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. 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.

Experimental design

Mock community datasets

We obtained four publicly available datasets for three mock communities (two with sequencing or Oxford Nanopore Technology (ONT) sequencing, and Illumina data). The mock community datasets generated using Pacific Biosciences (PacBio) HiFi profiling is accelerating, yet there is a lack of consensus regarding their relative advantages over short reads.

<table>
<thead>
<tr>
<th>Read Length</th>
<th>READ1</th>
<th>READ2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>R10</td>
<td>R10</td>
</tr>
<tr>
<td>15,000</td>
<td>R15</td>
<td>R15</td>
</tr>
<tr>
<td>25,000</td>
<td>R25</td>
<td>R25</td>
</tr>
</tbody>
</table>

Species: A, B, C, D, E

Figure 1. Taxonomic profiling overview. Metagenomics involves the sequencing of DNA extracted from a microbiome sample. Read-based profiling can be performed, which requires aligning or matching reads to references. The references can be species, species sequence, or species sequence, which is known as pairwise or whole-genome alignment. In this study, we used a database of species sequences and compared it to the sample read set. The percentages can be calculated based on the number of reads assigned to each taxon.

Results: relative abundance

The top performing methods shared several key characteristics.

• Use minimum threshold
• Use last common ancestor algorithm

The top performing methods are also highly accurate models based on nucleotide alignment.

Conclusions

We identified two methods that performed best for long-read datasets.

1. BugSeq
2. MEGAN-LR-Prot

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 nucleotide alignments.
• Longer proportions of shorter reads (LR) negatively impact analysis: filter out short reads

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 — LR methods are fundamentally limited.

Simultaneous improvements in metagenome assembly show value of long reads.

References