MBP-662
06/17/2023

Taxonomic classification and profiling of long-read shotgun metagenomics datasets: best practices and new pipelines

Daniel J. Nasko[1], Jeremy E. Wilkinson[1], C. Titus Brown[2], N. Tessa Pierce-Ward[2], and Daniel M. Portik[1]
1. PacBio, 1305 O'Brien Drive, Menlo Park, California USA; 2. Department of Population Health and Reproduction, University of California Davis, Davis, California USA

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 highly 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.

We recently published a benchmarking study of several taxonomic profiling/classification methods for long-read datasets[1]. Here, we outline the experimental design and key findings of our study. To improve accessibility to top-performing tools, we also developed comprehensive workflows for 1) sourmash[2-4] and 2) Diamond & MEGAN-LR[5,6] and describe them here.

Methods

Mock community datasets

We obtained four publicly available datasets for three mock communities (two with PacBio HiFi reads, two ONT). 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

Classification and profiling methods

We evaluated five long-read (LR) methods, five popular short-read (SR) methods and one generalized method, which cover many combinations of matching/assignment algorithms (Fig. 1).

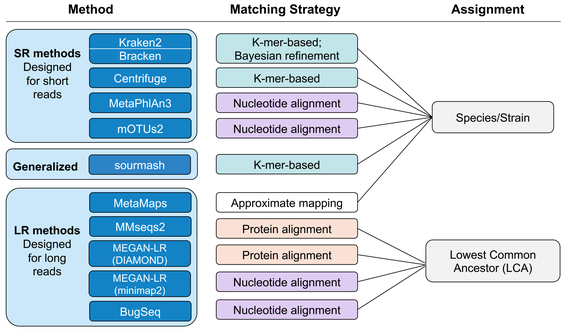


Figure 1. An overview of the taxonomic classification and profiling methods tested, showing different combinations of matching/alignment strategies and read assignment algorithms.

Comparative analysis

We evaluated performance based on the following categories.

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

Precision and recall

- SR methods generally display low precision, high recall, and low F1 scores, due to high false positives (Figs. 2, 3).
- Several LR-methods display high precision, high to moderate recall, and high F1 scores, and rarely produce false positives (Figs. 2, 3).
- **Sourmash** had the highest precision and recall for HiFi data, with detection down to 0.001% relative abundance (Figs. 2, 3).

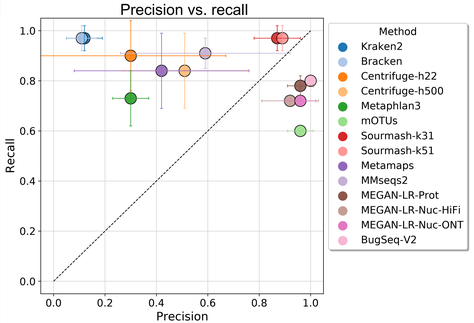


Figure 2. Average values (from two HiFi datasets) are shown for precision and recall based on the mock communities evaluated. Error bars around dots represent standard deviation.

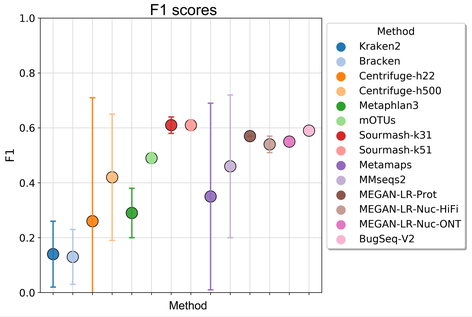


Figure 3. Average values (from two HiFi datasets) are shown for F1 scores based on the mock communities evaluated. Error bars around dots represent standard deviation.

Relative abundance

- Sourmash, DIAMOND & MEGAN-LR, and BugSeq generally had the highest accuracy, but results varied (Fig. 4).

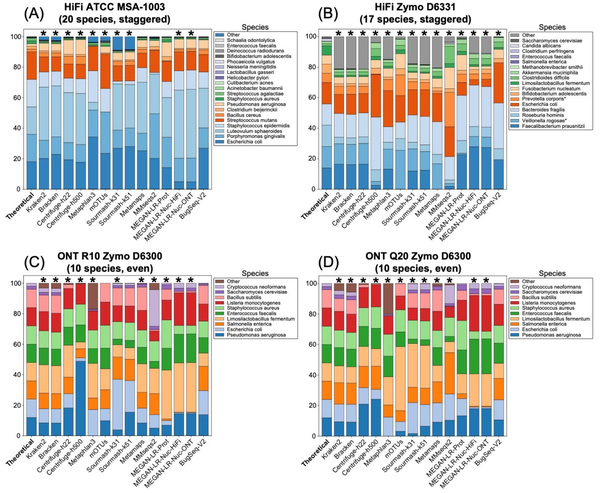


Figure 4. Theoretical distributions are shown on the left. Read counts for false positives were grouped into the "Other" category. Asterisks signify methods that failed the GOF test.

Conclusions

Three methods performed well for HiFi datasets.
- **Sourmash**[2-4] (workflow on PacBio github)
- **DIAMOND & MEGAN-LR**[5,6] (workflow on PacBio github)
- **BugSeq**[7] (Cloud platform: https://bugseq.com)

Differences in accuracy of reads influence performance.
- Higher accuracy reads (PacBio) perform better with methods using protein alignments or exact k-mer 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.

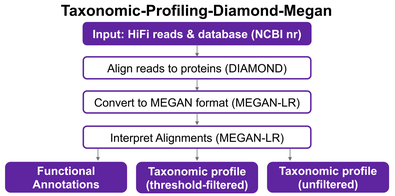
Workflows

- All PacBio metagenomics pipelines and tools are publicly available on github

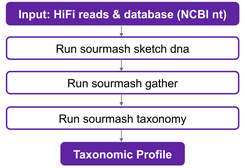
PacificBiosciences / **pb-metagenomics-tools**

- Pipelines implemented in snakemake, a Python-based workflow management system.
- Documentation available for all pipelines.
- See **datasets** page for 55+ publicly available HiFi metagenomic datasets from different sample types.

The **DIAMOND & MEGAN-LR** workflow uses DIAMOND for translation alignment of reads to a protein database (e.g., NCBI nr). MEGAN-LR is used to assign reads to taxonomy using an interval-union LCA algorithm. The pipeline outputs read-based classifications (taxonomic and functional) as well as a taxonomic profile (with and without threshold-filtering).



The **sourmash** workflow is a k-mer-based approach which runs three modules (sketch, gather, taxonomy). It differs from other k-mer methods by using combinatorial observations of k-mers to find the minimum set of reference genomes that cover all information (k-mers) in the metagenome query. Afterwards, it aggregates the taxonomic information from these genomes using an LCA approach and outputs a taxonomic profile.



References

1. Portik DM, et al. (2021). Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics*, 23: 541.
2. Brown CT & Irber L. (2016). Sourmash: a library for MinHash sketching of DNA. *Journal of Open Source Software*, 1: 27.
3. Pierce NT et al. (2019). Large-scale sequence comparisons with sourmash. *F1000 Research*, 8:1006.
4. Irber L et al. (2022). Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. *bioRxiv*, https://doi.org/10.1101/2022.01.11.475838
5. Buchfink B, et al. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60.
6. 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.
7. Fan J, et al. (2021). BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics*, 2021, 1–3.