New long-read metagenome assembly methods increase the number of high-quality MAGs from host-associated microbiomes

Daniel M. Portik, Jeremy E. Wilkinson
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

There are many challenges associated with metagenome assembly, which include:

- 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 make it possible to reconstruct full metagenome-assembled genomes (MAGs) for many high abundance species (fig. 1).

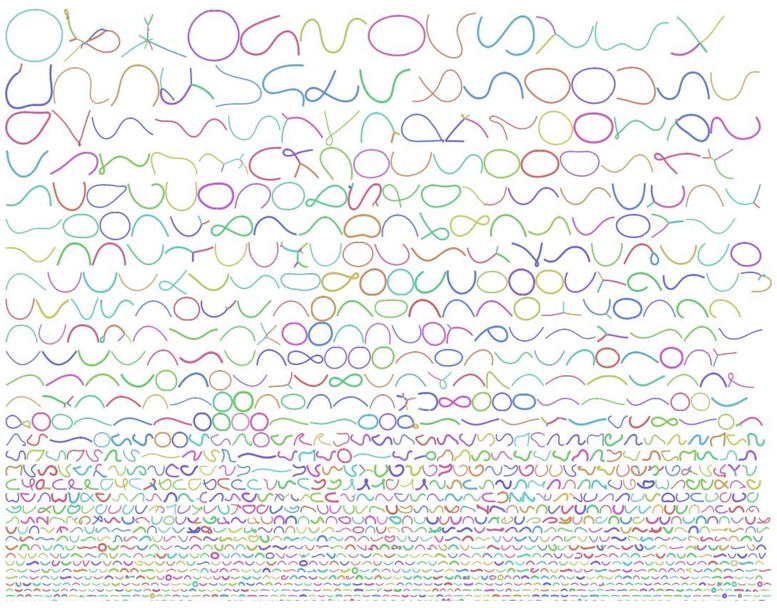

Figure 1. A partial hifiasm-meta assembly graph for a pooled human gut microbiome dataset. The graph reveals many large circular contigs (1–6 Mb) produced directly from assembly. However, large linear contigs are also produced in the assembly. These represent fragmented genomes and postprocessing is required to recover these additional high-quality MAGs.

However, discontiguous assemblies will occur for lower abundance taxa. Post-assembly tools incorporating binning methods are required to identify and extract additional MAGs. The HiFi-MAG-Pipeline (v2) is a comprehensive workflow for processing long-read assemblies, and includes major steps such as binning, quality filtering, and taxonomic identification.

Here, we demonstrate the performance of these methods using a variety of HiFi metagenomic datasets.

Methods

We selected 11 publicly available HiFi datasets to analyze (table 1). These include environmental samples, artificial environments, and plant- and animal-associated microbiomes. We used hifiasm-meta and metaMDBG for metagenome assembly, and processed each assembly using the HiFi-MAG-Pipeline (fig. 2).

Table 1. Set of publicly available HiFi metagenomic datasets selected.

Type	Sample	Accession	HiFi reads (million)	Total data (Gb)
Artificial	Digester / Bioreactor	ERR7015089	0.99	15.3
Artificial	Digester / Bioreactor	SRR24881069	3.15	26.9
Environmental	Lichen thallus	SRR24475746	3.46	11.6
Environmental	Lichen thallus	SRR24475747	3.58	24.7
Environmental	Seawater	ERR9769281	2.92	22.4
Environmental	Seawater	ERR9769303	2.55	20.6
Environmental	Hot spring sediment	DRR290133	2.69	27.9
Gut microbiome	Sheep gut	SRR14289618	18.46	206.5
Gut microbiome	Chicken gut	SRR19732729	5.87	111.9
Gut microbiome	Human gut	SRR15275211	1.90	18.8
Gut microbiome	Human gut	SRR15275213	1.79	18.5

Metagenome assembly workflow

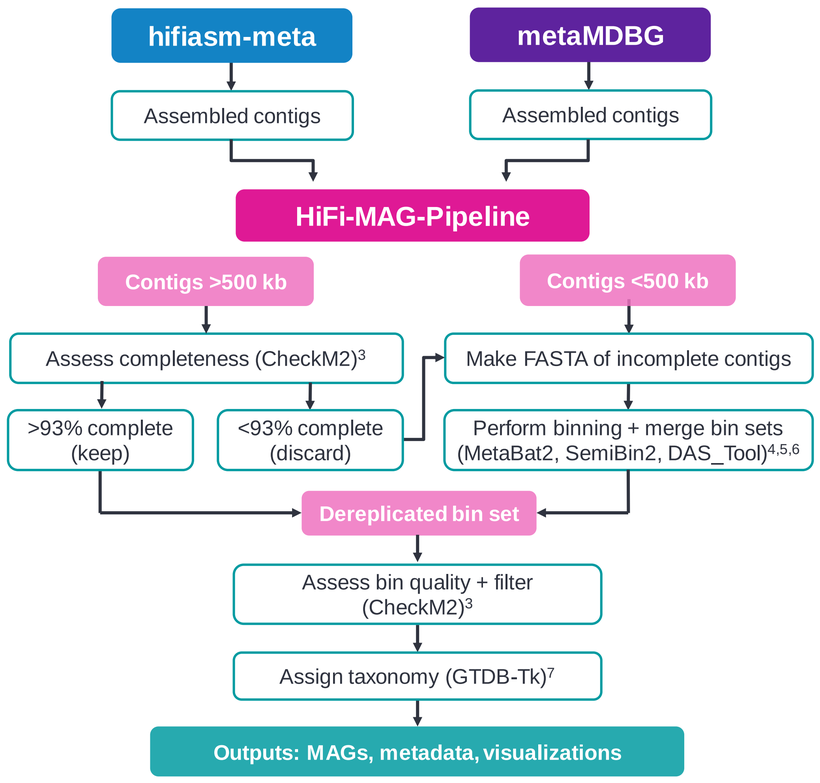

Figure 2. Visual overview of methods used for assembly and post-processing.

Results

HiFi assemblies produce many high-quality MAGs

- Both assembly methods produce hundreds of MAGs
- Recovered 38–1,036 MAGs per sample (fig. 3).
- Over ~2,700 MAGs recovered across all eleven samples, including ~1,400 high-quality MAGs (51%).

metaMDBG tends to outperform hifiasm-meta

- On average, metaMDBG results in a 90% increase in the number of MAGs recovered (with as much as a 218% increase in total MAGs).
- Both methods perform similarly well for lower diversity samples (e.g., lichen thallus).

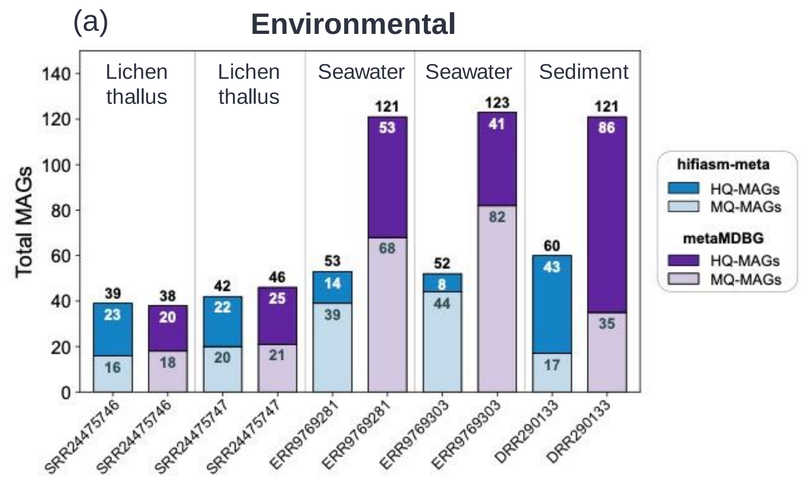

Figure 3. MAG yields across different sample types, including (a) environmental samples, (b) artificial environments, and (c) gut microbiomes. Results for hifiasm-meta shown in blue; metaMDBG shown in purple. Total MAG numbers are shown on top, with barplot colors showing medium (MQ; >50% completeness, <10% contamination) and high-quality categories (HQ; >90% completeness, <5% contamination). Sample information is available in table 1. Note we could not assemble the chicken gut microbiome dataset with hifiasm-meta due to memory limitations (despite having access to 1Tb total memory).

metaMDBG can assemble hundreds of single-contig, high-quality MAGs per sample

- The metaMDBG chicken and sheep gut assemblies resulted in 269 and 579 HQ-MAGs, respectively.
- We found 70% of the sheep gut HQ-MAGs are composed of a single contig (n = 404; fig. 4).
- Approximately 73% of the chicken gut HQ-MAGs are composed of a single contig (n = 199).

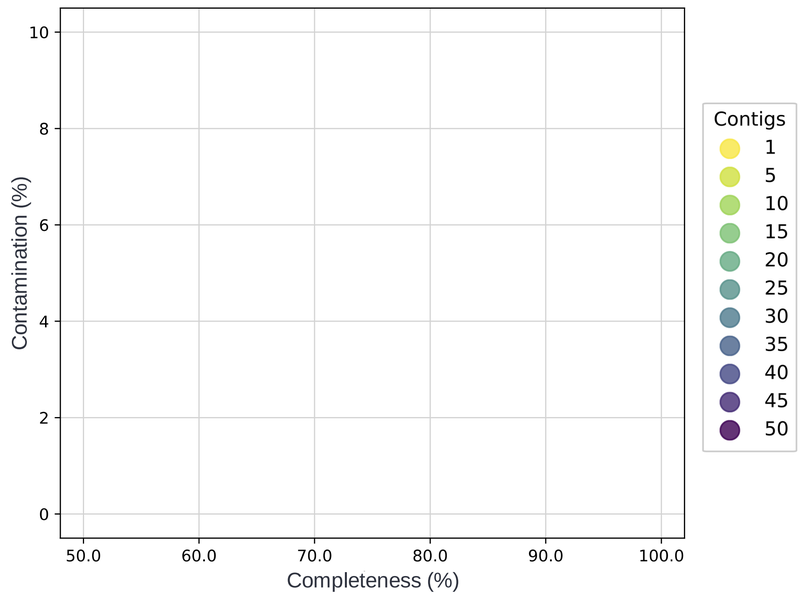

Figure 4. Completeness versus contamination scores for 579 high-quality MAGs for the sheep gut assembly with metaMDBG. Each dot represents a MAG, and colors indicate the number of contigs contained in the MAG. We found 404 MAGs displayed >90% completeness, <5% contamination, and are composed of a single contig.

Conclusions

- PacBio HiFi sequencing offers major advantages for metagenome assembly.
- Complete, single-contig MAGs can be routinely assembled from HiFi reads.
- The development of new tools, such as metaMDBG, continue improving HiFi metagenomic analyses.
- HiFi sequencing is an effective strategy for obtaining large numbers of high-quality MAGs, particularly for uncultured and uncharacterized species.
- HiFi metagenomics can be used to characterize host-associated microbiomes for plant and animal species.

All PacBio metagenomics pipelines 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.