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 taxonomic profiling, where the main goal is to identify the species present in a microbiome sample (typically bacteria, archaea, fungi, viruses) and their relative abundances. 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 PacBio HiFi sequencing or Oxford Nanopore Technology (ONT) sequencing, and Illumina data.

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 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

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. 1).

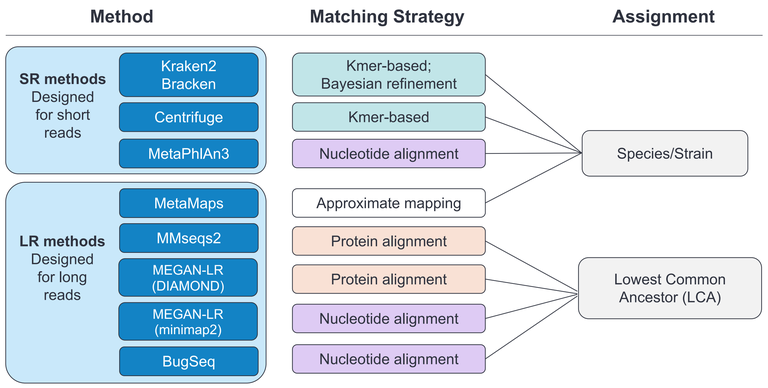

Figure 1. Profiling methods. An overview of the profiling methods tested, showing the different combinations of matching/alignment strategies and read assignment algorithms.

Comparative analysis

We evaluated performance based on the following categories.

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

Precision, recall, and F-scores
- Precision = 1: *only* detected species in community
- Recall = 1: detected *all* species in community

Relative abundance
- Pass/fail chi-squared goodness of fit to theoretical abundances

Results: read utilization

- SR methods generally assign more reads (Fig. 2)
- Several LR methods show clear effects of the LCA algorithm
- Assignment is higher for HiFi reads (80%) vs. ONT data (60%) for LR methods

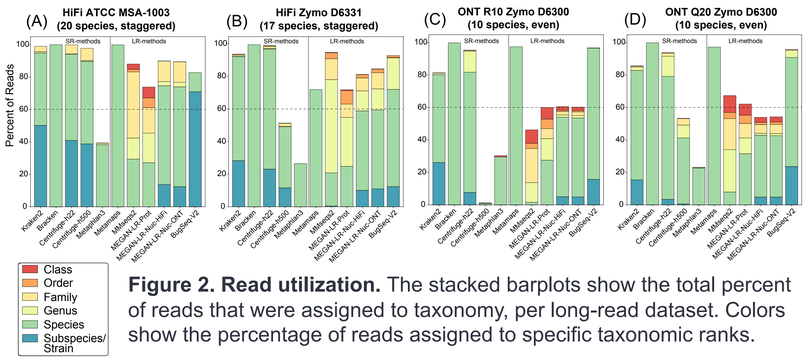

Figure 2. Read utilization. The stacked barplots show the total percent of reads that were assigned to taxonomy, per long-read dataset. Colors show the percentage of reads assigned to specific taxonomic ranks.

Results: precision, recall, F-scores

- SR methods display low precision, high recall and low F-scores (Fig. 3)
- Several LR-methods display high precision, moderate recall, and high F-scores

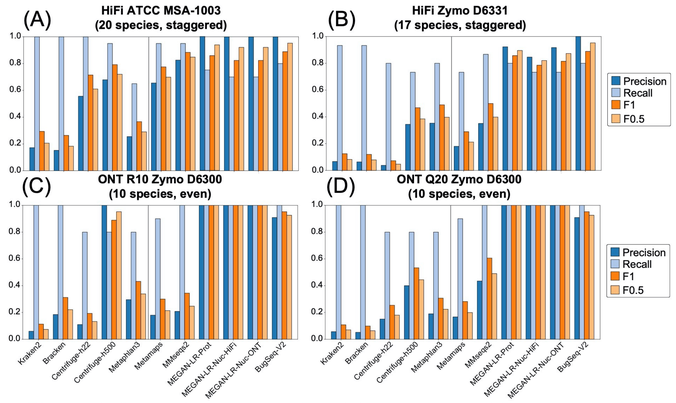

Figure 3. Detection results. Precision, recall, and F-scores are shown for the four long-read datasets.

Results: relative abundance

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

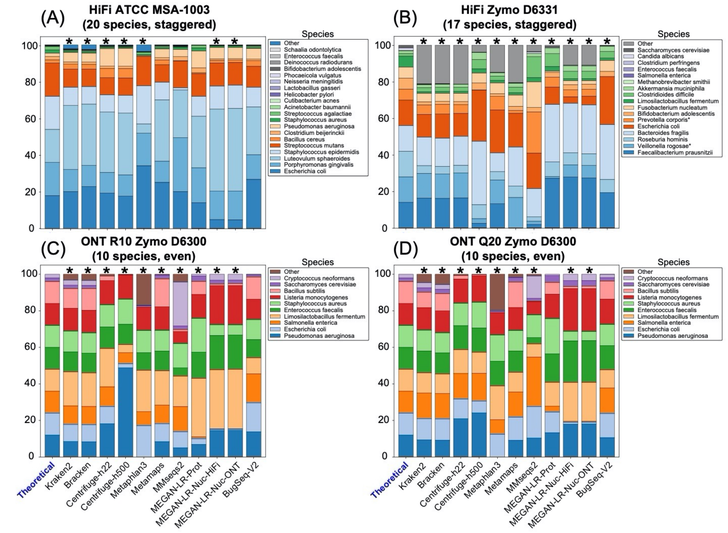

Figure 4. Relative abundances. Theoretical distributions are shown on the left. Read counts for false positives were grouped into the "Other" category. Asterisks signify methods that failed the GOF test.

Conclusions

Two methods that performed best for long-read datasets
- **DIAMOND & MEGAN-LR**[2,3]
  - PacBio github: PacificBiosciences/pb-metagenomics-tools
- **BugSeq**[4]
  - Cloud platform with online submission: https://bugseq.com

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

Differences in read quality have an effect on performance.
- Higher accuracy reads (PacBio) perform better with methods using protein alignments or exact kmer matching
- Shorter reads (<2 kb) negatively impact analysis – filter out!

Long reads perform better than short reads.
- *Any* long-read dataset analyzed with a LR method performed better than a comparable short-read dataset – SR methods are limited

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.