Journal of Biological Methods | **2023** | **Vol**. 10 | e99010003
DOI: 10.14440/jbm.2023.376

Benchmarking assembly free nanopore read mappers to classify complex millipede gut microbiota via Oxford Nanopore Sequencing Technology

Orlando J. Geli-Cruz[1], Carlos J. Santos-Flores[1], Matias J. Cafaro[1], Alex Ropelewski[2], Alex R. Van Dam[1]*

[1]Universidad de Puerto Rico, Recinto Universitario de Mayagüez, Call Box 9000 Mayagüez, PR 00681-9000
[2]Pittsburgh Supercomputing Center, 300 S. Craig Street, Pittsburgh, PA 15213

***Corresponding author:** Alex R. Van Dam, Universidad de Puerto Rico, Recinto Universitario de Mayagüez, Call Box 9000 Mayagüez, PR 00681-9000, Tel: (787) 832-4040, exts. 3900, 2405 E-mail: alex.vandam@upr.edu

Competing interests: The authors have declared that no competing interests exit

Abbreviation used: ONT, Oxford Nanopore Technologies; NGS, next generation sequencing

Received August 30, 2021; Revision received March 13, 2023; Accepted April 27, 2023; Published August 4, 2023

ABSTRACT

Millipedes are key players in recycling leaf litter into soil in tropical ecosystems. To elucidate their gut microbiota, we collected millipedes from different municipalities of Puerto Rico. Here we aim to benchmark which method is best for metagenomic skimming of this highly complex millipede microbiome. We sequenced the gut DNA with Oxford Nanopore Technologies' (ONT) MinION sequencer, then analyzed the data using *MEGAN-LR*, *Kraken2* protein mode, *Kraken2* nucleotide mode, *GraphMap*, and *Minimap2* to classify these long ONT reads. From our two samples, we obtained a total of 87,110 and 99,749 ONT reads, respectively. *Kraken2* nucleotide mode classified the most reads compared to all other methods at the phylum and class taxonomic level, classifying 75% of the reads in the two samples, the other methods failed to assign enough reads to either phylum or class to yield asymptotes in the taxa rarefaction curves indicating that they required more sequencing depth to fully classify this community. The community is hyper diverse with all methods classifying 20–50 phyla in the two samples. There was significant overlap in the reads used and phyla classified between the five methods benchmarked. Our results suggest that *Kraken2* nucleotide mode is the most appropriate tool for the application of metagenomic skimming of this highly complex community.

Keywords: Metagenomics, Myriapoda, Nanopore sequencing, Gut microbiota, DNA extraction

INTRODUCTION

To elucidate the biodiversity of microscopic organisms next generation sequencing (NGS) technologies have within the last decade become widely utilized in this effort. Metagenomic skimming is a NGS analyses method which utilizes a shotgun sequencing approach to deliver a low depth DNA profile of a metagenomic sample. This low depth DNA profile can be used to obtain a DNA fingerprint of the phylogenetic diversity contained within a sample. Metagenomic skimming uses low coverage sequencing to obtain a reliable phylogenetic placement of rare species. An additional consideration is bias associated with the distance of the query sequence to the reference database used to make molecular identifications [1,2].

An emerging NGS tool for metagenomic skimming that is rapid and relatively inexpensive is Oxford Nanopore Technologies (ONT) sequencing, which has been used on a variety of ecosystems to rapidly inventory their biodiversity. To survey the biodiversity of a sample, ONT sequencing reads are used to directly align to a reference database or, in an alternative approach, reads are matched to reference sequences via statistical similarity using a k-mer subsequence based approach designed in part for noisy error prone ONT reads [3-6]. Whether an alignment-based approach or "alignment free"-k-mer-based approach is constrained by the genetic distance between the query sequence and the database for making taxonomic read assignment. Benchmarking studies investigating the efficacy of ONT read alignment programs have used known communities or have primarily focused on genome assembly and contaminant removal [4,6-8]. Using a known community is an excellent way to benchmark and evaluate the performance of ONT based alignment methods. However, understanding how well ONT read binning methods perform when

How to cite this article: Geli-Cruz QJ, Santos-Flores CJ, Cafaro MJ, Ropelewski A, Van Dam AR. Benchmarking assembly free nanopore read mappers to classify complex millipede gut microbiota via Oxford Nanopore Sequencing Technology. *J Biol Methods* 2023;10:e99010003. DOI: 10.14440/jbm.2023.376

the community is divergent or an unknown genetic distance from the known reference database has not been fully explored.

Metagenomic studies involving arthropods have focused primarily on insect systems. However, there are few comprehensive metagenomic sequencing surveys of the microbiota of other non-insect arthropods, millipedes included. Millipedes (Class: Myriapoda) are key to the decomposition of leaf litter into soil, alongside other macroinvertebrates such as earthworms and isopods. Despite the significance of millipedes in nutrient cycling at the soil leaf litter interface, they are a comparatively understudied arthropod group. For this study, we focused on the microbiota that inhabits the gut of *Anadenobolus monilicornis* (von Porat, 1876), a species of millipede native to the Caribbean [9]. It should serve as a good proxy for other similar leaf litter inhabiting millipedes in the tropics and serve as a novel system to compare ONT read classification methods for metagenomic skimming of similar metagenomic systems. Via ONT sequencing we aim to benchmark the quantity of read assignment and relative overlap of taxonomic assignment of five popular ONT compatible k-mer read assignment programs using millipede gut microbiota to benchmark their performance on this novel system.

MATERIALS AND METHODS

Millipede Sampling

Anadenobolus monilicornis millipedes were collected from the Puerto Rican municipalities of Mayagüez and Añasco (Rincón). The millipedes were kept in small glass containers with moist filter paper without food for 24 hours (Mayagüez sample), and ten days (Rincón sample). This was done to eliminate intestinal contents, so as not to sequence ingested organisms. We had planned to sequence additional millipedes, but the additional enzymes and reagents were destroyed during Hurricane Maria in October,2017 due to lack of refrigeration.

Gut dissection and DNA extraction

The gut dissection and DNA extraction work were done in the Symbiosis laboratory at the University of Puerto Rico, Mayagüez Campus. Following workstation and lab material sterilization with 10% bleach, the head and the last two or three segments of the abdomen of the specimens were cut and removed with a scalpel. The abdomen was cut to facilitate gut extraction. The guts were removed and placed in 2 mL tissue disruption tubes, then liquified by manually shaking the tubes. We followed the Qiagen Fast DNA Tissue Kit (Cat. #51404, Qiagen, North Rhine-Westphalia, Germany) protocol to purify the DNA, samples.

DNA Fragmentation and Library Build

First a master mix of 14 µL of Fragmentase buffer and 2 µL of 10X NEBNext® dsDNA Fragmentase® (NEB Cat. #M0348s, New England Biolabs, Ipswich, MA) was made. In new tubes, we added 32 µL of the samples and 8 µL of the master mix to each. The new tubes were vortexed for two seconds and spun down; they were then placed on a thermocycler for five minutes at 37°C followed by approximately five minutes at 4°C. In order to heat kill the Fragmentase, 5 µL of EDTA was added and placed on a thermocycler for 15 minutes at 65°C followed by 10 minutes at 5°C.We aimed to construct libraries with 5,000–30,000 Kb DNA fragments. DNA quality was verified using 2 µL of each sample mixed with 3 µL of loading dye and then added to a 1X TAE electrophoresis gel set to 66 V for 30 minutes. After this step, the ONT 1D PCR barcoding genomic DNA (SQK-LSK108, Oxford Nanopore Technologies, Oxford, England) for version R9 chemistry via the PCR protocol was followed to construct the libraries [10].

Barcoding PCR

2 µL of PCR barcode from the PCR Barcoding Kit (ONT Cat. #SQK-PBK004, Oxford Nanopore Technologies), 2 µL of 10 ng/µL adapter ligated template, 50 µL of NEB LongAmp Taq 2X Master Mix (NEB Cat. #M0287, New England Biolabs), and 46 µL of nuclease-free water were mixed.

Ligation of Sequencing Adapter

20 µL of Adapter Mix, 50 µL of Blunt/TA Ligation Master Mix and 30 µL of end-prepped DNA were mixed. After ten minutes at room temperature, another round of Ampure XP bead cleanup was performed. The finished samples were then transferred to Eppendorf DNA LoBind tubes.

SpotOn Flow Cell Prep and Sequencing

We followed the ONT for the SpotOn Flow Cell version R9 chemistry (ONT Cat. #FLO-MIN 107 R9, Oxford Nanopore Technologies). Library Loading Bead kit (ONT Cat #EXP-LLB001, Oxford Nanopore Technologies) was used to help load samples onto the Flow Cell. The two libraries were quantified on a NanoDrop and allowed for library pooling at the DNA concentration level. A q-PCR machine was not available at the time making equimolar pooling impossible. The pooled barcoded libraries were sequenced on two SpotOn Flow Cells. Both sequencing runs were carried out for 48 hours using a MacBook (16GB RAM, e 2.9 GHz Intel Core i5 processor, model early 2015) laptop and ONT *MinKNOW* software.

Quality filtering and de-multiplexing

We used the *MinKNOW* software program for initial quality filtering of reads obtained from both flow cells. The HDF5-formatted data from the nanopore sequencer was moved from the MacBook laptop, to the Pittsburgh Supercomputing Center's (PSC) Bridges Supercomputer. Within the PSC and using the Anaconda Python environment, we installed the *albacore* base-caller v2.1.3 [11] to separate the different barcodes and convert the data to FASTQ format [10]. Further quality filtering was performed

via *Nanofilt* v2.2.0 [12] implemented with q-10 quality filtering. Read length, Phred quality scores, and other summary statistics were calculated using *Pauvre* [13].

Metagenomic Mapping Programs

The mapping programs that should work well with shotgun long-read data produced by the ONT MinION sequencer. Nanopore mapping program 3s used were: *MEGAN-LR* v6.15.2 via *Lastal* v759 [8,14,15], *Kraken2* v2.0.7-beta run in nucleotide and protein modes [16], *GraphMap* v0.5.2 [4], and *Minimap2* v2.9-r720 [17]. These programs were chosen as they all take into account the relatively high error rate in nanopore reads when performing alignments [3,4,8,16-18]. *MEGAN-LR* and *Kraken2* in protein mode, used the NCBI *nr* (March 2019) amino acid database [8], a custom nucleotide database was used for the other three methods. For each of these programs the default settings were used. Optimizing the settings of each of the five programs for this specific dataset is beyond the scope of this benchmarking study. For programs that assign or map to multiple equally good or nearly equivalent hits (*MEGAN-LR* via *Lastal*, *GraphMap*, and *Minimap2*) we used the read with the lowest Levenshtein distance (edit distance) from the read to the database mapping. As quality scores vary by method and cannot be compared directly, the Leventshtein distance is the most equivalent metric across these three methods. This "best-hit" approach would reduce the variance in the taxonomic read assignment. *Kraken2* delivers a "best-hit" taxonomic assignment but does not provide edit distance directly as part of its output. *Kraken2* uses the read with the greatest number of k-mer hits shared between the read and database to assign taxonomy to the lowest common ancestor of the read [16].

Reference Database

Based on our initial DNA extractions we could observe an abundance of protists with associated bacteria, nematodes, and plant material in the millipede guts. Additionally, from previous studies we also expected to find symbiotic gut fungi [19]. To try and cover this massive breadth of taxonomic diversity we attempted to fuse several pre-existing databases as well as added in novel genome assemblies to represent the putative hyper-diverse communities' taxonomic range.

As no pre-existing millipede genomic assembly was available on NCBI or EMBL we used the fastq reads from the rusty millipede genome assembly project [20] to assemble a draft millipede genome to add to our database. We used NCBI Biosample SAMN03048671 and followed the methods in Kenny *et al*. [20] for quality filtering and draft genome assembly via *Velvet* 1.2.10 [21]. The rusty millipede draft genome scaffolds were used as representative millipede reference database for *Kraken2*, *GraphMap*, and *Minimap2*.

For the other phyla there were a variety of pre-existing genome assemblies that we could use. For the nematode representatives we used the 32 pre-existing draft genome assemblies from WormBase WS275 [22], see **Table S1** for a full list of accession numbers. All of these scaffolds were first repeat masked via *Dustmasker* [23] implemented via *blast tools* v. 2.7.1 [24]. We then fused the accessions from WormBase with the *Kraken2* bacteria, archaea, protist, fungi, and plant databases into a single database. Finally, the fasta files from each of these draft genomes were prepared as a database for each of the three nucleotide mapping programs *GraphMap*, *Minimap2*, and *Kraken2*.

Evaluation criteria

1) Quantification of ONT reads assigned to phylum or class taxonomic classification by method

We evaluated the number of "best-hits" assigned by each ONT mapping program. *NanoComp* 1.1.0 via *NanoPack* [12] was used to compute summary statistics on the numbers of reads, and the number of shared reads between assignment methods was computed via *Jvenn* [25]. Since many of the organisms in this sample are undescribed and new to science, we only evaluated the reads assigned by phylum and reads assigned by taxonomic class. Summary taxonomic assignment at the phylum and class taxonomic level was visualized via *MEGAN-6* v18.5 [8] as a stacked bar chart for qualitative comparison (**Fig. S1–S2**).

2) Quantification of biodiversity and shared taxonomy statistics between methods

Under the "best-hit" criteria previously mentioned we used the NCBI taxonomy ID assigned by each method to each read, at the phylum and class taxonomic levels. *iNext* [26] was used to evaluate the thoroughness of taxonomic assignments by each method. We used Hill numbers to measure species diversity with 100 randomizations with replacement via *iNext* [26]. Species rarefaction curves using Hill numbers were calculated using *iNext* (26) to provide confidence intervals around extrapolation of the data to simulate taxonomic assignment with equivalent numbers of samples across the five methods.

The Morisita-Horn shared species index was used to compare the shared phyla and class taxonomy levels between methods with rare taxa having fewer than 10 assignments via *SpadeR* [27]. The Morisita-Horn index is robust and more reliable than other shared species indices because it is not strongly influenced by species richness and sample size [28].

RESULTS

Quantification of ONT reads assigned to phylum or class taxonomic classification by method

Sequencing results: Read length and Phred quality scores for both samples can be seen in **Fig. S3–S4**, and summary statistics can be seen in **Table S2**. For the Mayagüez sample, we obtained a

total of 87,110 reads post *Nanofilt* quality filtering; for the Rincón sample, a total of 99,749 reads. For a full summary of reads assigned by method for each class and phylum see **Tables S3–S7**.

Numbers of reads classified by method: Kraken2 in nucleotide mode assigned by far the most reads and taxonomic diversity across all methods, with 9 and 8.7 times as many reads at the phylum level for the Rincón and Mayagüez samples, respectively (**Fig. 1**, **Table S3**, **Fig. S5–S6**). At the class taxonomic level *Kraken2* assigned 5.6 and 4.1 times as many reads as second-best method for Rincón and Mayagüez samples, respectively (**Fig. 1**, **Table S3**, **Fig. S5–S6**). The other two methods that utilize the same nucleotide database (*Graphmap* and *MiniMap2*) were roughly equivalent in the number of reads classified to phyla or class (**Fig. 1**).

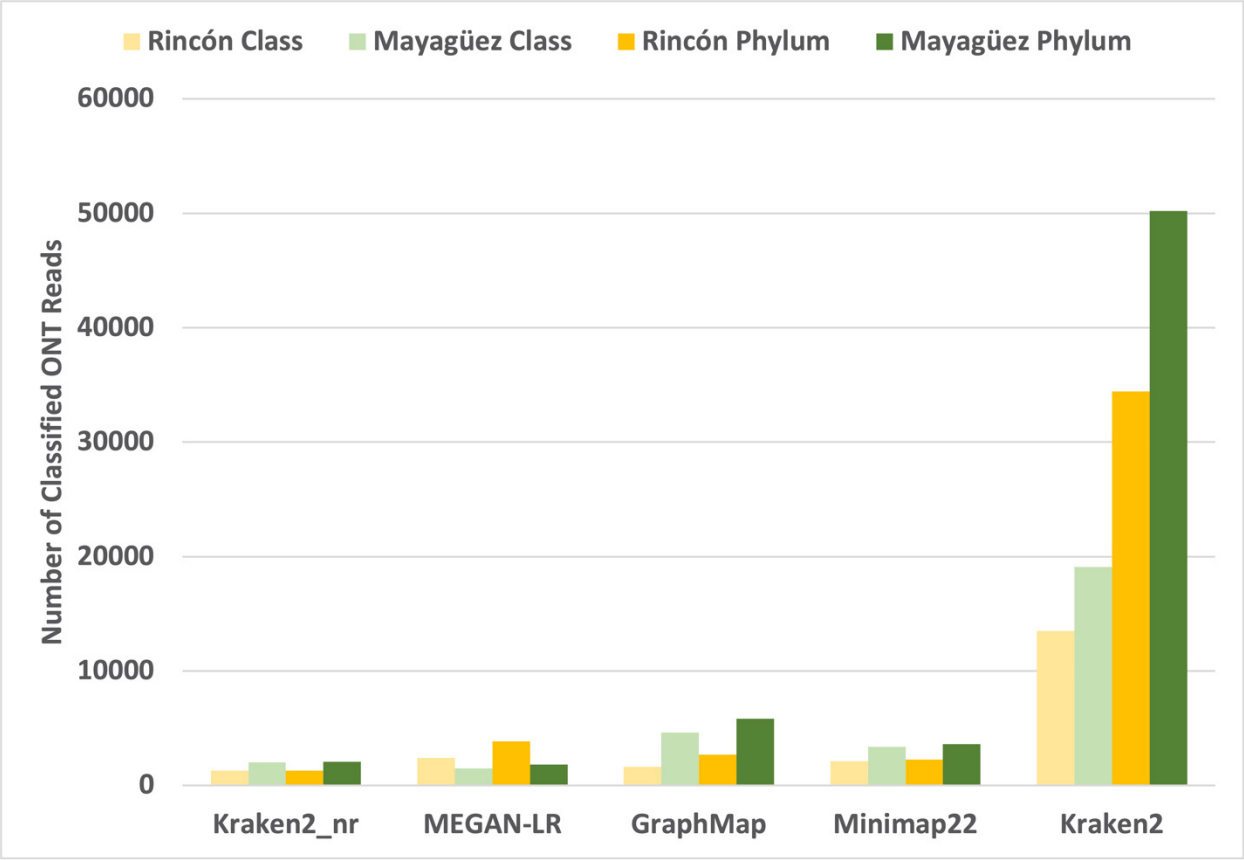

Figure 1 Phylum and class taxonomic level classification of ONT reads by method as absolute counts.

Between the two protein-based classification methods *Kraken2* in protein mode and *MEGAN-LR* assigned a phylum and class taxonomic ranking to a similar number of ONT reads (**Fig. 1**, **Fig. S1, S6**). Without a large sample size and no dramatic difference between the two methods run on the NCBI *nr* database it is difficult to draw a distinction between their efficacy at this stage.

Read Length by Method: The read lengths that each method assigned were all similar, although *MEGAN-LR* tended to assign taxonomy to longer reads (**Fig. S7**) and *GraphMap* had the most variance in read length (**Fig. S7**).

Quantification of biodiversity and shared taxonomy statistics between methods

Among methods utilizing nucleotide databases *Kraken2* classified the most taxonomic diversity with 50 and 41 phyla and 86 and 84 class taxonomy assignments in the Mayagüez and Rincón samples, respectively. The method that classified the most taxa in terms of diversity of classification for protein-based methods was *MEGAN-LR* with 45 and 39 phyla and 64 and 48 class level taxonomic assignments for Mayagüez and Rincón samples, respectively.

Via the Sørenson index, between all nucleotide methods there was at most a 78% and 62% overlap of phylum taxonomic assignment between *Kraken2* and *Minimap2* for the Rincón and Mayagüez samples via the Sørenson index (**Table 1**). Interestingly there was more similarity at the class taxonomic level via the Sørenson index between assignment methods *Kraken2* and *MEGAN-LR* despite using different databases than between *MEGAN-LR* and *Kraken2* protein mode that both use the NCBI *nr* database (**Table 1**).

The results from the Morisita-Horn index indicate that taxonomic assignments were very different from one another by method. Using the Morisita-Horn index to compare reads assigned

to each taxonomy between nucleotide methods, *Kraken2* and *Minimap2*, at the class level for Rincón and Mayagüez samples with 49% and 44% similarity, respectively (**Table 1**). There were many method pairs that had 1% or less overlap in proportional diversity as measured by the Morisita-Horn index indicating that each method's taxonomic assignments were quite different from one another, this can also be visualized in the stacked bar plots comparing each method (**Table 1** and **Fig. S1–S2**). Between protein methods, *Kraken2* in protein mode and *MEGAN-LR*, reads mapped to very different taxa at both phylum and class taxonomic levels for both samples (**Table 1**). A complete list of taxonomy assignments at the phylum and class taxonomy levels can be found in **Table S4–S7**.

Table 1 Shared species statistics between different ONT read annotation methods.

Rincon Phylum Morisita-Horn Similiarity Index						Rincon Phylum Sørenson Similiarity Index					
	kraken2_nr	meganLR	graphmap	minimap	kraken2		kraken2_nr	meganLR	graphmap	minimap	kraken2
kraken2_nr	1	0.106	0.029	0.122	0.009	kraken2_nr	1	0.203	0.463	0.607	0.579
meganLR		1	0.008	0.128	0.01	meganLR		1	0.163	0.4	0.504
graphmap			1	0.088	0.101	graphmap			1	0.608	0.422
minimap				1	0.043	minimap				1	0.775
kraken2					1	kraken2					1
Rincon Class Morisita-Horn Similiarity Index						Rincon Class Sørenson Similiarity Index					
	kraken2_nr	meganLR	graphmap	minimap	kraken2		kraken2_nr	meganLR	graphmap	minimap	kraken2
kraken2_nr	1	0.078	0.014	0.043	0.04	kraken2_nr	1	0.216	0.299	0.215	0.474
meganLR		1	0.005	0.089	0.056	meganLR		1	0.231	0.557	0.605
graphmap			1	0.033	0.055	graphmap			1	0.382	0.42
minimap				1	0.485	minimap				1	0.676
kraken2					1	kraken2					1
Mayagüez Phylum Morisita-Horn Similiarity Index						Mayagüez Phylum Sørenson Similiarity Index					
	kraken2_nr	meganLR	graphmap	minimap	kraken2		kraken2_nr	meganLR	graphmap	minimap	kraken2
kraken2_nr	1	0.426	0.009	0.157	0.007	kraken2_nr	1	0.401	0.503	0.594	0.65
meganLR		1	0.053	0.352	0.009	meganLR		1	0.261	0.381	0.541
graphmap			1	0.146	0.085	graphmap			1	0.443	0.365
minimap				1	0.053	minimap				1	0.617
kraken2					1	kraken2					1
Mayagüez Class Morisita-Horn Similiarity Index						Mayagüez Class Sørenson Similiarity Index					
	kraken2_nr	meganLR	graphmap	minimap	kraken2		kraken2_nr	meganLR	graphmap	minimap	kraken2
kraken2_nr	1	0.266	0.002	0.059	0.022	kraken2_nr	1	0.315	0.339	0.371	0.378
meganLR		1	0.009	0.03	0.01	meganLR		1	0.249	0.497	0.619
graphmap			1	0.106	0.079	graphmap			1	0.465	0.285
minimap				1	0.439	minimap				1	0.654
kraken2					1	kraken2					1

In terms of relative efficiency and completeness of taxonomic assignment only *Kraken2* appeared to classify enough reads at the Phylum or Class taxonomic level (**Fig. 2–3**) to have the rarefaction curve begin to plateau. For example, at the phylum level Hill number rarefaction curve plateaus for the Mayagüez and Rincón samples only for the *Kraken2* nucleotide-based method (**Fig. 2–3**). At the class taxonomic level, the rarefaction curve nearly plateaus for *Kraken2* and does not even begin to plateau for any of the other methods (**Fig. 2–3**). These results indicate that only *Kraken2* run in nucleotide mode was able to assign sufficient number of reads to taxonomic groupings in order to fully classify the diversity with any degree of completeness.

DISCUSSION

The benchmarking of these five methods would indicate that gut microbiota of *A. monolicornus* is a hyper-diverse community posing significant challenges to assessing its metagenomic composition. The only other paper on NGS sequencing of a millipede (*Telodeinopus aoutii* (Demange, 1971)) used RNA-seq Illumina based approach [29]. Despite employing a different methodology they also had similar results in the phyla level taxonomic diversity with more than 26 phyla reported [29]. Here we report nearly double that number. One caveat is that the millipedes sampled in this study were collected from their natural habitat whereas the other report contained millipedes from captivity, as well as being from an entirely different taxonomic orders of millipedes [29]. It should also be noted that we used a different reference database in this study which may also contribute to a different number of phyla.

Among methods for metagenomic filtering tested here *Kraken2* run in nucleotide mode was the best overall method for metagenomic skimming such a diverse community of microbiota. While *MEGAN-LR* assigned a similar number of taxonomic groups, and was the best protein-based method, the relative number of reads assigned to each group failed to plateau in the rarefaction analyses reducing confidence in this method. However, *MEGAN-LR* was quite efficient at classifying taxonomic

groups and this method would probably yield robust results with deeper sequencing efforts. *MEGAN-LR* and *Kraken2* were the two programs assessed in this study that were specifically designed to skim novel communities using long read technology. The other three programs while built with ONT reads in mind, were primarily designed to map back to assemblies with great accuracy and speed. Most benchmarking studies that assess performance of ONT read based mapping programs evaluate their performance using a database that contain genomes identical to those found in the query sample. The advantage of using a known reference is that performance can be evaluated with a high level of accuracy and statistical support. However, such approaches miss the challenges of mapping reads obtained from novel biological communities back to existing databases, such as the millipede gut microbiome. Here we turned to statistics normally used to evaluate different ecological communities and treated each method as if it was its own community replicated by two samples. The methods employed here in benchmarking does give us a very good qualitative idea about how well each of these different methods will perform on a novel community that has few close relatives in existing databases.

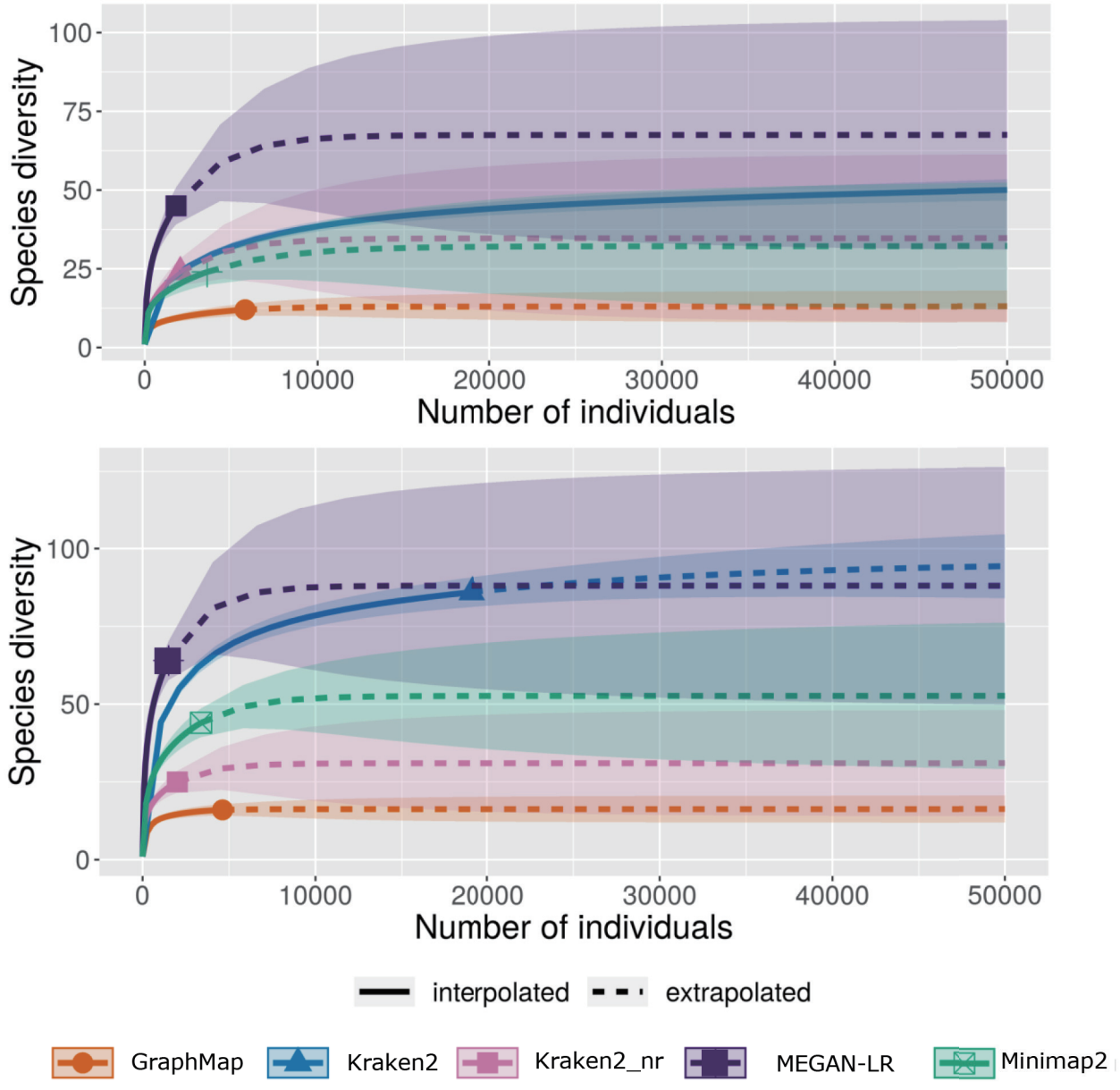

Figure 2 Hill number rarefaction accumulation curves implemented via *iNext* for Mayagüez *Anadenobolus monilicornis* gut microbiota samples sequenced via Oxford Nanopore MinION sequencer. Top panel is for rarefaction of phylum and bottom panel for class taxonomic classification.

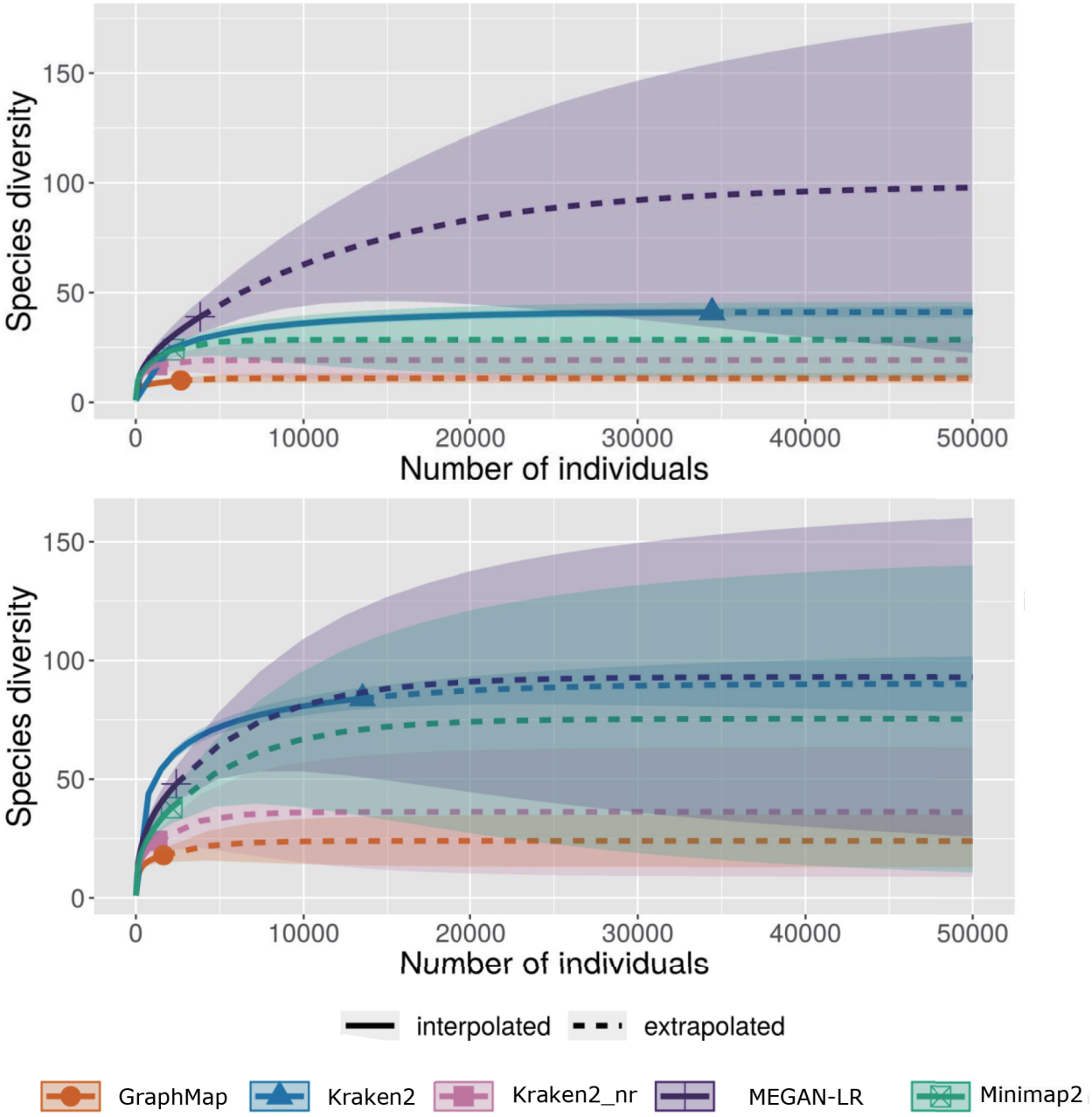

Figure 3 Hill number rarefaction accumulation curves implemented in *iNext* for Rincón *Anadenobolus monilicornis* gut microbiota samples sequenced via Oxford Nanopore MinION sequencer. Top panel is for rarefaction of phylum and bottom panel for class taxonomic classification.

In this study, we demonstrated evidence that the number of reads classified by each method affects both the number of taxa assigned but also our confidence in a binning method's ability to fully delimit a novel hyper-diverse community via ONT sequencing. The summary of our results would indicate that for novel biologically diverse samples *Kraken2* is the best overall method to evaluate taxonomic diversity using ONT based metagenomic skimming. The other insight is that the complexity of this community is such that it may not be appropriate to evaluate the taxonomic richness with ONT based metagenomic skimming using the sequencing depth of the present study. ONT based meta-barcoding would be an appropriate alternative or simply more ONT sequencing depth of each sample. These results would indicate that in general deeper sequencing is also required to fully capture the taxonomic complexity of the millipede gut microbiota community via metagenomic skimming in combination with *Kraken2* read classification, benchmarked here as the best performing method.

In closing the present study along with that of Sardar *et al*. [29] indicate that the millipede microbiome is a hyper-diverse system.

The milliped microbiome is an understudied studied system compared to its importance in the formation of leaf litter and the carbon cycle. Future work should identify if the gut symbionts are adventitious based on the environmental conditions and finally if different clades of millipedes contain host specific symbionts.

Author Contributions

Project conceiving: A.V.D.
Laboratory work: O.G.C., A.V.D.
Resources contribution: A.V.D. and M.J.C.
Bioinformatics analyses: A.V.D. and O.G.C.
Assisting in Bridges troubleshooting efforts: A.R.
Writing the manuscript: all authors

ACKNOWLEDGEMENT

Authors thank PSC staff Tom Maiden, Rick Costa, and Roberto Gomez for their help with installing albacore on the Bridges system. Laboratory work was funded by a NIH PR-INBRE Grant Contract #5P20GM102475, and the bioinformatics was funded by an NSF-XSEDE Grant TG-BIO170059 to A.V.D. Publications costs were supported by USDA-NIFA-HSI Grant #1016816 to A.V.D.

References

1. Rachtman E, Balaban M, Bafna V, Mirarab S. The impact of contaminants on the accuracy of genome skimming and the effectiveness of exclusion read filters. Mol Ecol Resour. 2020 May;20(3). https://doi.org/10.1111/1755-0998.13135. Epub 2020 Feb 4. PMID: 31943790.
2. Tedersoo L, Tooming-Klunderud A, Anslan S. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol. 2018 Feb 1;217(3):1370–85.
3. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46.
4. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. 2016 Sep 15;7(1):11307.
5. Jain C, Dilthey A, Koren S, Aluru S, Phillippy AM. A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases. J Comput Biol. 2018 Jul 1;25(7):766–79.
6. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016 Jul 15;32(14):2103–10.
7. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
8. Huson DH, Albrecht B, Ba ˘ Gcı C, Bessarab I, Górska A, Jolic D, et al. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biol Direct. 2018;13:6.
9. Vélez M. Los Gongolies, Gungulenes o Milpiés (Clase Diplopoda). In: In Joglar, R, Santos-Flores, C & Torres-Pérez, J Biodiversidad de Puerto Rico: Invertebrados. 2014. p. 240–53.
10. Oxford Nanopore Technologies. New basecaller now performs `raw basecalling', for improved sequencing accuracy [Internet]. 2017 [cited 2017 Sep 20]. Available from: https://nanoporetech.com/about-us/news/new-basecaller-now-performs-raw-basecalling-improved-sequencing-accuracy.
11. Oxford Nanopore Technologies. Albacore basecaller from Oxford Nanopore. Available from: https://community.nanoporetech.com. Accessed June 7, 2023.
12. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Berger B, editor. Bioinformatics. 2018 Aug 1;34(15):2666–9.
13. Schultz D. Pauvre: QC and genome browser plotting Oxford Nanopore and PacBio long reads.. Accessed Nov 11, 2018. https://github.com/conchoecia/pauvre.
14. Kiełbasa S, Wan R, Sato K, Horton P, Frith M. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21(3):487–93.
15. Firth MC. last-rna/last-long-reads.md at master · mcfrith/last-rna · GitHub [Internet]. [cited 2019 Nov 11]. Available from: https://github.com/mcfrith/last-rna/blob/master/last-long-reads.md. Accessed June 7, 2023.
16. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257. https://doi.org/10.1186/s13059-019-1891-0. Accessed June 7, 2023.
17. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. F1000Research. 2018;7:1755.
18. Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB. MinION ™ nanopore sequencing of environmental metagenomes: A synthetic approach. GigaScience. 2017 Mar 1;6(3):1-10. https://doi.org/10.1093/gigascience/gix007. PMID: 28327976; PMCID: PMC5467020.
19. Hernandez RJJ, Virella CR, Cafaro MJ. First survey of arthropod gut fungi and associates from Vieques , Puerto Rico. Mycologia. 2009;101(6):103–896.
20. Kenny NJ, Shen X, Chan TTH, Wong NWY, Chan TF, Chu KH, et al. Genome of the Rusty Millipede, Trigoniulus corallinus, Illuminates Diplopod, Myriapod, and Arthropod Evolution. Genome Biol Evol. 2015 Apr 21;7(5):1280-95. https://doi.org/10.1093/gbe/evv070. PMID: 25900922; PMCID: PMC4453065.
21. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 May 21;18(5):821–9.
22. Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, et al. WormBase: a modern Model Organism Information Resource. Nucleic Acids Res. 2020 Jan 8;48(D1):D762-D767. https://doi.org/10.1093/nar/gkz920. PMID: 31642470; PMCID: PMC7145598.
23. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006 Jun;13(5):1028–40.
24. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
25. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. Jvenn: An interactive Venn diagram viewer. BMC Bioinformatics. 2014 Aug 29;15(1):293.
26. Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). McInerny

G, editor. Methods Ecol Evol. 2016 Dec 6;7(12):1451–6.

27. Chao A, Ma KH, Hsieh TC, Chiu CH. Online Program SpadeR (Species-richness Prediction And Diversity Estimation in R). 2015. p. http://chao.stat.nthu.edu.tw/wordpress/software_do.
28. Magurran AE. Measuring Biological Diversity. Oxford: Blackwell Science Ltd; 2004. 256 p.
29. Sardar P, Šustr V, Chroňáková A, Lorenc F, Faktorová L. De novo metatranscriptomic exploration of gene function in the millipede holobiont. Sci Rep. 2022 Sep 28;12(1):16173.

Supplementary information

Supplementary information of this article can be found online at https://polscientific.com/jbm/index.php/jbm/article/view/376/485.