Abstract #3050
6/11/2022 (MBP11)

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

Daniel Portik[1], C. Titus Brown[2], N. Tessa Pierce-Ward[2]
1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025; **2.** Department of Population Health and Reproduction, University of California, Davis, CA 95616

Introduction

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 main goal of taxonomic profiling is to identify the species present in a microbiome sample (typically bacteria, archaea, fungi, viruses) and their relative abundances (Fig. 1). The development of long-read specific tools for taxonomic profiling is accelerating, yet there is a lack of consensus regarding their relative performance. We performed a critical benchmarking study using five long-read methods and four popular short-read methods[1]. We applied these tools to several mock community datasets generated using Pacific Biosciences (PacBio) HiFi sequencing or Oxford Nanopore Technology (ONT) sequencing, and Illumina data.

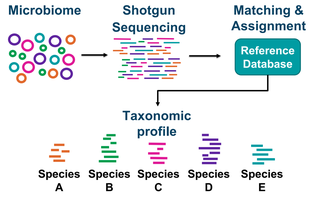

Figure 1. Taxonomic profiling overview. Metagenomics involves the sequencing of DNA extracted from a microbiome sample. Read-based profiling can then be performed, which requires aligning or matching reads to a reference database. The references can be kmers, nucleotide sequences, or protein sequences. Reads with reference matches are then assigned to a taxonomy based on various algorithms, and relative abundances can be calculated based on the number of reads assigned to each taxon.

Experimental design

Mock community datasets

We obtained four publicly available datasets for three mock communities (two with PacBio HiFi reads, two ONT)[1]. The mock communities differed in complexity (species and abundance design). We included Illumina data for two of the mock communities.

ZymoBIOMICS D6300
- **10 species, even**
- ONT R10.3
- ONT "Q20"
- Illumina

ZymoBIOMICS D6331
- **17 species, staggered**
- PacBio HiFi

ATCC MSA-1003
- **20 species, staggered**
- PacBio HiFi
- Illumina

The four long-read datasets differed in read lengths and quality scores (Fig. 2), with PacBio HiFi reads displaying higher accuracy (>99.95%) and longer mean lengths.

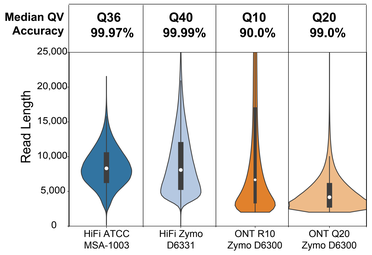

Figure 2. Dataset characteristics. Violin plots showing the distribution of read lengths, and associated median QV and accuracy scores. The two ONT datasets were previously filtered to remove all sequences <2 kb, which we found to have a strongly negative impact on profiling analyses.

Profiling methods

We evaluated five long-read (LR) methods and four popular short-read (SR) methods, which cover several combinations of matching and assignment algorithms (Fig. 3).

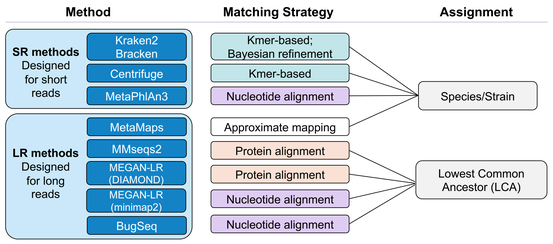

Figure 3. Profiling methods. An overview of the profiling methods tested, showing the different combinations of matching/alignment strategies and read assignment algorithms.

Comparative analysis

We evaluated the performance of each method based on the following categories.

Read utilization
- How many reads were assigned, and to which taxonomic ranks?

Precision, recall, and F-scores
- Precision = 1: *only* detected species in community; <1: detected false positives
- Recall = 1: detected *all* species in community; <1: failed to detect some species

Relative abundance
- Pass or fail a chi-squared goodness of fit to the theoretical abundances

Results: read utilization

- SR methods generally assign more reads (Fig. 4)
- Several LR methods show clear effects of the LCA algorithm
- Assignment is higher for HiFi reads (80%) vs. ONT data (60%) for LR methods

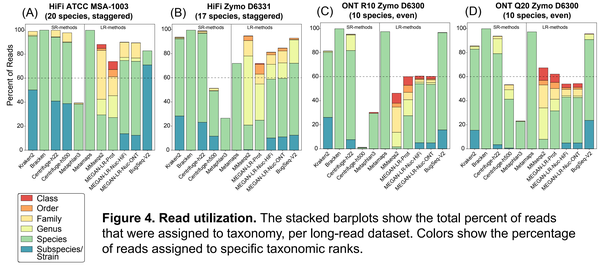

Figure 4. Read utilization. The stacked barplots show the total percent of reads that were assigned to taxonomy, per long-read dataset. Colors show the percentage of reads assigned to specific taxonomic ranks.

Results: precision, recall, F-scores

- SR methods display low precision, high recall and low F-scores (Fig. 5)
- Several LR-methods display high precision, moderate recall, and high F-scores

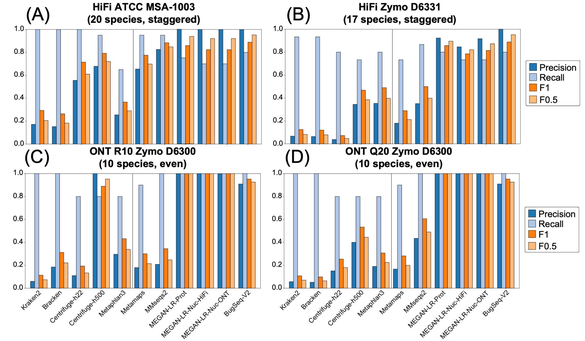

Figure 5. Detection results. Precision, recall, and F-scores are shown for the four long-read datasets.

Results: relative abundance

- Few methods passed the goodness of fit tests (Fig. 6)
- DIAMOND & MEGAN-LR[2,3] and BugSeq[4] had the highest accuracy

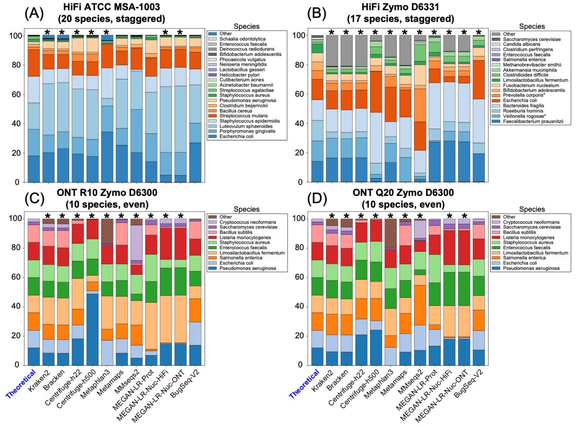

Figure 6. Relative abundances. Theoretical distributions are shown on the left. Read counts for false positives were grouped into the "Other" category. Asterisks signify methods that failed the GOF test.

Results: comparison to short reads

SR methods display the same characteristics with short-read (Illumina) datasets:
- Low precision, high recall, low F-scores, inaccurate relative abundances (Fig. 7)

This suboptimal performance occurs with long-read *and* short-read datasets.

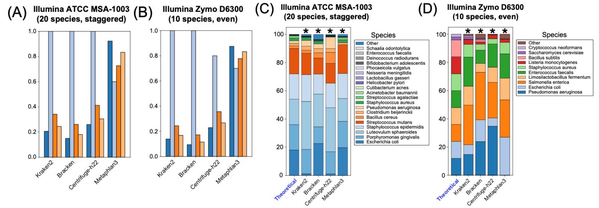

Figure 7. Short read results. Precision, recall and F-scores (dark blue, light blue, orange) for the short-read datasets (A, B). Relative abundance estimates for the short-read datasets (C, D).

Conclusions

We identified two methods that performed best for long-read datasets
- **DIAMOND & MEGAN-LR**[2,3]
  - PacBio workflow available on github: PacificBiosciences/pb-metagenomics-tools
  - A protein alignment method which also performs functional profiling simultaneously
- **BugSeq**[4]
  - Cloud platform with online submission: https://bugseq.com
  - A fast and highly accurate method based on nucleotide alignments

The top performing methods shared several key characteristics.
- Use full nucleotide or protein alignments
- Use last common ancestor algorithm
- Use minimum threshold-filtering for hits

Differences in long read quality have a clear effect on performance.
- Higher accuracy reads (PacBio HiFi sequencing) perform better with methods using protein alignments or exact kmer matching
- Large proportions of shorter reads (<2 kb) negatively impact analysis – filter out!

Long reads provide clear advantages over short reads for metagenomics.
- *Any* long-read dataset analyzed with a LR method performed better than a comparable short-read dataset – SR methods are fundamentally limited
- Simultaneous improvements in metagenome assembly show value of long reads

References

1. Portik DM, et al. (2021). Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets. *bioRxiv*, doi: 10.1101/2022.01.31.478527
2. Buchfink B, et al. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60.
3. Huson DH, et al. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, 13, 6.
4. Fan J, et al. (2021). BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics*, 2021, 1–3.