High MAG recovery and precision species profiling of a pooled human gut microbiome reference using PacBio HiFi sequencing

Kirren Kaur[1], Jeremy E. Wilkinson[1], Meredith Ashby[1], Siyuan Zhang[1], Khi Pin Chua[1], Kris Locken[2], Shuiquan Tang[2], Brett Farthing[2], Michael Weinstein[2], Martha Carlin[3], Raul Cano[3], Kyle Langford[4], Benjamin Auch[4], Ivan Liachko[4], and Daniel M. Portik[1]
[1]Pacific Biosciences; [2]Zymo Research Corporation; [3]The BioCollective, LLC.; [4]Phase Genomics, Inc.

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 fecal reference 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

- Shotgun metagenomics:
  - Sequel IIe system: 4 SMRT Cell 8M, 11.9 million HiFi reads, mean length of ~8 kb, total of 88.3 Gb of data. HiFi read **median QV was 41**
    - Down-sampled to investigate effects on analyses and ran additional SMRT Cell 8M to test 48-plex barcoded samples
  - Revio system: 2 SMRT Cell 25M, 17.6 million HiFi reads, mean length of ~8 kb, total of 135.7 Gb of data. HiFi read **median QV was 46**
- Full-length 16S rRNA sequencing (V1-V9)

Metagenome taxonomic + functional profiling

- Profiling was performed using **DIAMOND**[1] and **MEGAN-LR**[2] with the NCBI-nr protein database utilizing the PacBio **Taxonomic-Functional-Profiling-Protein** workflow (available on github: PacificBiosciences/pb-metagenomics-tools), with long-read settings and filtering optimized for high precision species detection[3]

Full-length 16S rRNA taxonomic profiling

- Profiling was performed using **QIIME 2**[4] and **DADA2**[5] with the GTDB database utilizing the PacBio **Full-length 16S analysis** workflow (available on github: PacificBiosciences/pb-16S-nf)

Metagenome assembly

- Assembly was performed using **hifiasm-meta**[6]
- The PacBio **HiFi-MAG-Pipeline** was used to identify high-quality metagenome assembled genomes (HQ MAGs) using a **completeness-aware binning** strategy

Phase Genomics ProximeMeta

- PacBio HiFi assembly + ProximeMeta Hi-C binning

Metagenome functional profiling

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

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

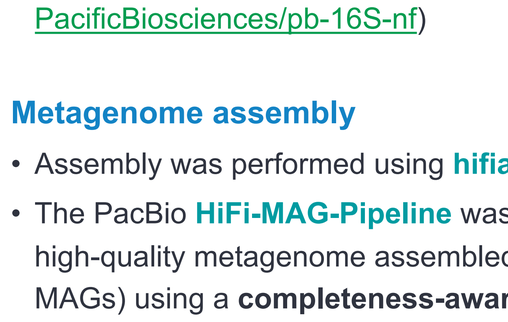

Figure 1. Total data vs. unique functional classes. A summary of the number of unique SEED classes assigned across reads, based on total data. At the lowest data level, 70% of the unique classes from the full dataset were detected. The trend was similar for the other three functional databases.

Metagenome taxonomic profiling

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

- Detection of 155 species (80 genera) in high precision mode (Fig. 2A)
- Detection of 7,184 species (~2,000 genera) with no filtering (Fig. 2B)
- Consistent profiles across data levels using high precision mode (Figs. 2, 3)

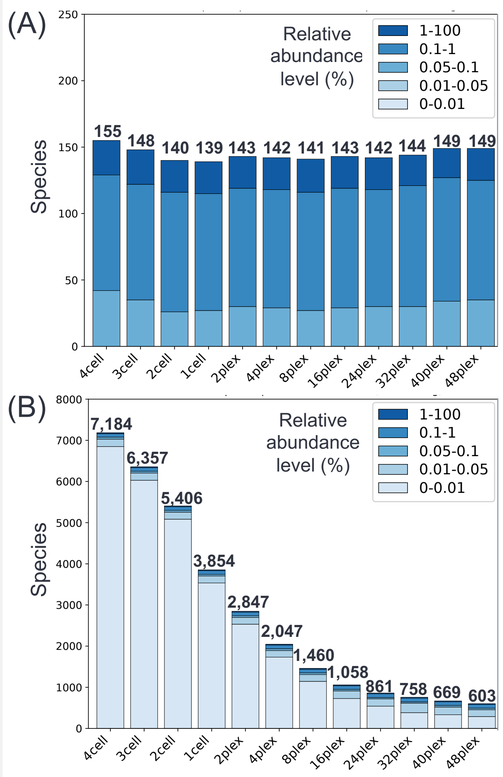

Figure 2. Species detection. Number of species detected across down-sampled data levels. Colors indicate the number of species detected in each relative abundance category, and the total number of species is shown on top.

(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 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).

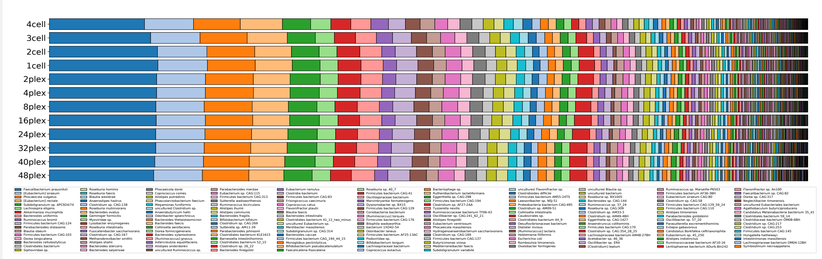

Figure 3. Relative abundances. A comparison of the relative abundances of species across different data levels using the high precision filtering mode. Within a row, each color represents a species, and the width of the bar indicates its relative proportion in the community. The abundances display high similarity across 0.5–88 Gb of data. The 48-plex barcoded run yielded similar results (not shown).

Full-length 16S rRNA taxonomic profiling

Taxonomic profiling using **QIIME 2** and **DADA2** resulted in:

- Detection of 205 species
- **140 species** classified in both dataset accounts for **88% of assigned 16S reads and 74% of assigned shotgun reads**
- Unique species are all <2%
- Spearman correlation is 0.85 at the species level (Fig. 4)
- Down-sampled 16S (denoised) reads to ~64k reads
- Shotgun data down-sampled to simulate 48-plex (~60k reads)

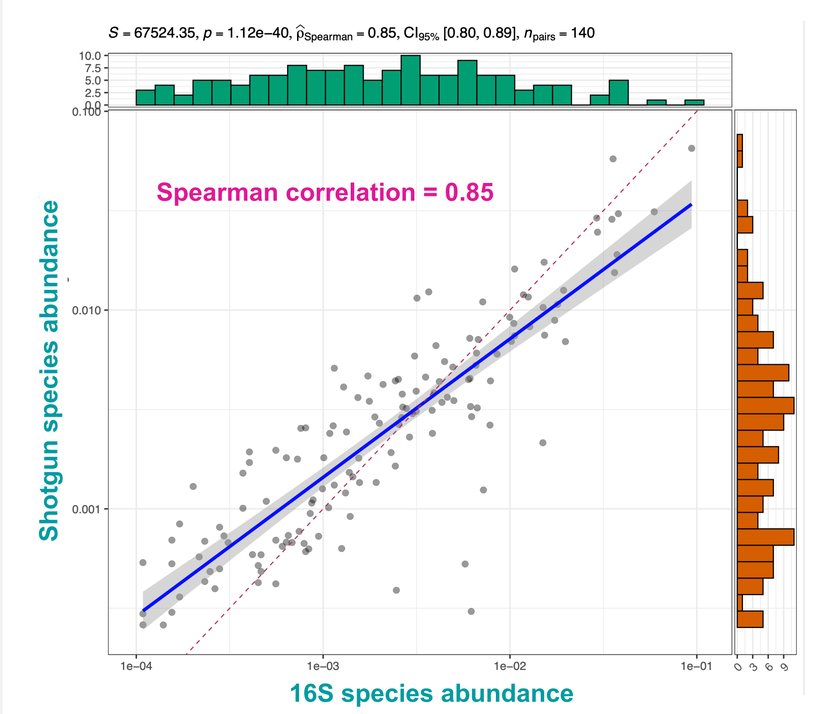

Figure 4. 16S species abundance vs shotgun species abundance. A correlation of the relative abundances of species between 16S and shotgun metagenomics classifications. At the species level, correlation is 0.85. At the genus level, correlation increases to 0.90, and account for 87% of assigned shotgun reads.

Metagenome assembly

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

- The assembly graph reveals many circular contigs were produced directly from hifiasm-meta, pre-binning (Fig. 5)
- 190 total HQ MAGs; 67 are single contig and >95% complete (reference quality)
- HQ MAGs from 154 species, 116 genera, and 42 families
- 19 species represented by 2–3 MAGs (strain-level variation)
- Higher throughput with the Revio system increases MAG recovery and reduces cost (Fig. 6)

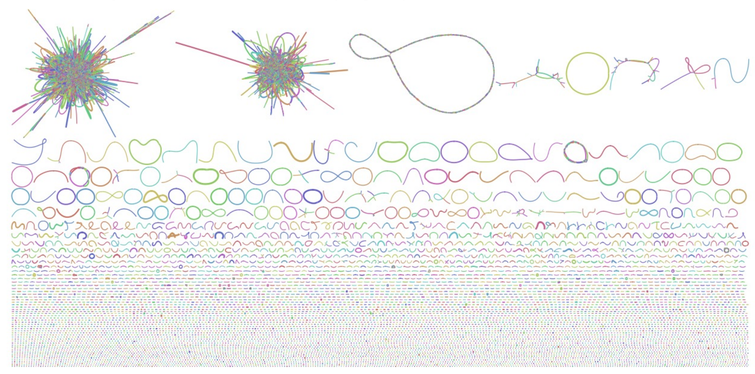

Figure 5. The hifiasm-meta assembly graph. Assembly graph depicting many circular contigs from the metagenome assembly of the HiFi data.

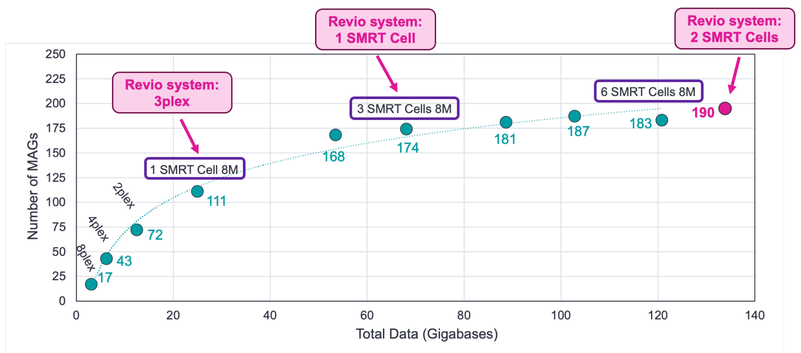

Figure 6. Effects of total data vs HQ MAG recovery. Sequel IIe SMRT Cell 8M in teal & purple; Revio SMRT Cell 25M in magenta.

HiFi + ProximeMeta metagenome deconvolution

HiFi assembly + binning with **ProximeMeta** results in:

- An increase in high-quality MAG yield by 44 – 64%
- Up to **313 HQ MAGs** recovered with Revio dataset (Fig. 7)
- Host associations: BGCs, AMR genes, viruses, plasmids, MGEs

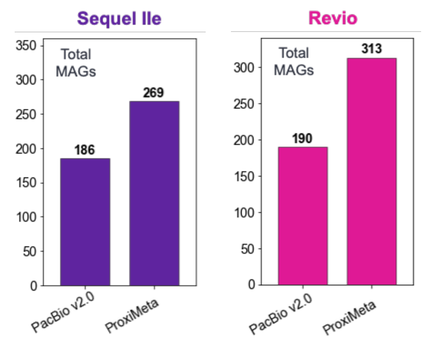

Figure 7. HQ MAG results for HiFi assembly + HiFi-MAG-Pipeline v2.0 binning vs HiFi assembly + Phase Genomics ProximeMeta binning. Sequel IIe in purple; Revio in magenta. First column in each is PacBio HiFi-MAG-Pipeline binning, 2nd column in each is ProximeMeta binning.

Summary

Metagenome taxonomic and functional profiling

- Detection of 155 species in high-precision filtered mode
- Up to 7,184 species detected without filtering
- 67 million functional annotations (from 12 million reads)
- Consistent taxonomic profiles across data levels

Full-length 16S rRNA taxonomic profiling

- High concordance with shotgun metagenome taxonomic profiles

Metagenome assembly

- Recovered 190 HQ MAGs, 67 are single contig and >95% complete
- With 1 SMRT Cell 8M, recovered 111 HQ MAGs; with 1 SMRT Cell 25M, recovered 174 HQ MAGs

PacBio HiFi + ProximeMeta Hi-C

- 313 high-quality MAGs
- Metagenome-host: BGCs, AMR genes, viruses, plasmids, MGEs

PacificBiosciences / **pb-metagenomics-tools**

References

1. Buchfink B, et al. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60.
2. 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.
3. Portik DM, et al. (2022). Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics*, 23:541.
4. Bolyen E, et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37, 852–857.
5. Callahan BJ, et al. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581–583.
6. Feng X, et al. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, 19, 671–674.