaded datasets to the BugSeq online platform: https://bugseq.com. For each dataset, we selected the NCBI nt reference database option, and submitted the analysis. After successful completion all results were available for download.

RESULTS

The kreport files produced from all taxonomic profiling methods, and the Jupyter notebooks used to generate the following results, are freely available on the Open Science Framework project page (https://osf.io/bqtdu/). These files can be used to replicate all results reported below.

Comparative Analyses

Read Utilization. Total read assignment differed drastically across the profiling methods (Fig. 2). In terms of short-read methods, Kraken, Bracken, and Centrifuge-h22 assigned the greatest number of reads (93–100% for HiFi, 81–99% for ONT). Centrifuge-h500, which required a minimum total length of 500 for partial hits, assigned a much smaller number of reads across datasets (1–53%), with the exception of HiFi ATCC MSA-1003 (which had 98% read assignment). Read assignment was exceptionally low for Centrifuge-h500 in ONT R10 Zymo D6300 (~1%; Fig. 2). We found MetaPhlAn3, which relies primarily on read mapping, consistently assigned the fewest reads of any method (23–39%; Fig. 2).

There was considerable variation in read assignments across the long-read methods and across different sequencing technologies (Fig. 2). Total read assignment in the HiFi datasets ranged from 71–99% (average = 85%) across all long-read methods, and in the ONT ranged from 46–97% (average = 71%). For the ONT datasets, MetaMaps and BugSeq-V2 assigned the greatest number of reads (95–97%), with all other methods assigning fewer reads (46–67%). Methods that rely on translation alignments to protein references assigned more reads in the HiFi datasets versus ONT datasets, including MMseqs2 (HiFi: 94–99%; ONT: 46–67%) and MEGAN-LR-prot (HiFi: 71–74%; ONT: 60–62%) (Fig. 2). There were no clear differences in total read assignment for MEGAN-LR-nuc-HiFi and MEGAN-LR-nuc-ONT within the ONT datasets or the HiFi datasets, suggesting read assignment was not sensitive to different minimap2 settings. The MEGAN-LR-nuc methods resulted in a higher number of reads assigned in HiFi datasets (81–90%) versus ONT datasets (54–60%). BugSeq-V2 assigned more reads in the ONT datasets (95–96%) versus HiFi datasets (82–93%). As expected, methods using LCA algorithms (MMseq2, all three MEGAN-LR workflows, BugSeq-V2) displayed a significant proportion of read assignments in taxonomic ranks beyond the strain and species level (Fig. 2). However, the MEGAN-LR-nuc methods showed a smaller proportion of reads assigned to higher ranks, relative to the protein-alignment methods.

Detection Metrics. The complete set of read counts per dataset used in the species and genus-level analyses are provided in Supplementary Tables S1–S8. Detection at different thresholds follows the minimum read counts in Table 3.

The species-level detection results based on the minimum threshold of 0.001% of the total reads are summarized in Figure 3 and Table 4. The clearest difference in performance

occurs between short-read and long-read methods. The short-read methods display very low precision and relatively high recall, and consequently very low F-scores (Fig. 3). These results for precision and F-scores are driven by the large number of false positives detected (40–300) despite the presence of few false negatives (Table 4). We note that Bracken did not significantly improve the results of Kraken2, based on these measures (Fig. 3). The Centrifuge-h500 analysis, which required longer matches, resulted in a lower number of false positives and consequently higher precision (Fig. 3, Table 4), though this improvement varied considerably across datasets. MetaPhlAn3 displayed values that were intermediate between Centrifuge-h500 and the other short-read methods.

The long-read methods outperformed the short-read methods in terms of precision, recall, and F-scores (Fig. 3, Table 4), but also displayed variation in performance. Some methods did not show consistent results and performed better for a particular dataset. For example, MetaMaps and MMseqs2 performed quite well for HiFi ATCC MSA-1003. However, these two methods performed worse for the other three datasets and more closely resembled the results for the short-read methods (e.g., very low precision, higher recall; Fig. 3, Table 4). Across all four datasets, MEGAN-LR-prot, MEGAN-LR-nuc-HiFi, MEGAN-LR-nuc-ONT, and BugSeq-V2 consistently displayed the best performance. These four methods detected most species in the communities (e.g., low false negatives) and rarely called any false positives (0–2). Consequently, they display high precision, moderate to high recall, and the highest F-scores (Fig. 3). The moderate recall scores for the HiFi datasets resulted from the failure to detect species at lower abundances, particularly for the 0.02% to 0.0001% abundance levels (Supplementary Table S9). For the ONT datasets, the species in Zymo D6300 had comparatively high abundances (12% and 2%), and this was reflected in perfect recall for nearly all long-read methods (Fig. 3, Table 4). We did not observe any difference in performance between MEGAN-LR-nuc-HiFi and MEGAN-LR-nuc-ONT for the ONT datasets or HiFi datasets, suggesting the profiling analyses are not sensitive to minimap2 alignment settings.

The genus-level analysis based on the minimum threshold of 0.001% of the total reads largely mirrored the species-level results, but with expected improvements in precision, recall, and F-scores (Fig. 4, Supplementary Table S10). Improvements were nearly guaranteed because reads assigned to multiple species within a genus are all considered correct at the genus level, and consequently the number of false positives (and potentially false negatives) decreased.

Despite improvements in precision, recall, and F-scores across all methods at the genus level, the long-read methods still outperformed short-read methods by a considerable margin (Fig. 4, Supplementary Table S10).

Requiring a moderate minimum threshold for detection (0.1% of total reads) for the species-level analysis had an overall positive effect on precision, but negative effect on recall (Supplementary Fig. S2, Supplementary Table S11). These changes were most dramatic for the short-read methods, in which the number of false positives was reduced from several hundred to ~10 or fewer, thereby increasing precision considerably (Supplementary Table S11). However, despite this improvement the long-read methods still performed better in terms of precision and F-scores (Supplementary Fig. S2). Precision increased for some long-read methods (MetaMaps, MMseqs2), but others were unaffected as they were already high at the lower detection threshold. As expected, this increase in minimum detection threshold most strongly impacted recall in the communities with staggered abundances (HiFi datasets) versus communities with even abundances (ONT datasets). In the HiFi datasets, the long-read methods displayed more false negatives which resulted in lower recall (Supplementary Fig. S6). At the 0.1% total reads detection threshold, all methods (long and short) failed to detect species with <0.02% abundance and missed several species with 0.1–1.8% abundance (Supplementary Table S12). Surprisingly, this detection threshold also reduced the recall of some methods for the ONT datasets, with a more noticeable reduction in recall values for ONT R10 Zymo D6300 (Supplementary Fig. S2, Supplementary Table S11). The patterns for the genus-level analysis using the 0.1% total reads detection threshold mirrored the species-level results (Supplementary Fig. S3). Precision increased in the short-read methods across all datasets, and recall was lowered in the staggered abundance communities (Supplementary Table S13).

The highest minimum threshold for detection used in our experiment (1% of total reads) exacerbated the effects described for the 0.1% detection threshold. The most noticeable effects were for the communities with staggered abundances: all methods displayed perfect precision (with one exception), but recall was drastically lowered (<0.6; Supplementary Fig. S4, Supplementary Table S14). In other words, false positives were completely eliminated, but at the cost of increased false negatives. Using 1% of total reads as the minimum detection threshold for HiFi ATCC MSA-1003 and Zymo D6331, all methods (long and short) failed to detect species with <1.8% relative abundance, and some species were not detected in the 1.5% and 6%

abundance levels (Supplementary Table S15). This higher threshold for detection also impacted results for the even abundance communities (ONT R10 and Q20 for Zymo D6300). Precision increased primarily for the short-read methods, yet perfect precision was not achieved by all methods (Supplementary Fig. S4, Supplementary Table S14). This higher detection threshold also caused recall to drop (<0.8) in these datasets for all methods except Kraken2, Bracken, and one instance of BugSeq V2, each of which maintained perfect recall (Supplementary Fig. S4). This indicates that multiple methods failed to detect several species at the 2% and 12% abundance levels in Zymo D6300. These effects were mirrored in the genus-level analysis with the 0.1% detection threshold (Supplementary Fig. S5, Supplementary Table S16).

Relative Abundance Estimates. The species-level and genus-level relative abundances are shown in Figures 5 and 6, respectively. The results of the chi-squared goodness of fit tests (GOF) are reported in Supplementary Tables S17 and S18 and highlighted in Figures 5 and 6. At the species level, abundance estimates by the long-read methods were more accurate than those produced by short-read methods (based on chi-squared test statistic values; Supplementary Table S17). For HiFi ATCC MSA-1003, MetaMaps, MMseqs2, MEGAN-LR-prot, and BugSeq-V2 all passed the GOF, and BugSeq-V2 had the lowest error. All methods failed the GOF for HiFi Zymo D6331 at the species level (which had two species missing from most databases, see methods), but MEGAN-LR-prot and BugSeq-V2 resulted in the lowest error. For ONT R10 Zymo D6300, only BugSeq-V2 passed the GOF, but error was relatively low in MetaMaps and MEGAN-LR-prot. Both BugSeq-V2 and MEGAN-LR-prot passed the GOF for ONT Q20 Zymo D6300, with Centrifuge-h500 having the next lowest abundance error. At the genus level we generally found more methods passed GOF for each dataset, except for HiFi Zymo D6331 for which only BugSeq-V2 passed (Supplementary Table S18). All methods that accurately estimated abundances at the species level also passed the GOF at the genus level (Figs. 5, 6). We additionally found Centrifuge (h22 and/or h500) and MetaMaps passed GOF at the genus level in some datasets in which they failed at the species level (Figs. 5, 6). Although only one method (BugSeq-V2) passed the GOF for HiFi Zymo D6331, most of the other long-read methods (MMseqs2, MEGAN-LR-nuc, MEGAN-LR-prot) had lower error relative to the short-read methods (Supplementary Table S18). Across all datasets and levels, we generally found that BugSeq-V2 had the lowest abundance error, followed closely by MEGAN-LR-prot

(Supplementary Tables S17, S18). Across datasets, the proportion of reads assigned to false positives ('Other', Figs. 5, 6) was generally highest for MetaPhlAn3, followed by Kraken2 and Bracken.

Analyses of Shorter ONT Reads. Comparisons of the length-filtered variations of each ONT dataset revealed that shorter reads (< 2kb) negatively impacted taxonomic profiling analyses. For each ONT dataset, we created a primary dataset which contained only longer reads (> 2kb) and a secondary dataset which had a large proportion of shorter reads (< 2kb; see methods). In the primary datasets, precision and F-scores were very high for long-read methods and low for short-read methods at the 0.001% reads detection threshold. In the secondary datasets, precision and F-scores were comparatively lower for the long-read methods and were similarly low for the short-read methods (Supplementary Fig. S6, Tables S19, S20). Based on Wilcoxon Signed-Rank tests, the observed differences in precision and F-scores between the primary and secondary datasets were not statistically significant. However, at the 0.1% reads detection threshold we found precision and F-scores were substantially lower in the secondary datasets at both the species and genus level, across all methods (Supplementary Fig. S6, Tables S19, S20). These differences in precision and F-scores were statistically significant ($p < 0.01$ for all comparisons). In contrast to most methods, BugSeq produced relatively consistent results in precision and F-scores between the primary and secondary datasets across the different filtering thresholds.

Relative abundance estimates appeared heavily skewed in the secondary datasets, and most methods greatly overestimated the abundance of *Limosilactobacillus fermentum* in the community (Supplementary Fig. S7). Interestingly, in the secondary datasets the abundance error at the species level decreased for the short-read methods but increased in the long-read methods. At the genus level, abundance error appeared to increase across all methods in the secondary datasets. Based on Wilcoxon Signed-Rank tests, we did not find a significant difference in abundance error between the primary and secondary datasets at the species level, but at the genus level abundance error was significantly higher in the secondary datasets ($p < 0.05$ for the R10 and Q20 comparison). In the secondary datasets, nearly every method failed the chi-squared goodness of fit test at the species level (21 of 22) and genus level (20 of 22; Supplementary Tables S21, S22). We found BugSeq and Centrifuge-h22 passed the GOF for the species level of ONT R10 Short, and BugSeq passed the GOF for ONT R10 Short at the genus level

(Supplementary Tables S21, S22). No methods passed the GOF for ONT Q20 Short at the species or genus level.

Analyses of Illumina and Artificial Short Reads. We evaluated the performance of Kraken2, Bracken, Centrifuge-h22, and MetaPhlAn3 for two types of short-read datasets for the ATCC MSA-1003 and Zymo D6300 mock communities. We found detection and abundance results were highly similar between the Illumina short-read datasets and the "simulated" short-read datasets (which were derived from the long reads). This indicates that differences in results between the long-read datasets and the Illumina short-read datasets are unlikely to be driven by confounding effects (such as DNA extraction methods or error profiles).

The precision, recall, and F-score values obtained from the short-read datasets strongly resembled those obtained from long reads for both communities (Fig. 7, Table 4, Supplementary Figure S8, Supplementary Tables S23–24, S27–28). This overall pattern included a high number of false positives, low precision, high recall, and low F-scores. For the more complex ATCC MSA-1003 community (20 species), Kraken2, Bracken, and Centrifuge-h22 displayed a high number of false positives (57–113), low precision (0.15–0.26), and high recall (1.0; Fig. 7). More stringent filtering (0.1% or 1% of total reads) dramatically reduced false positives, but also negatively impacted recall (Supplementary Table S23), and in many cases produced scores that were worse than the long-read scores for these method and filtering combinations (Supplementary Table S11, S14). The same patterns were present for the genus-level analyses of the short-read datasets of ATCC MSA-1003 (Supplementary Table S24). Notably, MetaPhlAn3 performed much better for the short-read ATCC MSA-1003 datasets, with high precision (0.92) and suitable recall (0.60). We found similar results for the less complex ZymoD6300 community (10 species). Kraken2, Bracken, and Centrifuge-h22 displayed a high number of false positives (27–96), low precision (0.09–0.23), and high recall (0.8–1.0), and suitable trade-offs between precision and recall were not obtained through filtering (Supplementary Table S27). MetaPhlAn3 also performed better for these datasets regarding precision and recall (0.88 and 0.70, respectively), relative to the other three profiling methods.

The short-read datasets failed to produce accurate relative abundance estimates (Fig. 7, Supplementary Figures S9–10, Supplementary Tables S25–26, S29–30). All short-read methods failed the chi-squared goodness of fit test at the species level in both communities (16 of 16), and

at the genus level only one method passed in one community (15 of 16; MetaPhlAn3for ATCC-MSA-1003).

DISCUSSION

With decreasing error rates in long reads and the recent introduction of new long-read read profiling methods, long reads are increasingly utilized for metagenomic applications. We used publicly available mock community datasets to perform a critical assessment of taxonomic profiling methods for long-read datasets, including five long-read methods and four short-read methods. While all methods displayed some trade-offs between precision and recall, our results suggest that methods designed for long reads generally outperformed short-read taxonomic profiling methods.

An ideal read profiling method should assign all assignable reads (e.g., those with database representation), display high precision and recall, and provide accurate abundance estimates. Profiling methods tended to fall into two categories: those with high precision and moderate recall, and those with low precision and high recall. Overall, we find that BugSeq, MEGAN-LR-prot, and MEGAN-LR-nuc, all of which leverage mapping to whole genome or proteome references, provide the best tradeoffs (e.g., high precision, suitable recall, more accurate abundance estimates) for long-read metagenomics data. Below, we summarize our findings for short-read methods, long-read methods, and provide recommendations for long-read datasets. Finally, we briefly summarize differences in method performance between PacBio HiFi and ONT datasets.

The short-read methods (Kraken2, Bracken, Centrifuge-h22, MetaPhlAn3) assigned a high proportion of reads and displayed high recall, but they produced poor abundance estimates. They also recovered a very high number of false positives (15–300 species) and consequently had very low precision and F-scores (Figs. 2, 3). False positives were not a trivial proportion of assigned reads; they comprised up to 25% of the reads assigned at the species level (Fig. 5). Previous studies have shown similar results for these methods with short-read datasets (McIntyre et al. 2017; Ye et al. 2019; Parks et al. 2021), but here we demonstrate the use of long reads does not significantly change these trade-offs. We also note that in contrast to the other short-read

methods, MetaPhlAn3 assigned the fewest reads of any method (Fig. 2). The mapping and coverage strategy used in MetaPhlAn3 is not suitable for long reads, at least in its current implementation. We attempted to apply long-read settings to Centrifuge (Centrifuge-h500) to improve detection results. Unfortunately, this setting reduced total read assignment and had unpredictable outcomes on precision, recall, and F-scores across the datasets (Figs. 2, 3). We also applied different minimum thresholds for detection (0.001%, 0.1%, and 1% of the total reads) in an effort to reduce false positives and improve precision. A moderate detection threshold (0.1% total reads) successfully reduced the false positive count of species from hundreds to fewer than 15, and without significantly reducing recall. However, precision in these methods was still below those produced by the long-read methods without any filtering. A stringent detection threshold (1% total reads) greatly improved precision for many short-read methods, but severely impacted recall by eliminating detection of many species at lower abundance levels (<2% abundance). Overall, filtering did not adequately improve the balance between precision and recall for the short-read methods. We found similar results (high false positives, low precision, inaccurate abundance estimates) when we applied the short-read profiling methods to Illumina short-read datasets for two of the mock communities. The consistency of results across short and long-read datasets for Kraken2, Bracken, Centrifuge, and MetaPhlAn3 suggests the outcomes are most likely driven by limitations of the methods themselves, rather than read type. Methods designed for short read data cannot currently utilize the additional long-range information present in long-read data. However, this is not limited to one particular classification strategy: Kraken, Bracken and Centrifuge all use k-mer based methods, either for pseudoalignment or reference index compression, while MetaPhlAn3 relies upon read mapping to clade-specific marker genes.

Several long-read profiling methods showed consistent and favorable characteristics across all datasets. These include MEGAN-LR-prot, MEGAN-LR-nuc (both mapping settings), and BugSeq displayed medium to high read assignment and very high precision (Figs. 2, 3, Table 4). Recall values from these methods differed between the staggered abundance and even abundance communities (0.7–0.8 and 1, respectively). This difference is explained by the failure to detect species with <0.02% abundance in the staggered community. In contrast to the short-read methods, several long-read methods estimated accurate species abundances for the complex communities (particularly ATCC MSA-1003; Fig. 5). Across all communities, we generally

found BugSeq displayed the lowest abundance error of any method, followed by MEGAN-LR-prot. Though abundance error was higher for Metamaps, MMseqs2, and MEGAN-LR-nuc, these methods still performed better than most short-read methods in most cases. We found that MetaMaps and MMseqs2 showed high read assignment and precision for one dataset (HiFi ATCC MSA-1003), but for all other datasets showed unfavorable qualities which resembled the short-read methods (e.g., high false positives and low precision, high recall). This contrasts with a recent study by Marić et al. (2020), who found MetaMaps performed better than MEGAN-LR. However, Marić et al. (2020) produced alignments for MEGAN-LR using a different method (LAST) and a reduced database, which may explain these differences. Several of the long-read methods displayed high or perfect precision (MEGAN-LR-prot, MEGAN-LR-nuc, BugSeq), and this did not change after applying a moderate detection threshold (0.1% of total reads). However, we observed a dramatic improvement in precision for MMseqs2 and MetaMaps, which was accompanied by a moderate reduction in recall (Supplementary Fig. S6). A more stringent detection threshold (1% total reads) resulted in perfect precision but severely reduced recall for all long-read methods. Overall, we found that filtering was not required for many long-read methods (MEGAN-LR-prot, MEGAN-LR-nuc, BugSeq), and that moderate filtering could be used to balance precision and recall for methods with higher false positive rates (MetaMaps, MMseqs2).

Our findings demonstrate the important trade-offs between precision, recall, and detection limits. Taxonomic profiling methods which have high recall (e.g., they find all the species present in a community) also tend to have low precision (e.g., they recover many false positives). In our experiment, methods with these characteristics tended to be the short-read methods (Kraken2, Bracken, Centrifuge-h22, MetaPhlAn3), but also included some long-read methods (MetaMaps, MMseqs2). By contrast, methods which have high precision (e.g., no false positives) tend to have lower recall (e.g., not all species are detected). In our experiment, this was represented by several long-read methods, including MEGAN-LR-prot, MEGAN-LR-nuc, and BugSeq. Notably, these high precision methods all involve mapping reads to whole-reference databases, and subsequently interpreting alignments across the entire length of reads. This strongly suggests that top-performing methods are those that can utilize long-range information available in long reads. If precision is the most important aspect of a long-read metagenomics experiment, we suggest using these three long-read methods, which do not require

any threshold filtering. These methods include the BugSeq V2 pipeline, the combination of minimap2, NCBI nt, and MEGAN-LR, and the combination of DIAMOND, NCBI nr, and MEGAN-LR. The choice among them could depend on which references will be used (proteins: MEGAN-LR-prot; nucleotide sequences: BugSeq, MEGAN-LR-nuc), computational skills/resource availability (BugSeq is an online service platform; the MEGAN-LR workflows require high resources and bioinformatics experience), and abundance estimation (BugSeq and MEGAN-LR-prot are considerably more accurate than MEGAN-LR-nuc). It is also possible to use MetaMaps and MMseqs2 after applying a moderate minimum detection threshold to reduce false positives. We do not recommend using any short-read methods for applications requiring high precision, as they require heavy filtering to achieve comparably high precision values (and at the cost of greatly reduced recall). Furthermore, as there is no consistent or universal filtering threshold that guarantees high precision, it is difficult to determine the optimal threshold for a new dataset. There may also be cases where recall is more important for an experiment. For this, it is important to know what the expected abundance level is for the target species (high vs. very low), and whether detecting false positives at similar abundance levels is relevant. For example, the long-read methods with high precision (MEGAN-LR-prot, MEGAN-LR-nuc, BugSeq) also had excellent recall for species with higher abundances. These three methods confidently detected species with 0.1% and greater abundance in all the mock communities, with no false positives detected at these higher abundance levels. However, they failed to detect any species with <0.02% abundance. Therefore, the detection limit for lower abundance species for these methods is between 0.1–0.02% relative abundance level. Other long-read methods (MetaMaps, MMseqs2) had higher detection power at lower abundance levels and detected nearly all species at 0.01–0.02% abundance, but none at <0.001% abundance. However, they also detected 15–45 false positives at similarly low abundances, making it difficult to determine which species should be considered ‘real’. The same general pattern is true for the short-read methods, but they display significantly more false positives at lower abundances, exacerbating this issue. An important caveat is that these detection limits are based on results from the PacBio HiFi staggered communities, which consist of 2–2.5 million reads and a minimum detection count of 20–25 reads (Table 3). In general, increasing the sequencing depth should also increase detection power for lower abundance taxa. However, none of the methods performed extremely well (in terms of balancing precision and recall) for very low abundance taxa (<0.02% abundance), and it is

unclear if increased sequencing depth can improve these results. If the goal of an experiment is to determine if a particular species is present at very low abundance, other methods may be more appropriate (such as targeted sequencing or mapping to a suitable reference genome).

Our comparisons of length-filtered datasets strongly suggest that including shorter long reads (< 2kb) can have an adverse effect on taxonomic profiling. We found that datasets with many shorter reads had significantly lower precision and F-scores compared to datasets containing only longer reads. We also found that the inclusion of shorter reads heavily skewed relative abundance estimates, which are based on read counts in our experiment. We acknowledge that calculating abundance estimates from the total number of aligned bases could potentially mitigate this effect, but this option is not available for all of the methods used here. Most importantly, we found that precision, F-scores, and relative abundances were affected in both the short-read and long-read profiling methods, suggesting these shorter read lengths may be a "gray" zone for both classes of methods. Therefore, we strongly recommend filtering these shorter long reads before attempting taxonomic profiling. This can be achieved bioinformatically after sequencing, but performing size selection during library preparation can also greatly reduce the number of shorter fragments that are sequenced.

We included mock community datasets sequenced with PacBio HiFi and ONT, allowing for limited comparisons of methods across sequencing technologies. Given that long-read methods outperformed short-read methods, we focus on long-read methods only. One noticeable difference occurs in read utilization for methods that perform translation alignments to protein references. Specifically, more reads were assigned in HiFi versus ONT datasets for MMseqs2 (94–99% vs. 46–67%) and to a lesser extent MEGAN-LR-prot (71–74% vs. 60–62%). This result could be related to differences in the mock communities sequenced, however the species in all three mock communities are expected to have adequate representation in the databases (except two species in HiFi Zymo D6331). It is more likely that differences in error profiles explain these results, as even slightly higher error rates are expected to negatively impact translation alignment (broken reading frames, premature stop codons). This is idea is supported by two observations. First, this effect was more pronounced for MMseqs2, which uses Prodigal for translation rather than a frameshift-aware method such as DIAMOND. Second, the ONT data include an R10.3 dataset with Guppy basecalling (mean = Q10.5; reported at data source) and the newest "Q20" chemistry release with Bonito v0.3.5 basecalling (expected modal quality ~Q20),

and we found fewer reads were assigned in the R10.3 dataset versus the Q20 dataset for MMSeqs2 (46% vs. 67%, respectively). We note the same pattern was present for Centrifuge-500, which requires 500 matched k-mer bases to the reference; read assignment improved dramatically from ONT R10.3 to Q20 (1% vs. 53%, respectively). However, despite the improvement in accuracy for the ONT Q20 dataset, it still had lower read assignment for protein alignment methods compared to the HiFi datasets (Fig. 2), for which all reads are >Q20 (median = Q36–40). Together, these results suggest that read quality remains critical for high-quality taxonomic profiling with long-read methods.

Different mock communities were available for PacBio HiFi (ATCC MSA-1003, Zymo D6331) and ONT (Zymo D6300), which prevents a direct comparison of detection metrics (precision, recall, and F-scores) and detection limits across sequencing technologies. The mock community sequenced with ONT is simpler than the HiFi mock communities in terms of the total number of species (10 vs. 17/20) and relative abundances (even vs. staggered). The simpler mock community design also prevented us from estimating recall and detection limits for lower abundance species with ONT data; our conclusions about detection power at low abundances are based exclusively on PacBio HiFi data. In their study, Marić et al. (2020) found that ONT pseudo-mock datasets displayed lower classification accuracy, higher false positives, and higher relative abundance error relative to PacBio pseudo-mock datasets. However, the pseudo-mock datasets for ONT and PacBio included in their study contained different numbers of species and abundance designs, meaning they were not direct comparisons. We caution against this type of approach, and instead propose that an objective comparison of detection metrics should be performed by sequencing the same mock community standard using both technologies. We also propose that a mock standard with high species diversity and staggered abundances will provide the most meaningful information for future benchmarking studies comparing methods and sequencing technologies.

CONCLUSION

With increasing quality and prevalence of long-read datasets, it is critical to assess the utility of these data for taxonomic profiling of metagenomic samples. Here, we evaluated several popular and new profiling methods for mock communities sequenced with PacBio HiFi and ONT. We

also included Illumina short read data for these communities as a comparison. Our results demonstrate there are clear precision and recall trade-offs associated with each profiler. We found that using short-read methods (Kraken2, Bracken, Centrifuge, and MetaPhlAn3) to profile long-read datasets resulted in many false positives, particularly at lower abundance levels. Filtering can increase precision for these methods, but comes at the cost of severely reducing recall. Importantly, we determined this pattern of low precision and high recall occurred for these methods using both long-read and short-read datasets. This suggests the methods themselves, rather than differences in read lengths, are driving these outcomes. By contrast, we found several long-read methods displayed high precision and suitable recall without any filtering necessary. These alignment-based methods include BugSeq (nucleotide alignments), and MEGAN-LR using translation alignments (DIAMOND to NCBI nr) or nucleotide alignments (minimap2 to NCBI nt). In terms of ease of use and minimal resource requirements, BugSeq ranks the highest among these top-performing methods. Our comparisons between long-read sequencing technologies indicate that read quality remains critical for taxonomic profiling performance. We found that read accuracy impacts the success of methods relying on protein predictions or exact kmer matches. We also found a high proportion of shorter long reads (<2kb) can result in lower precision and inaccurate abundance estimates, relative to length-filtered datasets. However, we emphasize that for any given mock community, the long-read dataset (analyzed with any long-read method) produced significantly better results than the short-read datasets. Methods which utilize long-range information present in long reads provide clear improvements in taxonomic profiling and abundance estimation, and demonstrate a clear advantage over short-read sequencing. To continue studying these effects, we propose that cross-platform sequencing of standardized mock communities would be most useful for future benchmarking studies.

COMPETING INTEREST STATEMENT

DMP is an employee and shareholder of Pacific Biosciences of California, Inc.

DATA AVAILABILITY STATEMENT

The mock community datasets are publicly available from the National Center for Biotechnology Information (NCBI), European Nucleotide Archive (ENA), or public lab websites: HiFi ATCC MSA-1003 (NCBI: PRJNA546278: SRX6095783), HiFi Zymo D6331 (NCBI: PRJNA680590: SRX9569057), Illumina ATCC MSA-1003 (NCBI: PRJNA510527: SRX5169925), Illumina Zymo D6300 (NCBI: PRJNA648136: SRX8824472), ONT Q20 Zymo D6300 (ENA: PRJEB43406: ERR5396170), and ONT R10 Zymo D6300 (https://lomanlab.github.io/mockcommunity/r10.html). The kreport output files for all methods and datasets, along with Jupyter notebooks and results files, are freely available on the Open Science Framework: https://osf.io/bqtdu/.

LITERATURE CITED

Arumugam, K., Bağci, C., Bessarab, I., Beier, S., Buchfink, B., Górska, A., Qiu, G., Huson, D.H., and R.B.H. Williams. (2019). Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. *Microbiome*, 7: 61.

Beghini, F., McIver, L.J., Blanco-Miguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Thomas, A.M., Manghi, P., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E.A., and N. Segata. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10:e65088.

Breitwieser, F.P., Lu, J., and S.L. Salzberg. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 20: 1125–1139.

Buchfink, B., Xie, C., and D.H. Huson. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12: 59–60.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and C. Van Broeckhoven. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34: 2666–2669.

Dilthey, A.T., Jain, C., Koren, S., and A.M. Phillippy. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nature Communications*, 10: 3066.

Escobar-Zepeda, A., Godoy-Lozano, E.E., Raggi, L., Segovia, L., Merino, E., Gutiérrez-Rios, R.M., Juarez, K., Licea-Navarro, A.F., Pardo-Lopez, L., and A. Sanchez-Flores. (2018). Analysis of sequencing strategies and tools for taxonomic annotation: defining standards for progressive metagenomics. *Scientific Reports*, 8: 12034.

Fan, J., Huang, S., and S.D. Chorlton. (2021). BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics*, 22: 160.

Govender, K.N., and D.W. Eyre. (2022). Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications. *bioRxiv*, https://doi.org/10.1101/2022.01.11.475979.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and R. Tappu. (2016). MEGAN Community Edition – interactive exploration and analysis of large-scale microbiome sequencing data. *PLOS Computational Biology*, 12: e1004957.

Huson, D.H., Albrecht, B., Bağci, C., Bessarab, I., Górska, A., Jolic, D., and R.B.H. Williams. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, 13: 6.

Kim, D., Song, L., Breitwieser, F.P., and S.L. Salzberg. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26: 1721–1729.

Leidenfrost, R.M., Pöther, D.-C., Jäckel, U., and R. Wünschiers. (2020). Benchmarking the MinION: evaluating long reads for microbial profiling. *Scientific Reports*, 10: 5125.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34: 3094–3100.

Lindgreen, S., Adair, K.L., and P.P. Gardner. (2016) An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6: 19233.

Lu, J., Breitwieser, F.P., Thielen, P., and S.L. Salzberg. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3: e104.

Marić, J., Križanović, K., Riondet, S., Nagarajan, N., and M. Šikić. (2020). Benchmarking metagenomic classification tools for long-read sequencing data. *bioRxiv*, https://doi.org/10.1101/2020.11.25.397729.

McIntyre, A.B.R., Ounit, R., Afshinnekoo, E., Prill, R.J., Hénaff, E., Alexander, N., Minot, S.S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N.A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R.R., Rosen, G.L., and C.E. Mason. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, 18: 182.

Meyer, F., Bremges, A., Belmann, P., Janssen, S., McHardy, A.C., and D. Koslicki. (2019). Assessing taxonomic metagenome profiles with OPAL. *Genome Biology*, 20: 51.

Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., and E.L. Karin. (2021). Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics*, 2021: 1–3.

Nicholls, S.M., Quick, J.C., Tang, S., and N.J. Loman. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience*, 8: 1–9.

Parks, D.H., Rigato, F., Vera-Wolf, P., Krause, L., Hugenholtz, P., Tyson, G.W., and D.L.A. Wood. (2021). Evaluation of the microba community profiler for taxonomic profiling of metagenomic datasets from the human gut microbiome. *Frontiers in Microbiology*, 12: 643682.

Pearman, W.S., Freed, N.E., and O.K. Silander. (2020). Testing the advantage and disadvantages of short- and long-read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*, 21: 220.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T.S.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M., Hansen, L.H.H., Sørensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Don Kang, D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W.W., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H.-H.H., Liao, Y.-C.C., Silva, G.G.G.Z., Cuevas, D.A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H.-P.P., Göker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A.E., Rattei, T., and A.C. McHardy. (2017) Critical assessment of metagenome interpretation - a benchmark of metagenomics software. *Nature Methods*, 14: 1063–1071.

Sui, H.-Y., Weil, A.A., Nuwagira, E., Qadri, F., Ryan, E.T., Mezzari, M.P., Phipatanakul, W., and P.S. Lai. (2020). Impact of DNA extraction method on variation in human and built environment microbial community and functional profiles assessed by shotgun metagenomics sequencing. *Frontiers in Microbiology*, 11: 953.

Tamames, J., Cobo-Simón, M., and F. Puente-Sánchez. (2019). Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC Genomics*, 20: 960.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.A., and

M.W. Hunkapiller. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37: 1155–1162.

Wood, D.E., and S.L. Salzberg. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15: R46.

Wood, D.E., Lu, J., and B. Langmead. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20: 257.

Ye, S.H., Siddle, K.J., Park, D.J., and P.C. Sabeti. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*. 178: 779–794.

Figure 1. Violin plots showing the read length distributions for the four mock community datasets included in this study, after length-filtering was applied to remove shorter reads (see methods). Interiors of plots contain white dots representing median values, black bars represent interquartile values, and black lines represent minimum and maximum range values. Read sizes range up to 50,000 bp in length, but the plot is clipped at 25,000 bp to show the core size distributions.

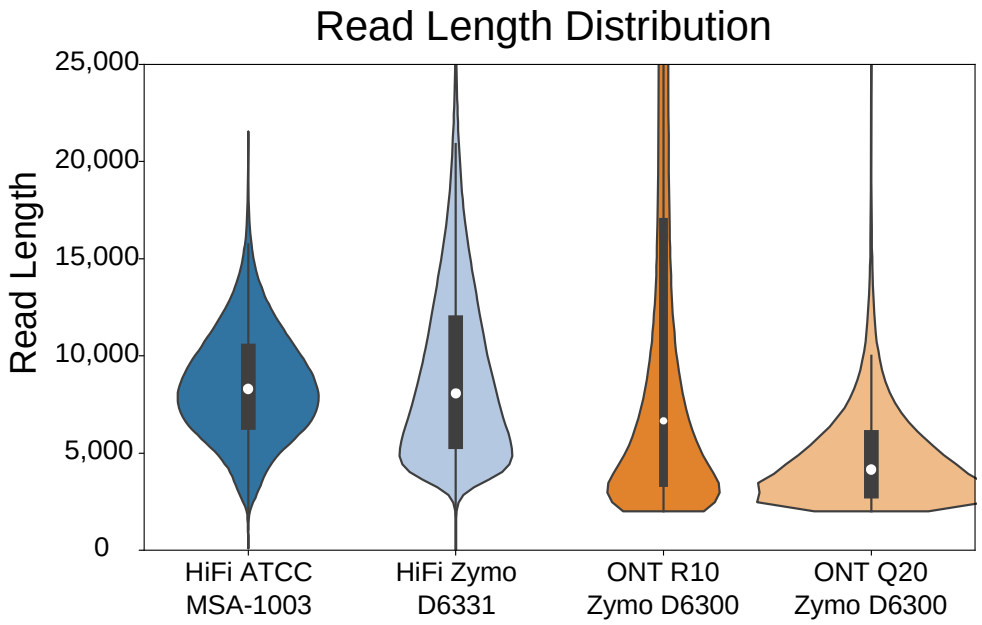

Figure 2. Read utilization for (A) HiFi ATCC MSA-1003, (B) HiFi Zymo D6331, (C) ONT R10 Zymo D6300, and (D) ONT Q20 Zymo D6300. The stacked barplots show the total percent of reads that were assigned to taxonomy. Different colors show the percentage of reads assigned to specific taxonomic ranks.

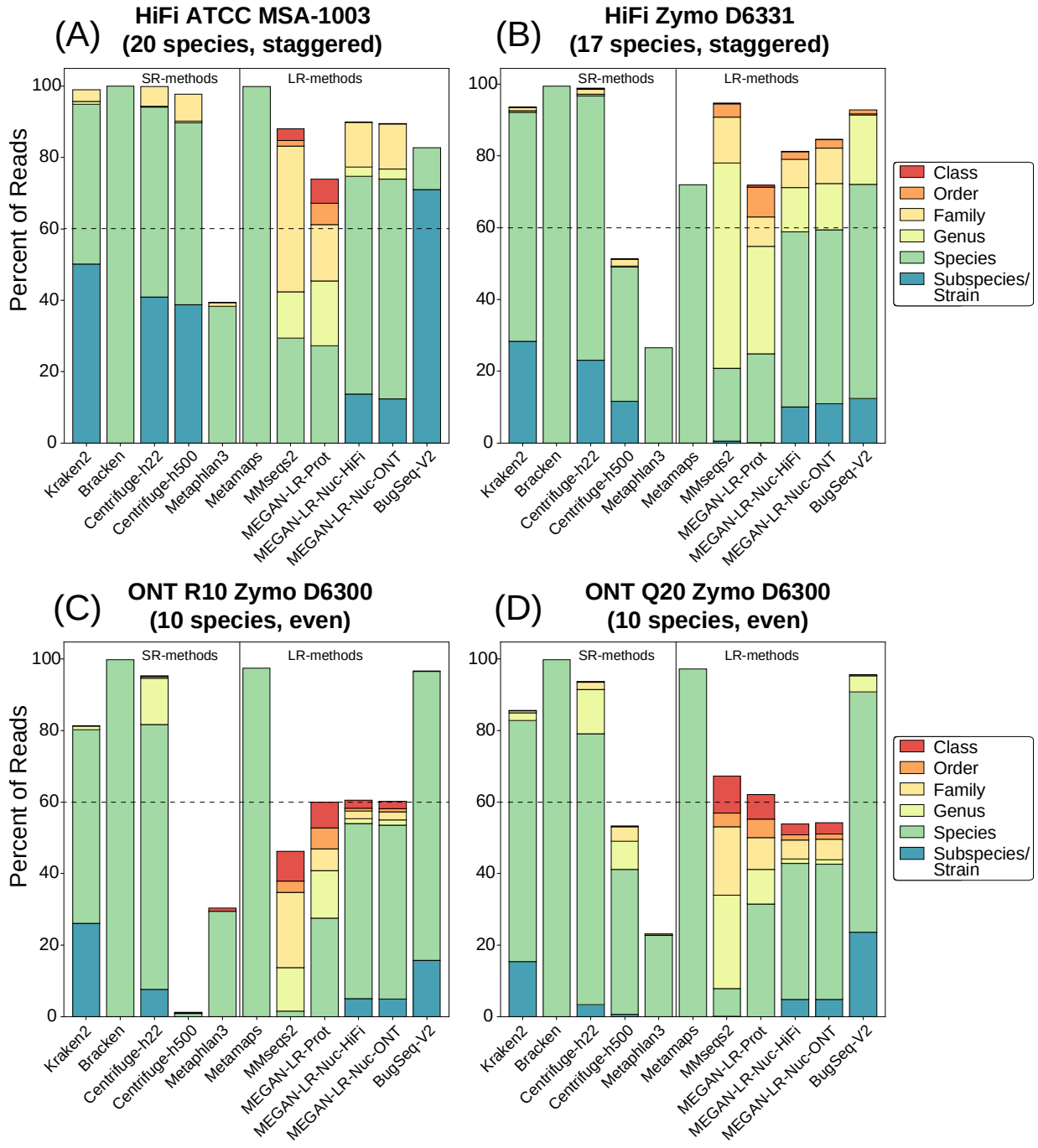

Figure 3. Precision, recall and F-scores for the species-level analysis based on a minimum threshold of 0.001% of the total reads for (A) HiFi ATCC MSA-1003, (B) HiFi Zymo D6331, (C) ONT R10 Zymo D6300, and (D) ONT Q20 Zymo D6300.

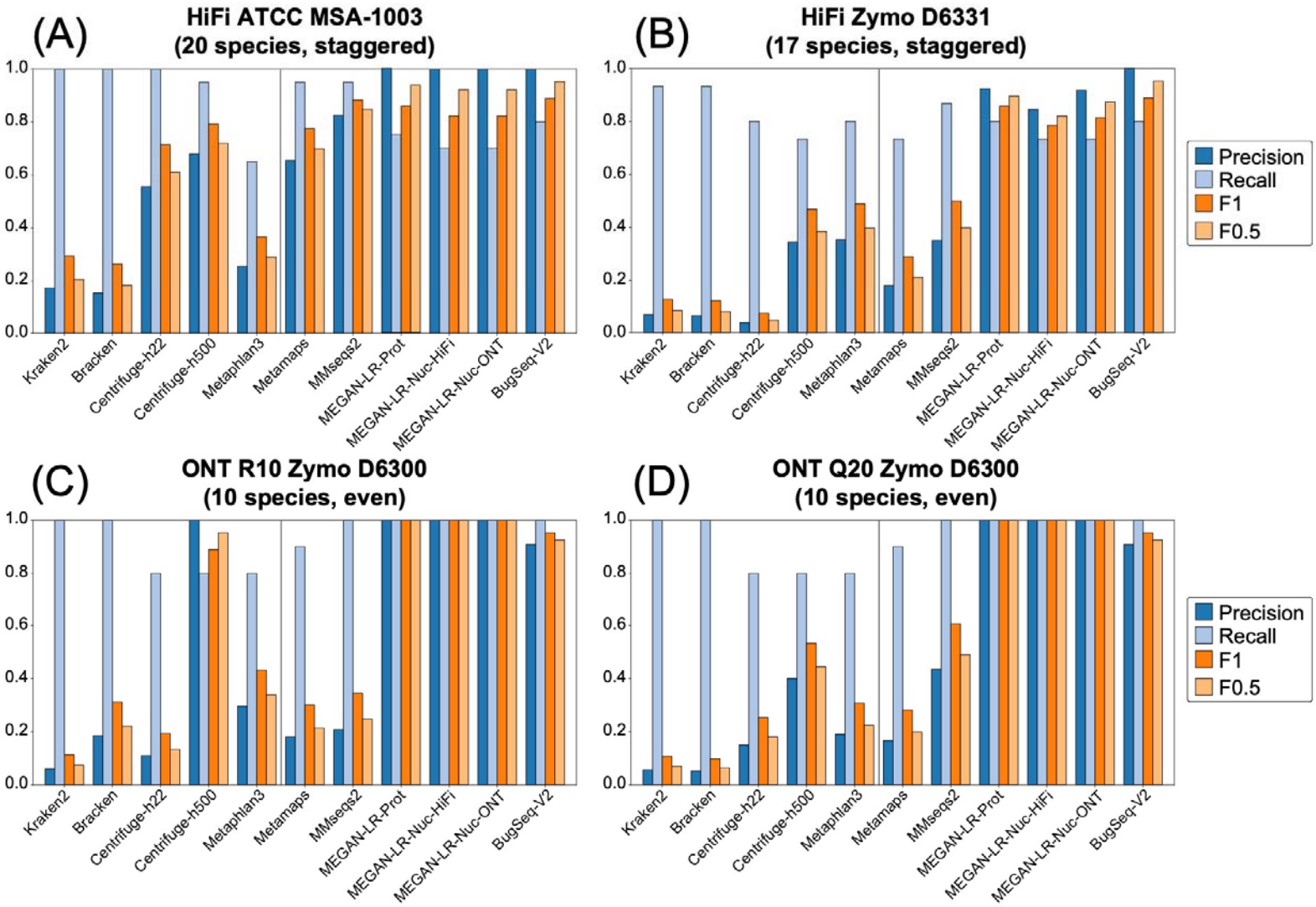

Figure 4. Precision, recall and F-scores for the genus-level analysis based on a minimum threshold of 0.001% of the total reads for (A) HiFi ATCC MSA-1003, (B) HiFi Zymo D6331, (C) ONT R10 Zymo D6300, and (D) ONT Q20 Zymo D6300.

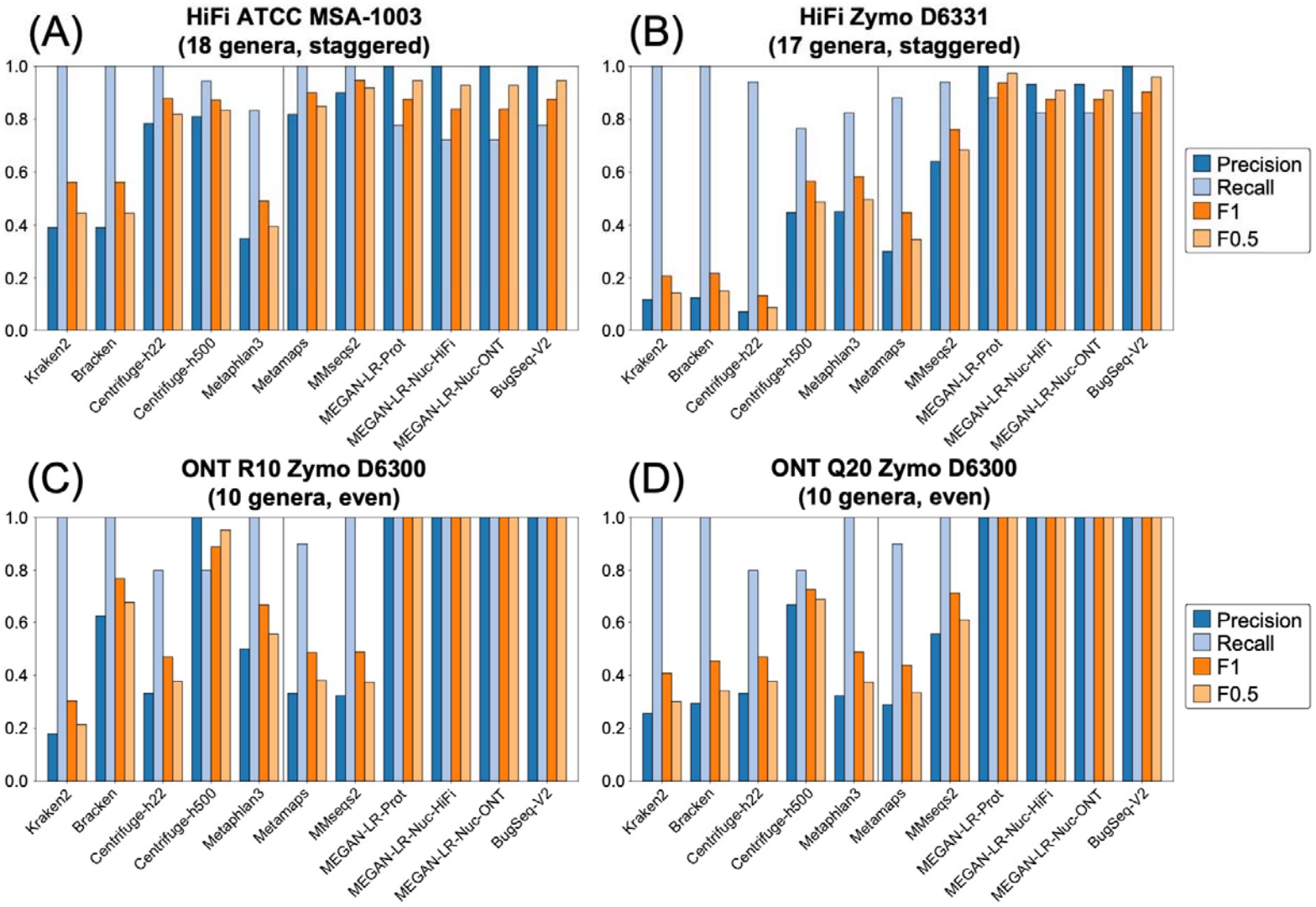

Figure 5. Species-level relative abundance estimates for (A) HiFi ATCC MSA-1003, (B) HiFi Zymo D6331, (C) ONT R10 Zymo D6300, and (D) ONT Q20 Zymo D6300. The theoretical distributions are shown on the left and are based on the manufacturer's specifications. The read counts for all species-level false positives were grouped in a category labeled 'Other'. For HiFi Zymo D6331, all species assignments within the genera *Prevotella* and *Veillonella* were counted towards *Prevotella corporis* and *Veillonella rogosae*, due to the absence of these species from several databases (see methods). Asterisks signify methods that failed the chi-squared goodness of fit test (e.g., the abundance estimates were significantly different from the theoretical values).

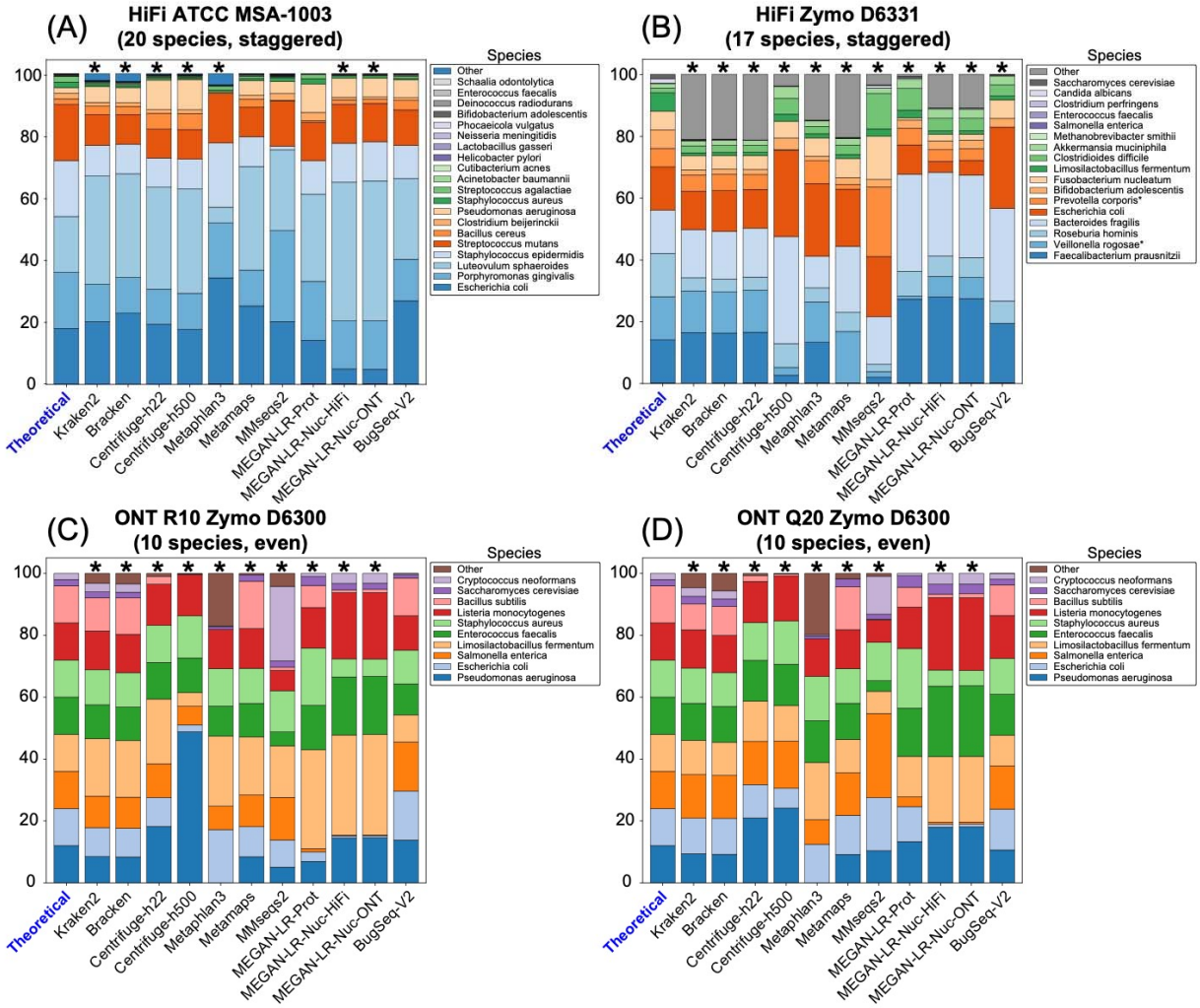

Figure 6. Genus-level relative abundance estimates for (A) HiFi ATCC MSA-1003, (B) HiFi Zymo D6331, (C) ONT R10 Zymo D6300, and (D) ONT Q20 Zymo D6300. The theoretical distributions are shown on the left and are based on the manufacturer's specifications. The read counts for all genus-level false positives were grouped in a category labeled 'Other'. Asterisks signify methods that failed the chi-squared goodness of fit test (e.g., the abundance estimates were significantly different from the theoretical values).

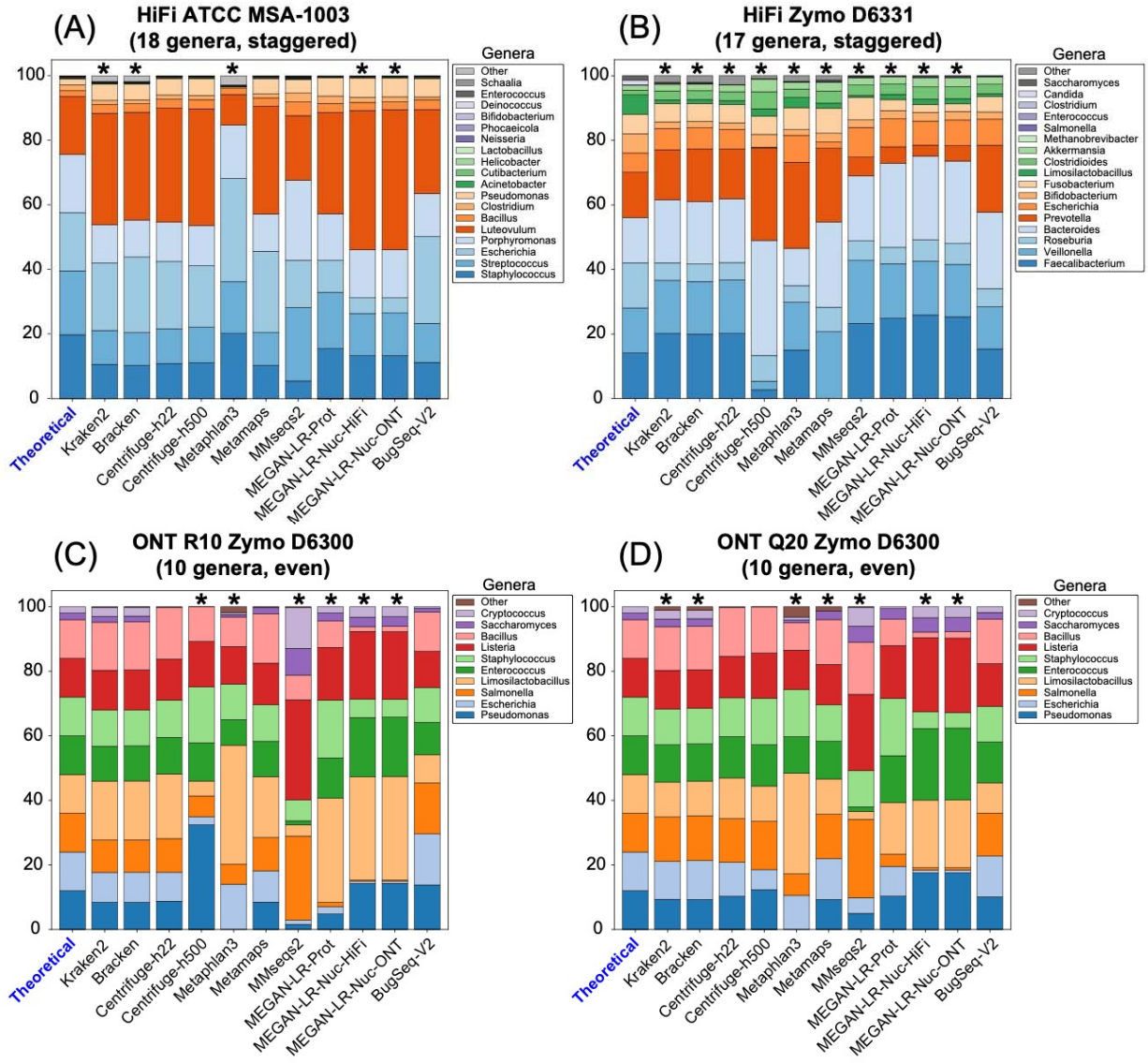

Figure 7. Results of the short-read profiling methods for the two Illumina short-read datasets. Precision, recall and F-scores for the species-level analysis based on a minimum threshold of 0.001% of the total reads for (A) Illumina ATCC MSA-1003 and (B) Illumina Zymo D6300. Species-level relative abundance estimates for (C) Illumina ATCC MSA-1003 and (D) Illumina Zymo D6300. The theoretical distributions are shown on the left and are based on the manufacturer's specifications. The read counts for all species-level false positives were grouped in a category labeled 'Other'. Asterisks signify methods that failed the chi-squared goodness of fit test.

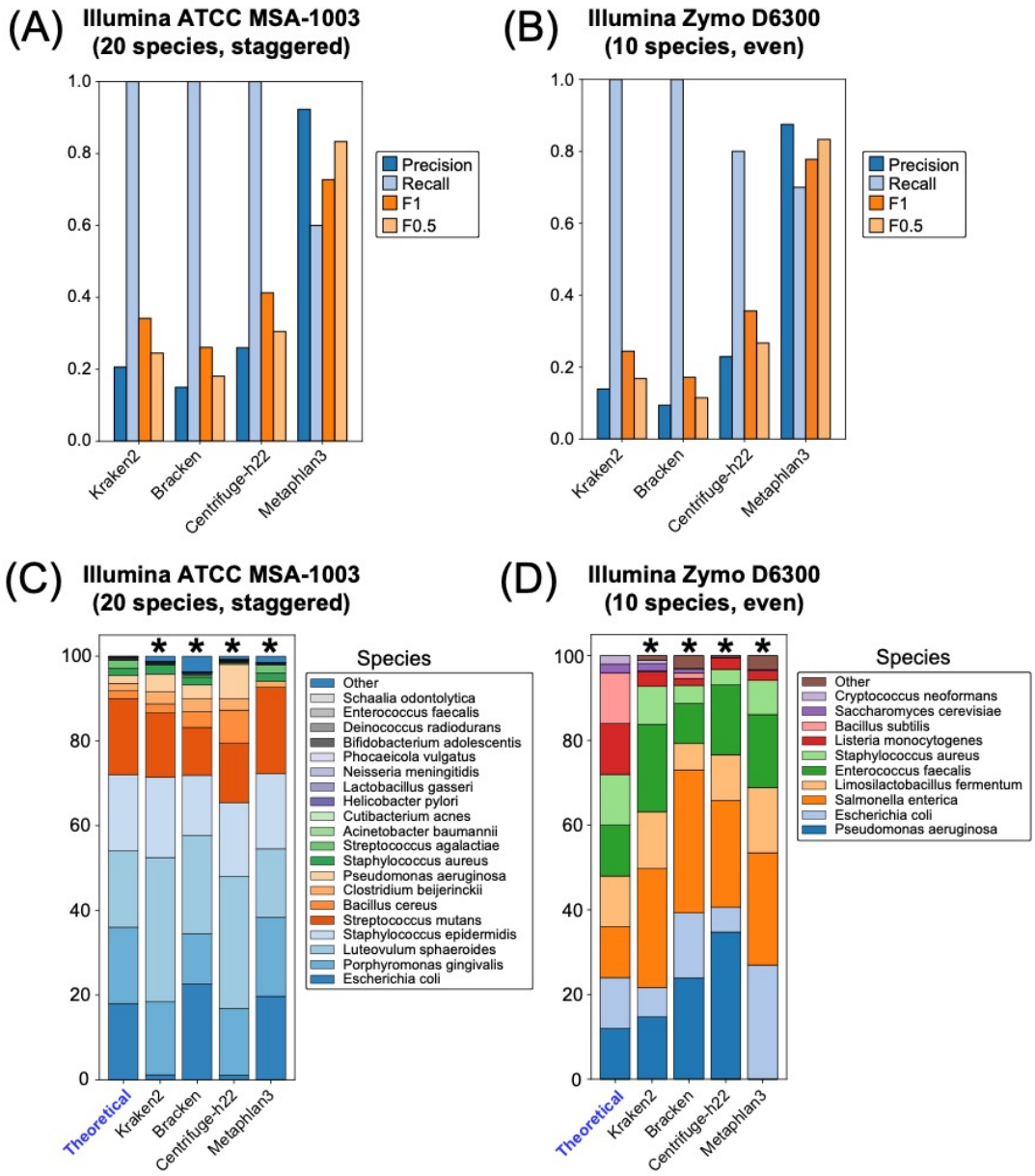

Table 1. Description of the publicly available mock community datasets used for this experiment.

Label	Technology	Mock Community	Species	Abundances	Reads Used	Median Length	Mean Length	Total Bases	Median QV	Release Date	Source
HiFi ATCC MSA-1003	PacBio HiFi	ATCC MSA-1003	20 [a]	Staggered (14–0.02%)	2,419,037	8,310	8,492	20.54 Gb	36	6/4/19	NCBI: SRX6095783
HiFi Zymo D6331	PacBio HiFi	ZymoBIOMICS D6331	17 [b]	Staggered (18–0.0001%)	1,978,852	8,077	9,092	17.99 Gb	40	11/25/20	NCBI: SRX9569057
ONT R10 Zymo D6300	Oxford Nanopore Technologies	ZymoBIOMICS D6300	10 [c]	Even (12%, 2%)	275,318[d]	6,664	12,022	3.31 Gb	10	2/7/20	https://lomanlab.github.io/mockcommunity/r10.html
ONT Q20 Zymo D6300	Oxford Nanopore Technologies	ZymoBIOMICS D6300	10 [c]	Even (12%, 2%)	2,000,000[d]	4,160	4,805	9.61 Gb	N/A	3/23/21	ENA: ERR5396170
Illumina ATCC MSA-1003	Illumina	ATCC MSA-1003	20 [a]	Staggered (14–0.02%)	10,038,314	125	125	1.25 Gb	37	12/2018	NCBI: SRX5169925
Illumina Zymo D6300	Illumina	ZymoBIOMICS D6300	10 [c]	Even (12%, 2%)	20,000,000[e]	150	150	2.99 Gb	37	7/2020	NCBI: SRX8824472

a: 20 bacteria; **b:** 14 bacteria, 1 archaea, 2 yeasts; **c:** 8 bacteria (at 12% abundance), 2 yeasts (at 2% abundance); **d:** length-filtered to eliminate reads < 2 kb and > 50 kb from starting set of 1.16 million reads (ONT R10) and 5.4 million reads (ONT Q20); **e:** subsampled from ~103 million available reads.

Table 2. Overview of taxonomic profiling methods used in this experiment.

Intended Use	Method	Variations	Reference Database	Strategy
Short reads	Kraken2	-	“PlusPF”	Kmer-based
	Bracken	-	“PlusPF”	Bayesian Refinement
	Centrifuge	h22, h500	Refseq ABVF	BW transform, FM-index
	MetaPhlAn3	-	mpa_v30_CHOCOPhlAn_201901	Read mapping, coverage scores
Long reads	MetaMaps	-	MiniSeq+H	Approximate mapping
	MMseqs2	-	NCBI nr	Translation alignment, LCA algorithm
	MEGAN-LR-prot	-	NCBI nr	Translation alignment, LCA algorithm
	MEGAN-LR-nuc	HiFi, ONT	NCBI nt	Nucleotide alignment, LCA algorithm
	BugSeq-V2	-	NCBI nt	Nucleotide alignment, LCA algorithm

Table 3. Minimum detection thresholds used to score the presence/absence of mock community taxa at the species or genus level.

Dataset	**Number of Reads**	**0.001% Threshold**	**0.1% Threshold**	**1% Threshold**
HiFi ATCC MSA-1003	2,419,037	24	2,419	24,190
HiFi Zymo D6331	1,978,852	19	1,978	19,788
ONT R10 Zymo D6300	275,318	2	275	2,753
ONT Q20 Zymo D6300	2,000,000	20	2,000	20,000
Illumina ATCC MSA1003	10,038,314	100	10,038	100,383
Illumina Zymo D6300	20,000,000	200	20,000	200,000

Table 4. Species-level detection results based on the minimum 0.001% of total reads threshold.

Dataset	Method Type	Profiling Method	True Positives	False Positives	False Negatives	Precision	Recall	F_1	$F_{0.5}$
HiFi ATCC MSA1003	Short read	Kraken2	20	96	0	0.17	1.00	0.29	0.21
(20 species, staggered)		Bracken	20	112	0	0.15	1.00	0.26	0.18
		Centrifuge-h22	20	16	0	0.56	1.00	0.71	0.61
		Centrifuge-h500	19	9	1	0.68	0.95	0.79	0.72
		MetaPhlAn3	13	38	7	0.26	0.65	0.37	0.29
	Long read	MetaMaps	19	10	1	0.66	0.95	0.78	0.70
		MMseqs2	19	4	1	0.83	0.95	0.88	0.85
		MEGAN-LR-Prot	15	0	5	1.00	0.75	0.85	0.94
		MEGAN-LR-Nuc-HiFi	14	0	6	1.00	0.70	0.82	0.92
		MEGAN-LR-Nuc-ONT	14	0	6	1.00	0.70	0.82	0.92
		BugSeq-V2	16	0	4	1.00	0.80	0.89	0.95
Illumina ATCC MSA1003	Short read	Kraken2	20	77	0	0.21	1.00	0.34	0.24
(20 species, staggered)		Bracken	20	113	0	0.15	1.00	0.26	0.18
		Centrifuge-h22	20	57	0	0.26	1.00	0.41	0.31
		MetaPhlAn3	12	1	8	0.92	0.60	0.73	0.83
HiFi Zymo D6331	Short read	Kraken2*	14	196	1	0.07	0.93	0.13	0.08
(17 species, staggered)		Bracken*	14	204	1	0.06	0.93	0.12	0.08
		Centrifuge-h22*	12	307	3	0.04	0.80	0.07	0.05
		Centrifuge-h500*	11	21	4	0.34	0.73	0.47	0.38
		MetaPhlAn3*	12	22	3	0.35	0.80	0.49	0.40
	Long read	MetaMaps*	11	50	4	0.18	0.73	0.29	0.21
		MMseqs2*	13	24	2	0.35	0.87	0.50	0.40
		MEGAN-LR-Prot*	12	1	3	0.92	0.80	0.86	0.90
		MEGAN-LR-Nuc-HiFi*	11	2	4	0.85	0.73	0.79	0.82
		MEGAN-LR-Nuc-ONT*	11	1	4	0.92	0.73	0.81	0.87
		BugSeq-V2*	12	0	3	1.00	0.80	0.89	0.95

ONT R10 Zymo D6300	Short read	Kraken2	10	156	0	0.06	1.00	0.11	0.07
(10 species, even)		Bracken	10	44	0	0.19	1.00	0.31	0.22
		Centrifuge-h22	8	65	2	0.11	0.80	0.19	0.13
		Centrifuge-h500	8	0	2	1.00	0.80	0.89	0.95
		MetaPhlAn3	8	19	2	0.30	0.80	0.43	0.34
	Long read	MetaMaps	9	41	1	0.18	0.90	0.30	0.21
		MMseqs2	10	38	0	0.21	1.00	0.34	0.25
		MEGAN-LR-Prot	10	0	0	1.00	1.00	1.00	1.00
		MEGAN-LR-Nuc-HiFi	10	0	0	1.00	1.00	1.00	1.00
		MEGAN-LR-Nuc-ONT	10	0	0	1.00	1.00	1.00	1.00
		BugSeq-V2	10	1	0	0.91	1.00	0.95	0.93
ONT Q20 Zymo D6300	Short read	Kraken2	10	166	0	0.06	1.00	0.11	0.07
(10 species, even)		Bracken	10	184	0	0.05	1.00	0.10	0.06
		Centrifuge-h22	8	45	2	0.15	0.80	0.25	0.18
		Centrifuge-h500	8	12	2	0.40	0.80	0.53	0.44
		MetaPhlAn3	8	34	2	0.19	0.80	0.31	0.22
	Long read	MetaMaps	9	45	1	0.17	0.90	0.28	0.20
		MMseqs2	10	13	0	0.44	1.00	0.61	0.49
		MEGAN-LR-Prot	10	0	0	1.00	1.00	1.00	1.00
		MEGAN-LR-Nuc-HiFi	10	0	0	1.00	1.00	1.00	1.00
		MEGAN-LR-Nuc-ONT	10	0	0	1.00	1.00	1.00	1.00
		BugSeq-V2	10	1	0	0.91	1.00	0.95	0.93
Illumina Zymo D6300	Short read	Kraken2	10	62	0	0.14	1.00	0.24	0.17
(10 species, even)		Bracken	10	96	0	0.09	1.00	0.17	0.11
		Centrifuge-h22	8	27	2	0.23	0.80	0.36	0.27
		MetaPhlAn3	7	1	3	0.88	0.70	0.78	0.83

*Two species were unavailable in several reference databases for HiFi Zymo D6331, and the set of taxa was adjusted to 15 species to calculate the species metrics (see methods).