Abstract #3009
6/11/2022 (MBP11)

A new standard: high MAG recovery and precision species profiling of a pooled human gut microbiome reference using PacBio HiFi sequencing

Daniel Portik[1], Meredith Ashby[1], Siyuan Zhang[1], Kris Locken[2], Shuiquan Tang[2], Brett Farthing[2], Michael Weinsten[2], Martha Carlin[3], Raul Cano[3], **Jeremy Wilkinson[1]**
1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025; **2.** Zymo Research Corporation, 17062 Murphy Ave, Irvine, CA 92614; **3.** The BioCollective, LLC., 5650 Washington St, Denver CO 80216

Introduction

Advancements in sequencing technologies have made metagenomic analyses of complex microbial samples routine and accessible. Mock communities of known composition are often run in parallel to allow for accurate data evaluation and to facilitate cross-study and inter-lab comparisons, yet they lack the microbial diversity of real-world samples. The **ZymoBIOMICS Fecal Reference with TruMatrix Technology** (D6323) is a highly diverse pooled human gut microbiome standard that provides a truly complex alternative to mock communities. However, the microbial content of this standard is only partially characterized, and species level composition remains underexplored. Here, we explore the content of this sample using highly accurate long-read sequencing.

Methods

PacBio HiFi sequencing

We performed PacBio HiFi sequencing using four SMRT Cells (8M) on the Sequel IIe system. This resulted in 11.9 million HiFi reads with a mean length of 8.9 kb, for a total of 88.3 Gb of data. The HiFi read median QV was 43, representing >99.99% accuracy. We also downsampled the full dataset to investigate effects on analyses.

Metagenome assembly

- Assembly was performed using **hifiasm-meta**[1]
- The PacBio **HiFi-MAG-Pipeline** was used to identify high-quality metagenome assembled genomes (HQ-MAGs) using a **circular-aware binning** strategy (Fig. 1). This workflow is available on github: PacificBiosciences/pb-metagenomics-tools

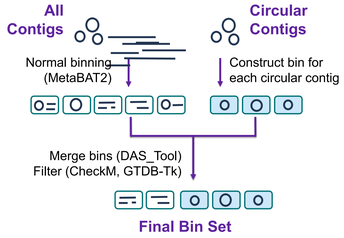

Figure 1. Circular aware binning. With typical binning, circular contigs are sometimes mis-binned with linear contigs. This inflates the numbers of single copy genes and the contamination score, leading to removal of the bin (and circular contig) during filtering. Here, in addition to typical binning, all circular contigs are detected and placed in individual bins. The bin sets are compared and merged into a non-redundant set, which effectively prevents mis-binning. Bins are filtered using completeness and contamination scores, and taxonomy is assigned.

Taxonomic and functional profiling

- Profiling was performed using **DIAMOND**[2] and **MEGAN-LR**[3] with the NCBI-nr protein database
- Analysis was automated with the PacBio **Taxonomic-Functional-Profiling-Protein** workflow, with long-read settings and filtering optimized for high precision species detection[4]

Metagenome assembly

Assembly with **hifiasm-meta** and evaluation with **HiFi-MAG-Pipeline** produced:

- ~2600 genome bins
- 199 total HQ-MAGs; 72 are single contig, 102 are >95% complete (Fig. 2)
- HQ-MAGs from 164 species, 114 genera, and 43 families
- 28 species represented by 2–3 MAGs (e.g., strain-level variation; Fig. 3)
- Predictable relationships between total data and MAG recovery (Fig. 4)

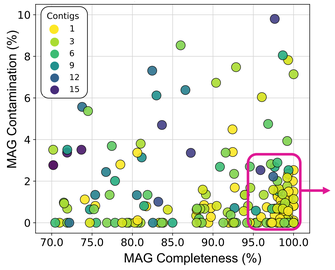

Figure 2. MAG characteristics. Completeness versus contamination scores for the 199 HQ-MAGs, as evaluated by CheckM. Each dot represents a MAG, and colors indicate the number of contigs the MAG contains. We found 102 HQ-MAGs (51%) displayed >95% completeness. Furthermore, 54 HQ-MAGs (27%) were >95% complete and composed of a single (and often circular) contig.

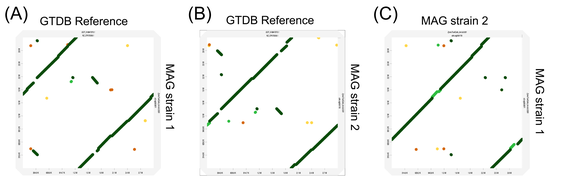

Figure 3. Strain variation. We assembled two highly complete MAGs for *Akkermansia muciniphila_B*. The D-GENIES dotplots show strain variation across comparisons, including between (A) the Genome Taxonomy Database (GTDB) reference and MAG 1, (B) GTDB and MAG 2, and (C) between the MAGs.

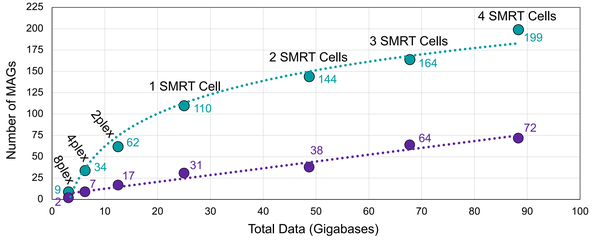

Figure 4. Total data vs. MAG recovery. Based on the downsampling, we found a logarithmic relationship between total data and HQ-MAGs (R^2 = 0.97), and a linear relationship between total data and single contig MAGs (R^2 = 0.96). With one SMRT Cell we recovered 110 HQ-MAGs (31 as single contigs), and even at 4plex level we recovered 34 HQ-MAGs (7 as single contigs).

Taxonomic profiling

Taxonomic profiling using **DIAMOND** and **MEGAN-LR** resulted in:

- Detection of 155 species (80 genera) in high precision mode (Fig. 5A)
- Detection of 7,184 species (~2,000 genera) in low precision mode (no threshold filtering; Fig. 5B)
- Consistent profiles across data levels using high precision mode (Figs. 5, 6)

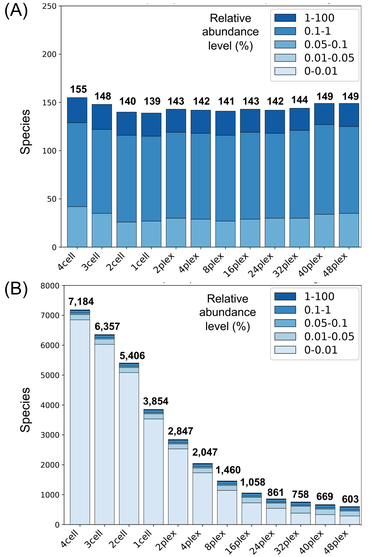

Figure 5. Species detection. A summary of the number of species detected across downsampled data levels. Colors indicate the number of species detected in each relative abundance category, and the total number of species is shown on top. The full dataset is shown on the left, with decreasing data levels to the right.

(A) Results for the high precision filtering mode, in which a minimum threshold of reads must be assigned to a particular species to report it. The lower limit for detection in this mode is ~0.03% relative abundance.

(B) Results for the low precision mode, in which no threshold filtering is applied. Here, species can be represented by a single matched read. Across data levels, nearly 95% of assignments are at the ultra-low (<0.01%) or very low (0.01–0.05%) abundance levels (85% and 10%, respectively).

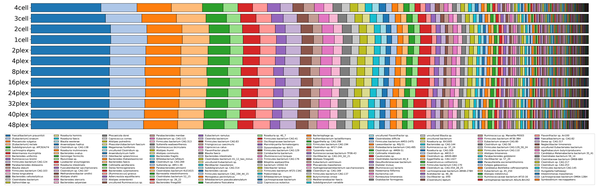

Figure 6. Relative abundance profiles. A comparison of the relative abundances of species across different data levels. Within a row, each color represents a distinct species and the width of the bar indicates its relative proportion in the community. The abundance profiles result from the high precision filtering mode and display high similarity across 0.5–88 Gb of data.

Functional profiling

Functional profiling using **DIAMOND** and **MEGAN-LR** resulted in:

- Approximately 92% of reads received at least one functional annotation
- Over 66.9 million functional annotations across all databases (Table 1)
- Clear decrease in number of unique classes with decreasing data levels (Fig. 7)

Database	Total annotations	Unique classes	Annotations per read (mean)
EC	13.1 million	2,714	2.3
eggNOG	11.3 million	2,714	3.2
InterPro2GO	25.1 million	17,428	4.1
SEED	17.4 million	722	2.2

Table 1. Functional annotations. A summary of functional annotations derived from the four databases used by MEGAN-LR, based on protein alignments inferred with DIAMOND.

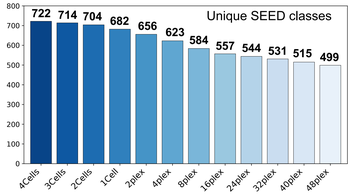

Figure 7. Total data vs. unique functional classes. A summary of the number of unique SEED classes assigned across reads, based on total data. At the lowest data level, 70% of the unique classes from the full dataset were detected. The trend was similar for the other three functional databases.

Conclusions

We used PacBio HiFi sequencing to obtain highly accurate long reads to characterize the **ZymoBIOMICS TruMatrix Fecal Reference,** resulting in:

- Assembly of 199 high-quality MAGs, including 54 assembled into single contigs and with >95% completeness.
- Detection of 155 species with high confidence (at >0.03% abundance), and up to 7,200 species without filtering (most at <0.01% abundance).
- Detection of 67 million functional annotations (from 12 million HiFi reads).
- Consistent taxonomic profiles across downsampled datasets (using high precision filtering), and predictable decreases in performance for metagenome assembly and functional profiling.
- Successful results even at the lowest data levels (0.5–3 Gb), including assembly of multiple MAGs and hundreds of thousands of functional annotations.

PacBio HiFi sequencing has emerged as a powerful and cost-effective approach for performing shotgun metagenomics studies.

References

1. Feng X, et al. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, Online Early, doi: 10.1038/s41592-022-01478-3
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. 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