Maximizing MAGs from long-read metagenomic assemblies: a new post-assembly pipeline with circular-aware binning

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

Introduction

There are many challenges involved with metagenome assembly, including the presence of multiple species, uneven species abundances, and conserved genomic regions that are shared across species. Highly accurate long reads offer clear advantages over short reads and can overcome many of the obstacles associated with metagenome assembly (Fig. 1). PacBio HiFi sequencing of metagenomic samples with the Sequel IIe system regularly produces reads 8–15 kb in size with a median QV ranging from 30–45 (99.9–99.99% accuracy).

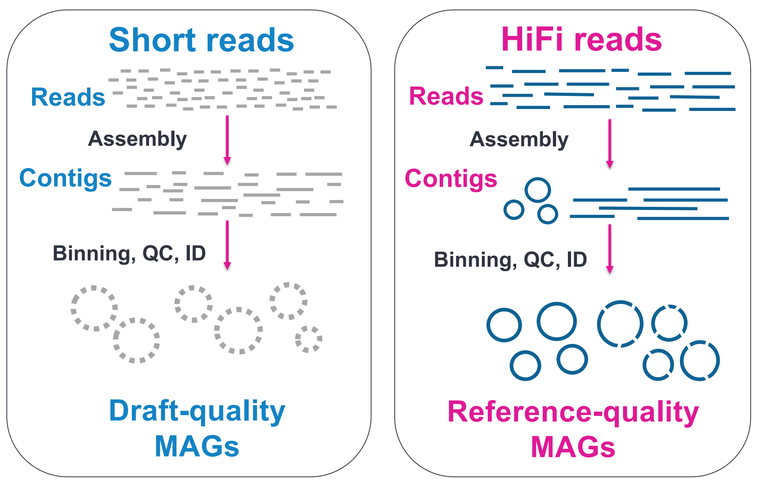

Figure 1. Metagenome assembly. Short-read assemblies produce smaller contigs, rely heavily on binning methods, and produce MAGs composed of dozens to hundreds of contigs. HiFi reads are similar in size (or larger) to the short-read based contigs, and overcome challenges associated with repeats, conserved regions, and ribosomal genes (including 16S). HiFi MAGs routinely include single-contig complete genomes and MAGs composed of a handful of contigs, which may be considered reference-quality.

New metagenome assembly algorithms have been developed for HiFi reads, including **hifiasm-meta**[1] and **metaFlye**.[2] With these methods, it is now possible to reconstruct full metagenome assembled genomes (MAGs) for many high-abundance species.[1,3,4,5] These MAGs can be composed of a single circular contig, representing a complete genome (Fig. 2). However, discontiguous assemblies still occur for lower abundance taxa, and post-assembly tools are required to process MAGs in this category. Here, we present the newest version of a workflow (v1.6) for processing long-read metagenome assemblies, **HiFi-MAG-Pipeline**.

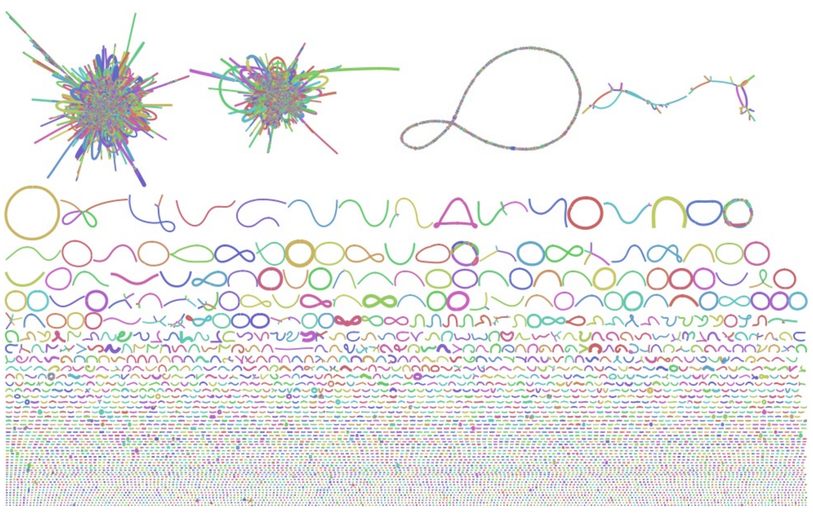

Figure 2. Metagenome assembly graph. A **hifiasm-meta** assembly graph for a human gut microbiome dataset. The graph reveals many large (>1 Mb) circular contigs produced directly from assembly. These represent complete MAGs and do not require binning methods to be discovered. However, many large linear contigs are also produced in the assembly. These often represent fragmented genomes, and postprocessing is required to recover these additional high-quality MAGs.

PacBio metagenomics pipelines

- **HiFi-MAG-Pipeline** and other workflows are freely available on github:

PacificBiosciences / **pb-metagenomics-tools**

Implemented in snakemake, a Python-based workflow management system

- Scalable to HPC, cloud compatible, and can also be run locally
- Automates all workflow steps and includes checkpoints
- Conda installs environments and dependencies for all steps

Workflow overview

- Assembly is performed prior to running the workflow.

HiFi-MAG-Pipeline

Inputs: HiFi reads & assembly contigs

↓

Map reads to contigs (minimap2)[6]

↓

Binning: circular-aware + standard methods (MetaBAT2, MaxBin2, CONCOCT, DAS_Tool)[7,8,9,10]

↓

Assess bin quality and filter (CheckM)[11]

↓

Assign taxonomy (GTDB-Tk)[12]

↓

Outputs: MAGs, metadata, visualizations

Circular-aware binning

- Standard binning assumes genomes are fragmented.
- This can cause unexpected behavior for HiFi assembly: complete contigs can be mis-binned with linear contigs.
- Mis-binning inflates contamination scores and causes removal of the bin during filtering (Fig. 3).
- The circular-aware strategy uses standard binning in combination with manual binning of circular contigs.
- The different bin sets are combined and de-replicated to produce the final bins, rescuing circular contigs.

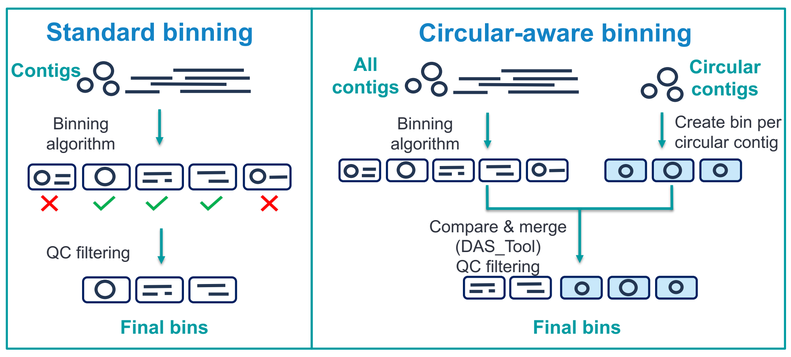

Figure 3. Comparison of binning strategies. Standard binning can cause the mis-binning of complete circular contigs, resulting in their removal. Circular-aware binning is a simple strategy that results in the retention of complete circular contigs.

Binning methods comparison

We compared a standard binning pipeline (MetaBAT2) to **HiFi-MAG-Pipeline** v1.6. We assembled 12 publicly available HiFi metagenomic datasets with hifiasm-meta (Table 1) and then performed binning analyses.

Organism	Dataset	HiFi reads	Avg read length	Total data	Median QV
Human	Omnivore gut 1	1.79 M	10.3 kb	15.2 Gb	Q40
	Omnivore gut 2	1.68 M	9.2 kb	15.5 Gb	Q40
	Vegan gut 1	1.90 M	9.8 kb	18.8 Gb	Q39
	Vegan gut 2	1.76 M	8.6 kb	18.5 Gb	Q39
	French gut	1.64 M	7.9 kb	13.0 Gb	Q35
	Korean gut	2.01 M	14.6 kb	29.6 Gb	Q34
	Pooled gut	11.89 M	7.4 kb	88.3 Gb	Q41
Animal	Sheep gut	18.45 M	11.2 kb	206.5 Gb	Q35
Environmental	Activated sludge	0.99 M	15.4 kb	15.3 Gb	Q35
	Hot spring sediment	2.69 M	10.3 kb	27.9 Gb	Q31
	Photobioreactor	1.41 M	3.2 kb	4.6 Gb	Q40

Table 1. HiFi datasets. Summary of HiFi metagenomic datasets used for analyses. A larger list of 50+ available datasets and their associated publications can be found on the PacBio metagenomics GitHub repo.

Results

HiFi assemblies produce many high-quality MAGs

- Recovered 50–285 MAGs per sample
- Found 13–151 MAGs (32–53%) are single-contig (Fig. 4)

Circular-aware binning yields more total MAGs

- Found 15–64% increase in total MAGs (Fig. 4)
- Gain of 9–80 total MAGs per sample

Circular-aware binning rescues complete circular MAGs

- Found 8–100% increase in single-contig, complete circular MAGs (Sheep gut, Fig. 4)
- Incomplete circular contigs are successfully binned

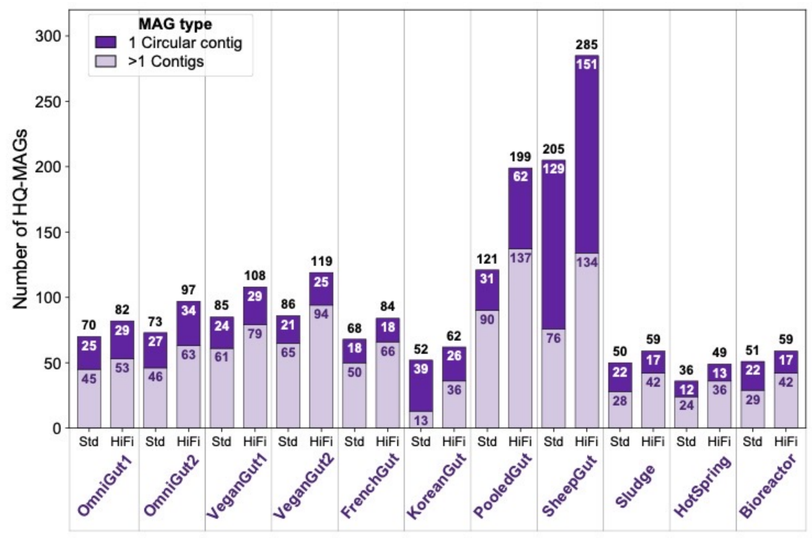

Figure 4. Binning results. MAG yields from standard binning with MetaBAT2 (Std) vs HiFi-MAG-Pipeline (HiFi). Dark purple represents single-contig circular MAGs and light purple represents MAGs containing >1 contigs. Numbers in the stacked bars represent each category, and numbers above represent total MAGs.

Visualizations and metadata

HiFi-MAG-Pipeline produces several informative figures displaying quality characteristics for MAGs recovered (Fig. 5), and provides metadata from CheckM and GTDB-Tk.

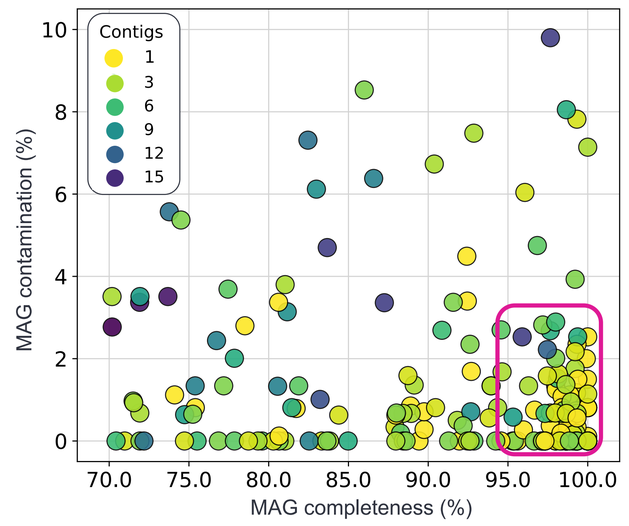

Figure 5. MAG characteristics. Completeness versus contamination scores for 199 high-quality MAGs found by HiFi-MAG-Pipeline for the human pooled gut assembly. Each dot represents a MAG, and colors indicate the number of contigs contained in the MAG. We found 102 MAGs (51%) were exceptionally high quality and displayed >95% completeness (pink outline), with 54 being single contig.

Conclusions

- **PacBio HiFi sequencing offers major advantages for metagenome assembly and MAG recovery.**
- Recent studies demonstrate single-contig, complete MAGs can be assembled from HiFi reads.[1,3,4,5]
- Most binning methods assume genomes are fragmented. HiFi metagenome assemblies can violate this assumption, leading to unexpected behavior.
- **HiFi-MAG-Pipeline** automates key steps to obtain HQ MAGs from long-read metagenome assemblies.
- Up to 64% increase in total MAGs and 100% increase in single-contig circular MAGs using the custom binning strategy in **HiFi-MAG-Pipeline**.
- The mis-binning of single-contig, complete circular MAGs is a pervasive problem for long-read assemblies.
- Higher sequencing throughput with Revio is expected to improve MAG yields.

Literature

1. Feng, X., et al. 2022. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, 19: 671–674.
2. Kolmogorov, M., et al. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17: 1103–1110.
3. Bickhart, D.M., et al. 2022. Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology,* 40: 711–719.
4. Gehrig, J.L., et al. 2022. Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microbial Genomics*, 8: 000794.
5. Saak, C.C., et al. 2022. Longitudinal, multi-platform metagenomics yields a high-quality genomic catalog and guides an in vitro model for cheese communities. *bioRxiv*, https://doi.org/10.1101/2022.07.01.497845
6. Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094–3100.
7. Kang, D.D., et al. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7: e7359.
8. Wu, Y.-W., et al. 2015. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32: 605–607.
9. Alneberg, J., et al. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11: 1144–1146.
10. Sieber, C.M.K., et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3: 836–843.
11. Parks, D.H., et al. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25: 1043–1055.
12. Chaumeil, P.-A., et al. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 35: 1925–1927.