Long-read metagenome assembly produces hundreds of high-quality MAGs from different soil types

Daniel M. Portik[1], Luis E. Valentin-Alvarado[2,3], Jillian F. Banfield[2], Boyke Bunk[4], Jorg Overmann[4], and Jeremy E. Wilkinson[1]

1. PacBio, 1305 O'Brien Dr, Menlo Park, California 93025 USA;
2. Innovative Genomics Institute, University of California, Berkeley, California 94720 USA;
3. Department of Plant and Microbial Ecology, University of California, Berkeley, California USA;
4. Leibniz Institute, DSMZ-German Collection of Microorganisms and Cell Cultures GmbH, Inhoffenstraße 7B, 38124 Braunschweig, Germany

Introduction

Metagenome assembly has many challenges:

- the presence of multiple species
- uneven and unknown species abundances
- conserved genomic regions shared across species
- strain-level variation within species

PacBio HiFi sequencing produces highly accurate long reads (>Q20, >99% accuracy) which provide major advantages for metagenome assembly. New metagenome assembly algorithms have been developed specifically for HiFi reads, including **hifiasm-meta**[1] and **metaMDBG**[2]. These methods can reconstruct full metagenome-assembled genomes (**MAGs**) for many higher abundance species.

Metagenome assembly of soil has been historically difficult using short reads. The combination of high species diversity and ultra-low relative abundances poses a challenge and requires a higher sequencing depth to achieve success. Here, we demonstrate that the amount of HiFi data from the high-throughput **Revio system** is sufficient to assemble high-quality MAGs in complex microbiomes such as wetland soil.

Methods

PacBio HiFi sequencing on the Revio system

Wetland soil. A soil core was obtained from a northern California wetland and sampled along six depths. The six samples were prepped and sequenced on the Revio system using 3 SMRT Cells.

Grassland and woodland soil. Soil samples were obtained from two field sites in Germany. Libraries were prepped and each sample was sequenced using 4 SMRT Cells on the Revio system.

Soil type	HiFi reads (million)	Total data (gigabases)	Avg read length (kb)	Media n QV
Wetland	20.6 M	195 Gb	8.4 kb	Q41
Grassland	30.5 M	283 Gb	8.7 kb	Q40
Woodland	17.4 M	151 Gb	8.2 kb	Q42

Metagenome assembly and postprocessing

Full datasets. The combined datasets for each of the three soils were assembled using **hifiasm-meta** and **metaMDBG.** Each resulting contig set was processed using the PacBio **HiFi-MAG-Pipeline** (v2+)[3], which performs long-read-specific binning, QC, and taxonomic annotation steps. We evaluated the number of MAGs produced for different quality categories, including single-contig high-quality (SC-HQ) MAGs, high-quality (HQ) MAGs and medium-quality (MQ) MAGs. We show results from the best assembly method, per sample.

Downsampling. We downsampled the full datasets to produce several smaller datasets. We performed assembly and postprocessing as described above to obtain MAGs. We then assessed if the number of MAGs could be predicted by the total data (Gb) per sample, either by a linear relationship (by performing linear regression) or a non-linear relationship such as a saturation curve (by log-transforming the total data values before linear regression).

Results

HiFi metagenomics yields fully resolved MAGs

Large numbers of SC-HQ-MAGs were produced directly from the assembly step. These represent complete genomes and are often visible as circular contigs in the assembly graph (Fig. 1).

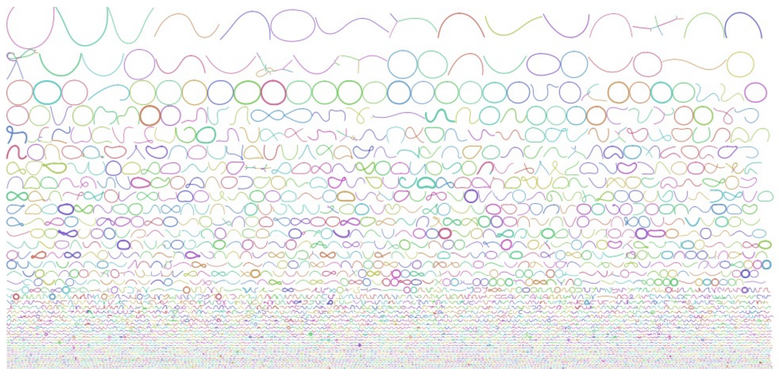

Figure 1. A partial view of the hifiasm-meta assembly graph for the wetland soil showing many large circular contigs (0.5–13 Mb).

Hundreds of HQ-MAGs assembled from soils

- The number HQ-MAGs produced per sample ranged from 133–563 (Fig. 2) and included 79–391 single-contig HQ-MAGs (e.g., the highest quality).

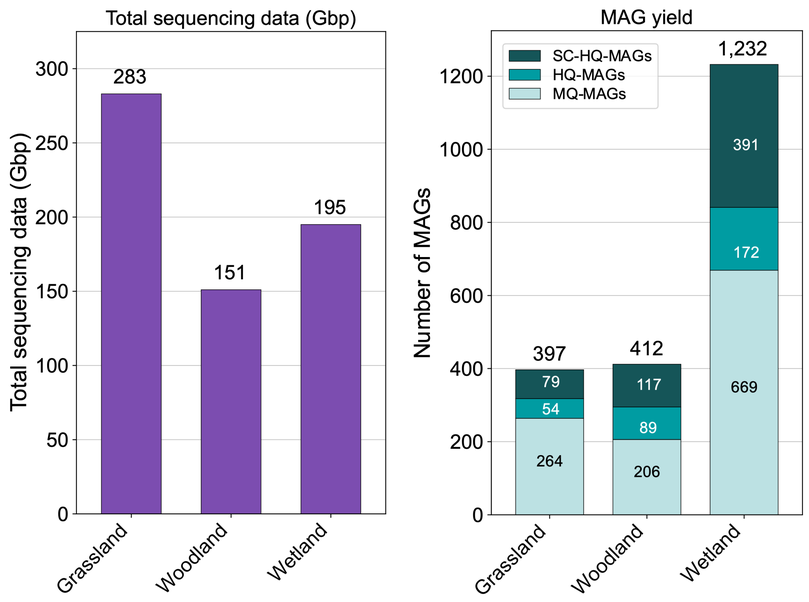

Figure 2. For each sample, the (a) total data and (b) number of MAGs across quality categories are shown.

Sequencing depth directly impacts MAG yield

- Human gut displays a saturation curve, suggesting a majority of MAGs are recovered.
- All soils display a linear relationship between total data (Gb) and MAGs recovered (Fig. 3).
- This suggests a large proportion of the total genomes in the soil samples are not captured.

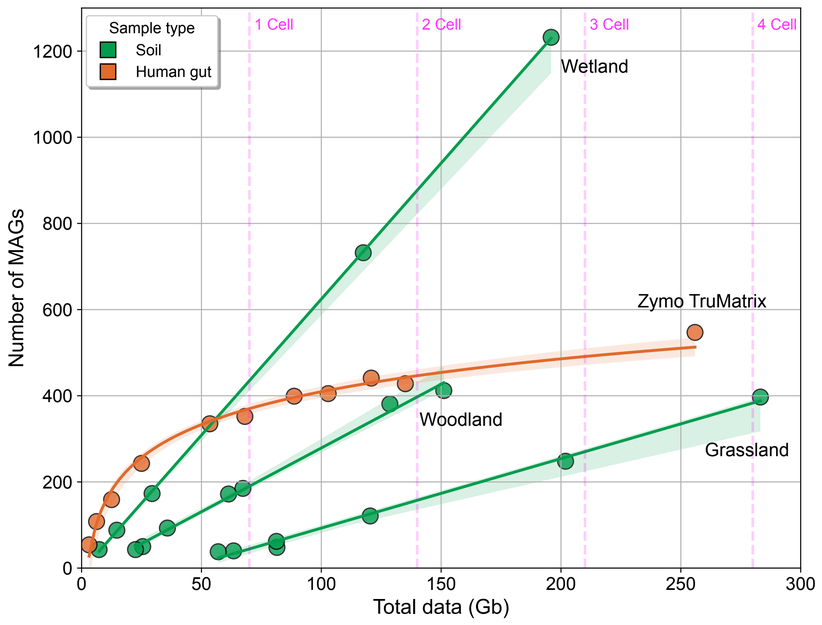

Figure 3. Relationship between total data and MAG yield for two HiFi deep-sequencing datasets. The human gut sample is from Portik et al. (2024)[3]. Revio yields vary and are estimated for 7kb reads (typical for metagenomics).

New genomes obtained for hundreds of uncultured species

None of the MAGs could be assigned to the species level by GTDB-Tk, indicating high levels of novel diversity present in all soil types.

Soil type	Bacterial MAGs	Archaeal MAGs	Total MAGs
Wetland	1,097	135	1,232
Grassland	388	9	397
Woodland	411	1	412

A large number of archaeal MAGs (n=135) were found in the wetland soil dataset. Phylogenetic analyses indicate they belong to 8 phyla (Fig. 4). Of the 135 MAGs, 50 are HQ-MAGs and represent the first genomes available for these understudied taxa.

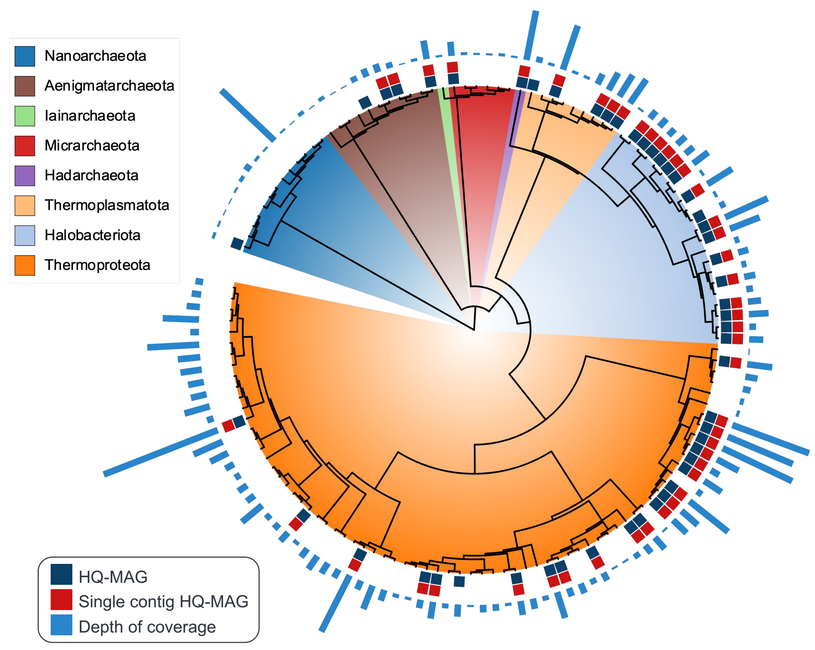

Figure 4. Phylogeny of 135 archaeal MAGs from wetland soil. The most abundant phyla include Thermoproteota, Halobacteriota, and Nanoarchaeota.The 50 HQ-MAGs here represent the first HQ genomes available for each taxonomic group.

Conclusions

- PacBio HiFi sequencing is effective for obtaining large numbers of high-quality MAGs from different soil types.
- Over 2,000 MAGs were recovered from three soil samples, including 902 HQ-MAGs.
- Deep-sequencing is essential for assembling highly complex microbiomes; this is now cost-effective using the PacBio Revio system.

All PacBio metagenomics workflows are open-source and publicly available on Github:

PacificBiosciences / pb-metagenomics-tools

References

1. Feng et al. 2022. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, 19: 671–674.
2. Benoit et al. 2024. High-quality metagenome assembly from long accurate reads with metaMDBG. *Nature Biotechnology*, doi: 10.1038/s41587-023-01983-6
3. Portik et al. 2024. Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods. *bioRxiv*, doi: 10.1101/2024.05.10.593587