EEB-1132

6/14/2024

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

Abstract # 5766

Daniel M. Portik[1], Luis E. Valentin-Alvarado[2,3], Jeremy E. Wilkinson[1], Jillian F. Banfield[2]
1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025; **2**. Innovative Genomics Institute, University of California, Berkeley, California 94720 USA; **3**. Department of Plant and Microbial Ecology, University of California, Berkeley, California USA

Introduction

There are many challenges associated with metagenome assembly:

- 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 microbiomes 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 with long reads. Here, we demonstrate that the amount of HiFi data from the high-throughput PacBio Revio system is sufficient to assemble high-quality MAGs in complex microbiomes, including soil.

Methods

PacBio HiFi sequencing

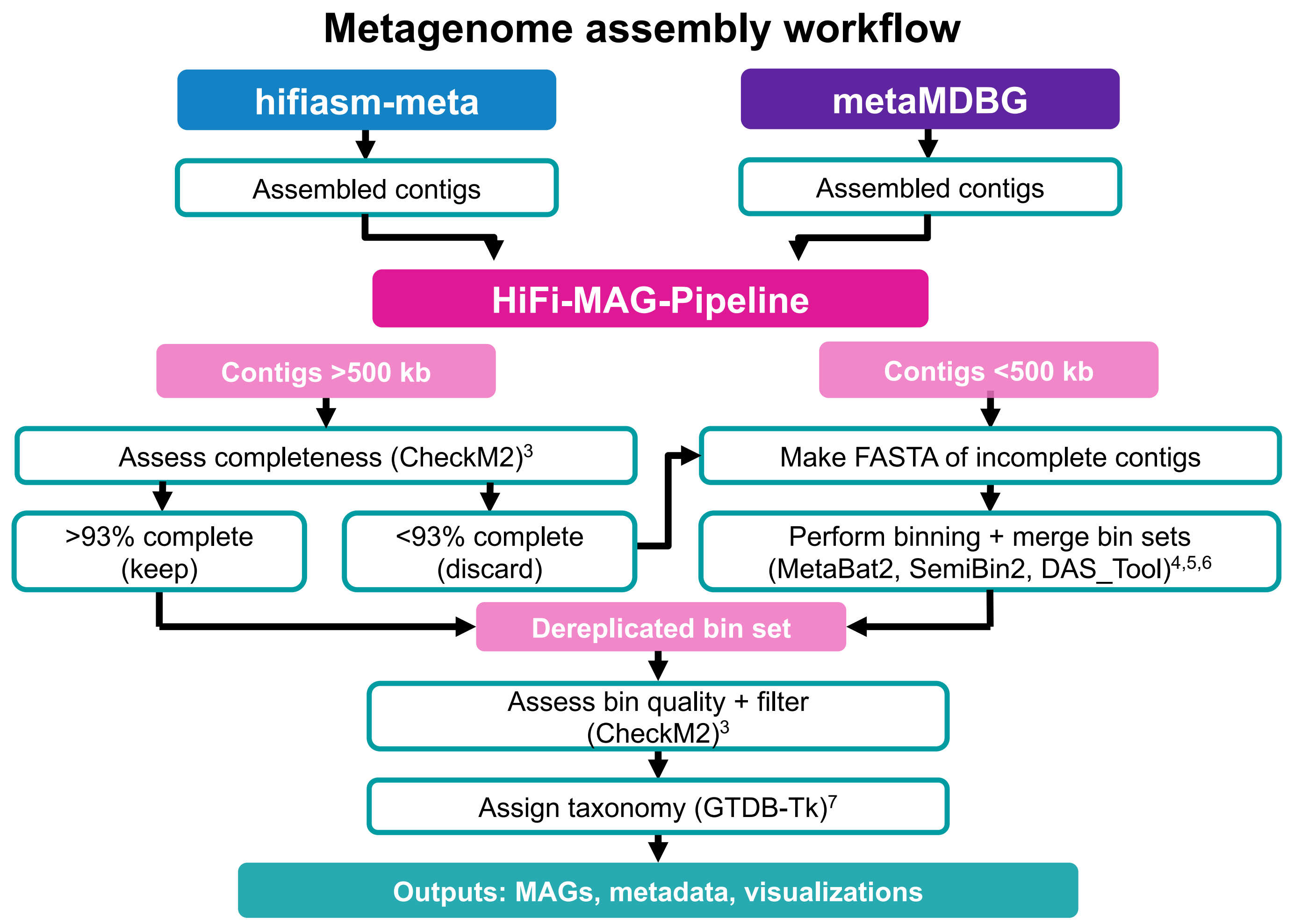
A soil core was obtained from a northern California wetland and sampled along six depths. The six samples were prepped and sequenced on a PacBio Revio system using three 25M SMRT Cells.

Sequencing resulted in a total of 20.6 million HiFi reads, 196 Gb total data, and a median read QV of 41 (representing >99.99% accuracy).

Sequencing	HiFi reads (million)	Total data (gigabases)	Average read length (kb)	Median QV
SMRT Cell 1	8.21 M	79.83 Gb	9.7 kb	Q40
SMRT Cell 2	6.65 M	57.15 Gb	8.6 kb	Q43
SMRT Cell 3	5.76 M	58.90 Gb	10.2 kb	Q41

Metagenome assembly and postprocessing

The combined sequencing dataset of 20.6M HiFi reads was assembled using **hifiasm-meta** and **metaMDBG**, and each contig set was processed using the PacBio **HiFi-MAG-Pipeline**. The complete analysis workflow is shown visually in Figure 1.



Metagenome assembly workflow

hifiasm-meta → Assembled contigs
metaMDBG → Assembled contigs

HiFi-MAG-Pipeline

Contigs >500 kb → Assess completeness (CheckM2)[3] → >93% complete (keep) / <93% complete (discard)

Contigs <500 kb → Make FASTA of incomplete contigs → Perform binning + merge bin sets (MetaBat2, SemiBin2, DAS_Tool)[4,5,6]

Dereplicated bin set → Assess bin quality + filter (CheckM2)[3] → Assign taxonomy (GTDB-Tk)[7] → Outputs: MAGs, metadata, visualizations

Figure 1. Visual overview of methods used for assembly and post-processing.

MAGs were categorized based on quality scores from CheckM2:

- Medium-quality (MQ): >50% completeness and <10% contamination
- High-quality (HQ): >90% completeness and <5% contamination
- Single-contig high-quality (SC-HQ): same criteria as high-quality, but the MAG also consists of one contig

Results

HiFi metagenomics routinely produces fully resolved MAGs

Both **hifiasm-meta** and **metaMDBG** produced over 350 single-contig, HQ-MAGs directly from the assembly step, which often appeared as circular genomes in the assembly graphs (Fig. 2).

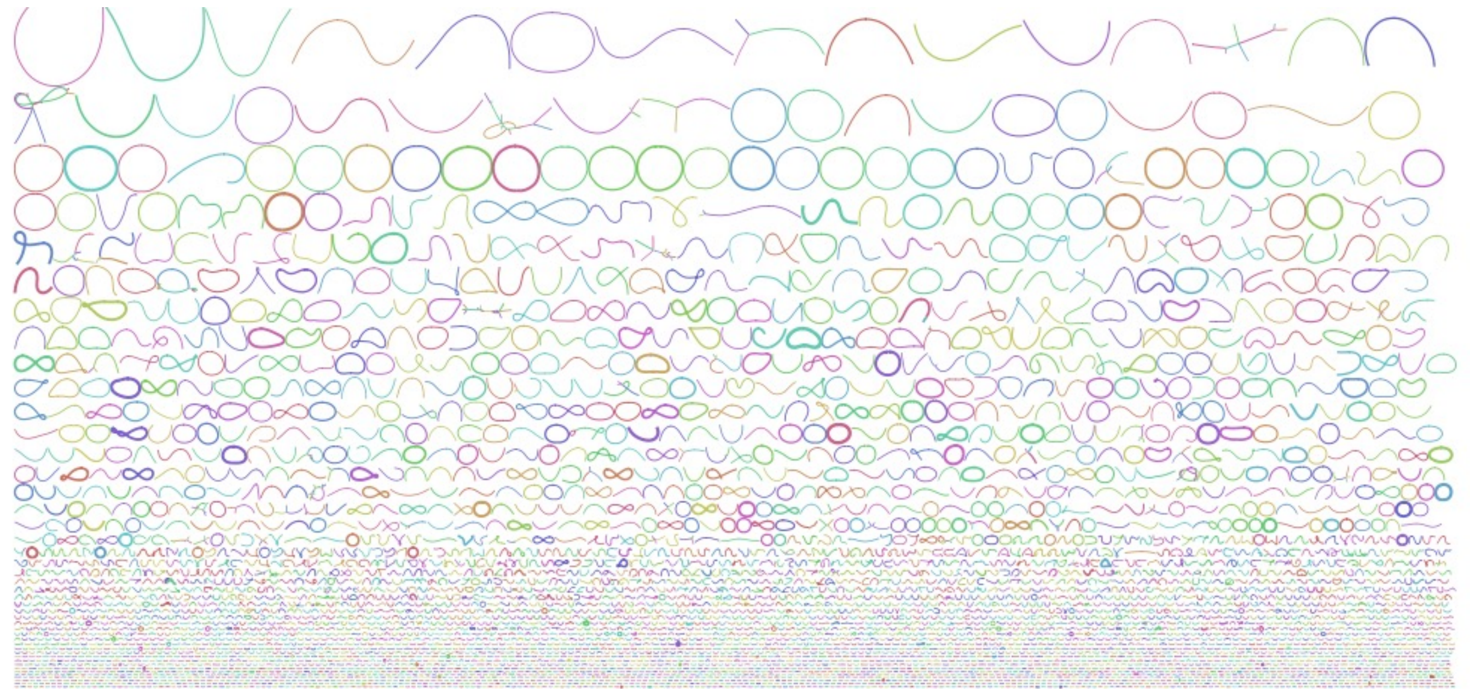


Figure 2. A partial view of the hifiasm-meta assembly graph for the combined soil dataset. The graph reveals many large circular contigs (0.5–13 Mb) produced directly from assembly.

Over 500 high-quality MAGs assembled from a soil sample

- **hifiasm-meta** produced over 1,200 total MAGs, with 550 HQ-MAGs (391 are single-contig; Fig. 3)
- **metaMDBG** resulted in 1,000 total MAGs, with over 500 HQ-MAGs (355 are single-contig; Fig. 3)

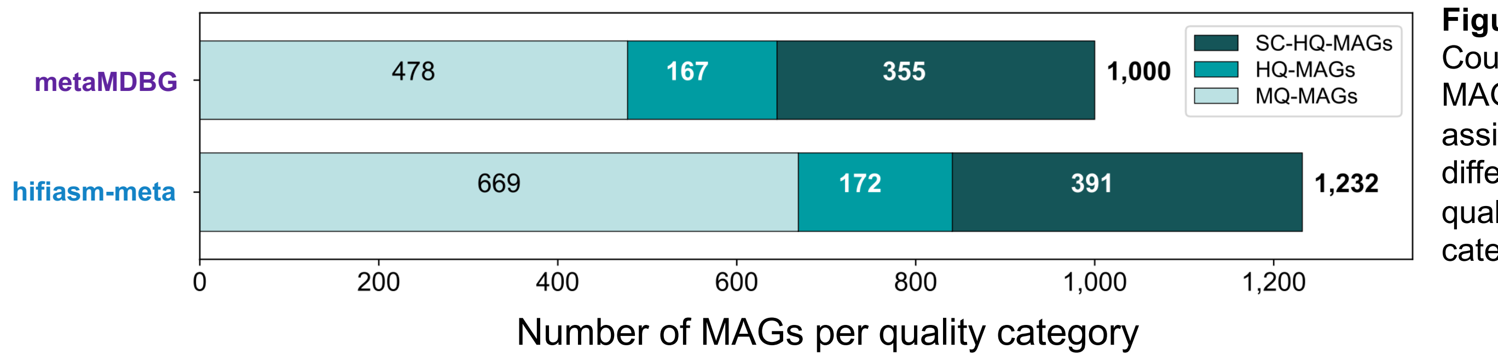


Figure 3. Counts of MAGs assigned to different quality categories.

Assembling genomes for hundreds of uncultured species

None of the MAGs could be assigned to the species level by GTDB-Tk.

Recovered **1,097 bacterial** MAGs from 51 phyla (513 HQ-MAGs; Fig. 4).

Recovered **135 archaeal** MAGs from 8 phyla (50 HQ-MAGs; Fig. 5).

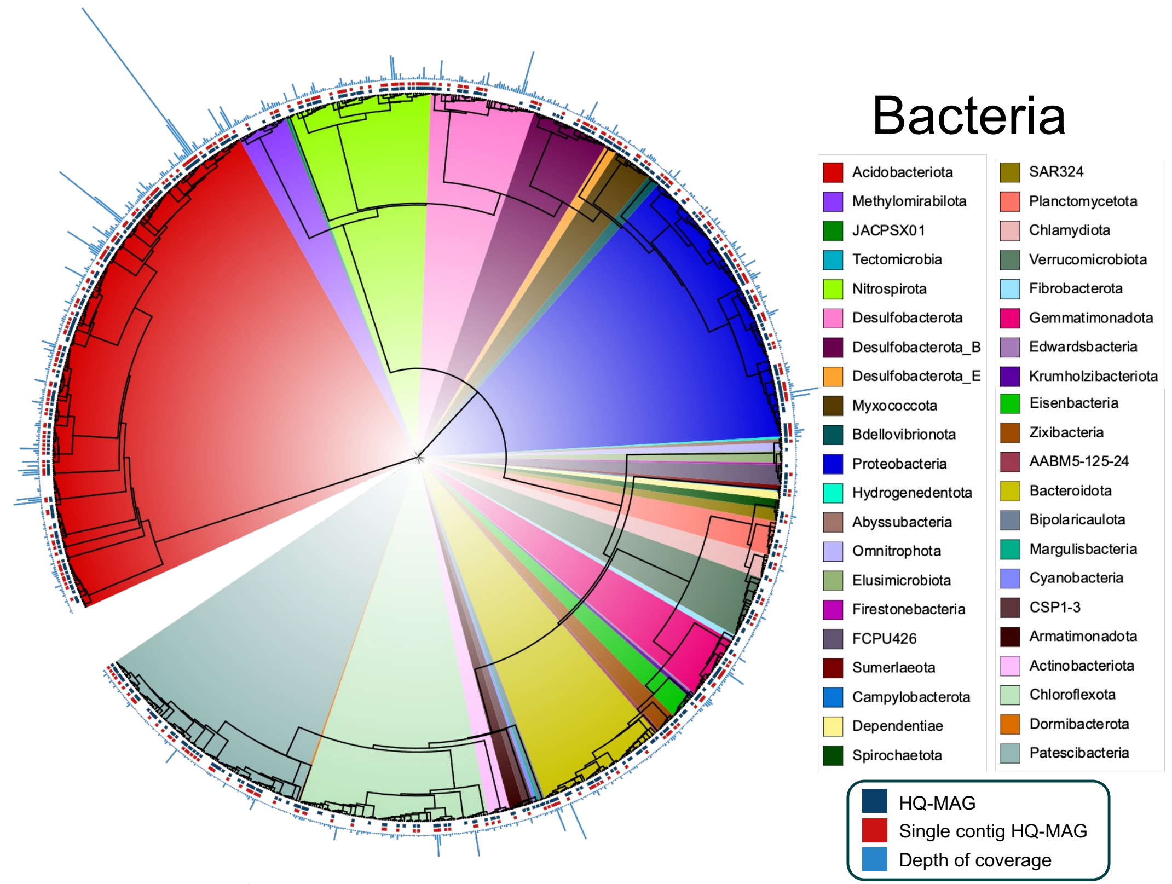


Figure 4. Phylogeny of 1,097 bacterial MAGs that were assembled.

The most abundant phyla include Acidobacteriota, Proteobacteria, Patescibacteria, and Chloroflexota.

Outside edges of the phylogeny show whether a MAG is classified as HQ (navy), SC-HQ (red), and the relative depth of coverage (light blue).

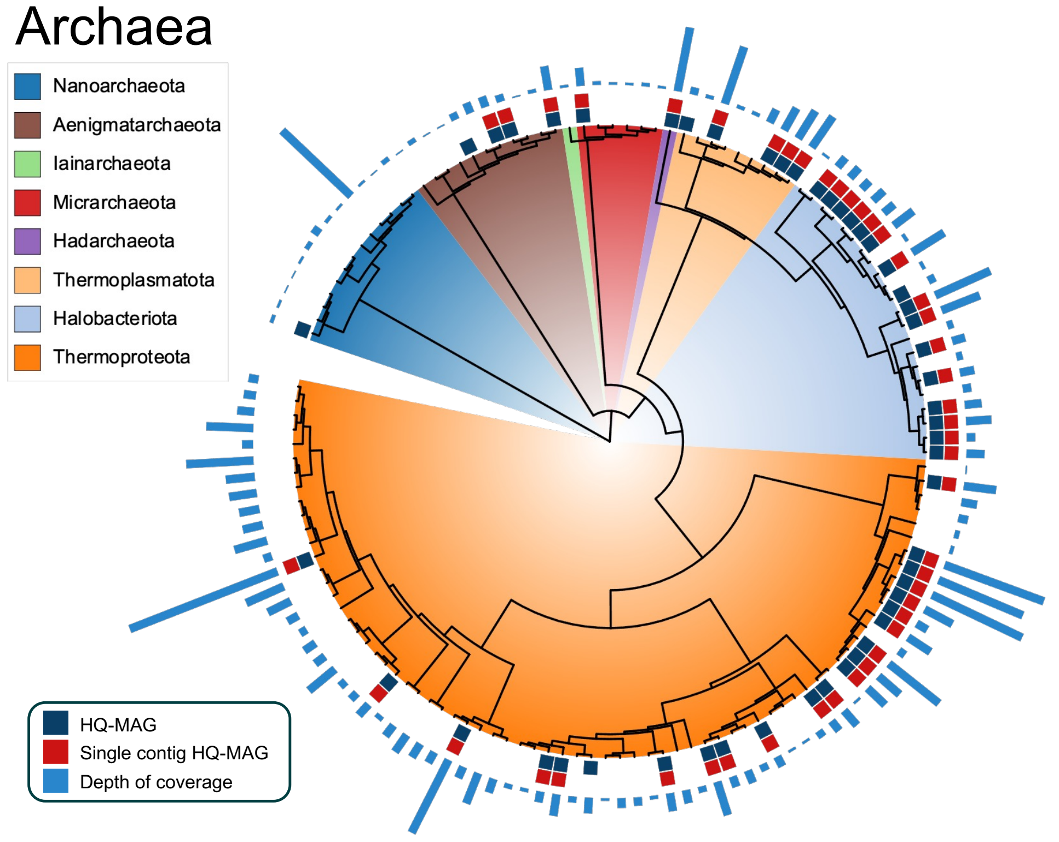


Figure 5. Phylogeny of 135 archaeal MAGs that were assembled in this study. The most abundant phyla include Thermoproteota, Halobacteriota, and Nanoarchaeota.

The 50 HQ-MAGs here are in many cases the first high-quality genomes available for a given taxonomic group.

Outside edges of the phylogeny show whether a MAG is classified as HQ (navy), SC-HQ (red), and the relative depth of coverage (light blue).

Long-read sequencing recovers genomes with diverse properties

- Size range across HQ-MAGs was 0.5–12Mb (median 4Mb)
- HQ-MAGs display a range of 35–73% GC content (Fig. 6)
- HQ-MAGs displayed 4X–400X depth of coverage (median 17X)

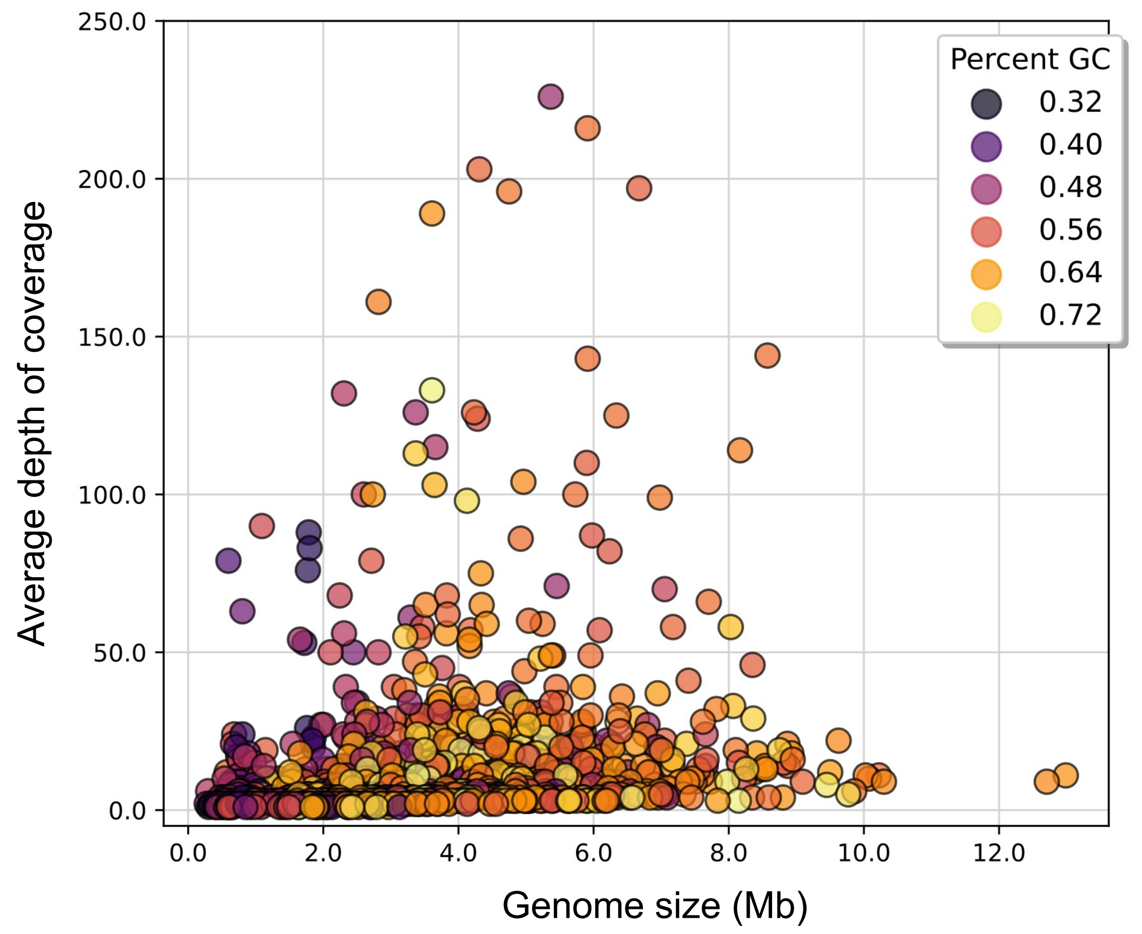


Figure 6. Characteristics of the 1,232 MAGs that were recovered by hifiasm-meta, including the genome size (x-axis) and average depth of coverage per genome (y-axis). Each point represents an individual MAG, and they are color-coded based on their estimated percent GC content. Three high coverage genomes were excluded form the plot (400-600X).

Conclusions

- PacBio HiFi sequencing offers major advantages for metagenome assembly, particularly for difficult environmental samples.
- Single-contig HQ-MAGs are routinely assembled with HiFi reads.
- HiFi sequencing is effective for obtaining large numbers of high-quality MAGs from uncultured species in complex microbiomes.

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*, https://doi.org/10.1038/s41587-023-01983-6
3. Chklovski et al. 2023. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *bioRxiv*, https://doi.org/10.1101/2022.07.11.499243
4. Kang et al. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7: e7359.
5. Pan et al. 2023. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics*, 39: i21–i29.
6. Sieber et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3: 836–843.
7. Chaumeil et al. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 35: 1925–1927.

Research use only. Not for use in diagnostic procedures. Information in this document is subject to change without notice. PacBio assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of PacBio products and/or third-party products. Refer to the applicable PacBio terms and conditions of sale and to the applicable license terms at pacb.com/license. Pacific Biosciences, the PacBio logo, PacBio, Circulomics, Omniome, SMRT, SMRTbell, Iso-Seq, Sequel, Nanobind, SBB, Revio, and Onso are trademarks of PacBio.