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). With the development of new metagenome assembly algorithms specific to HiFi reads (hifiasm-meta[1], metaFlye[2]), it is now possible to reconstruct full metagenome assembled genomes (MAGs) for many high abundance species[1,3,4,5]. These MAGs are often composed of a single circular contig, representing high-quality complete bacterial genomes. However, discontiguous assemblies still occur for lower abundance taxa, and post-assembly tools are required to identify MAGs in this category. Here, we present the HiFi-MAG-Pipeline, a comprehensive workflow for processing long-read metagenome assemblies.

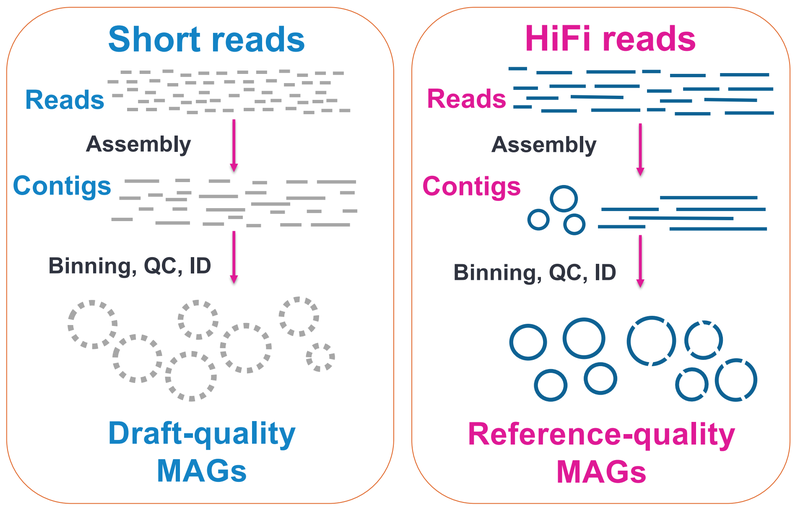


Figure 1. Metagenome assembly. Differences in metagenome assembly and MAG quality are technology specific. Short read assemblies rarely produce single-contig, complete genomes and rely heavily on binning methods to reconstruct putative genomes. Resulting MAGs are composed of dozens to hundreds of contigs, representing draft-quality genomes. HiFi reads are similar in size (or larger) to many contigs assembled using short reads, and overcome challenges associated with repeat regions and intraspecific conserved regions. HiFi MAGs routinely include single-contig complete genomes and MAGs composed of a handful of contigs, which may be considered reference-quality genomes.

hifiasm-meta assembly graph

Visualizing the assembly graph reveals many large (>1 Mb) circular contigs were produced directly from hifiasm-meta[1] (Fig. 2). These circular contigs represent truly complete MAGs and do not require binning methods to be discovered. However, many large linear contigs are also produced, representing fragmented genomes, and these require postprocessing to recover additional high-quality MAGs.

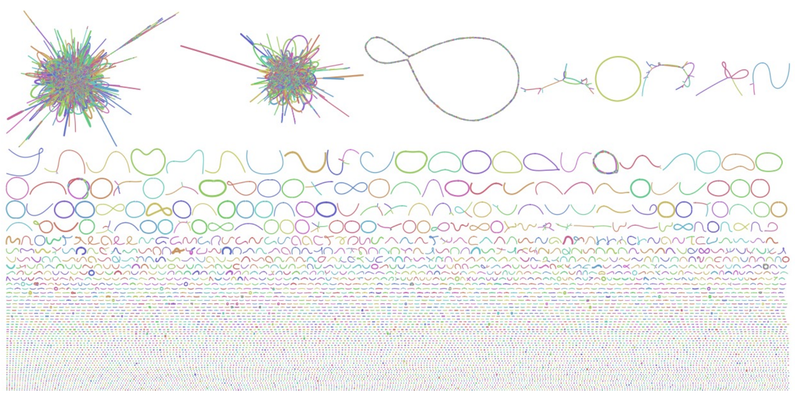


Figure 2. The hifiasm-meta[1] assembly graph. Assembly graph of the human pooled gut dataset depicting many circular contigs from the metagenome assembly of the HiFi data.

HiFi-MAG-Pipeline

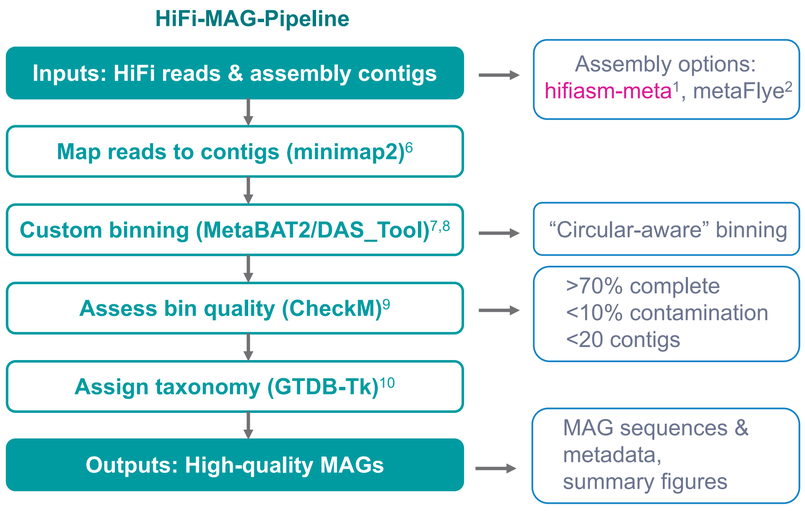
- PacBio metagenomics pipelines and documentation are freely available on github:

  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 workflow steps and includes checkpoints
  - Conda installs environments and dependencies for all steps

Workflow overview

The inputs to the workflow include a set of assembled contigs and the HiFi reads used to generate the assembly. Contigs can be assembled using any of the recommended methods. The outputs include all high-quality (HQ) MAG sequences, metadata (quality metrics, taxonomy), and visualizations.



Circular-aware binning

- Many standard binners assume genomes are fragmented.
- This can lead to unexpected behavior in long-read assembly.
- Circular contigs can be mis-binned with linear contigs. This results in inflated contamination scores and the subsequent removal of the bin (Fig. 3).
- Circular-aware binning uses standard binning and manual binning of circular contigs. The bin sets are combined and de-replicated to produce the final bins.

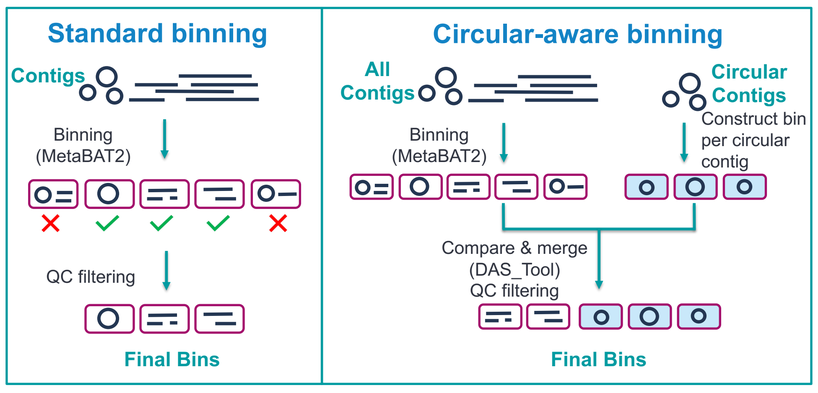


Figure 3. Comparison of binning strategies. MetaBAT2 can result in the mis-binning of complete circular contigs, resulting in their removal. Circular-aware binning is a simple and effective strategy that results in retention of complete circular contigs in the final bin set.

Binning methods comparison

We performed a direct comparison of MetaBAT2 binning to circular-aware binning. We used several publicly available HiFi metagenomic datasets (Table 1) to perform analyses. We implemented each binning strategy in the HiFi-MAG-Pipeline, with otherwise identical steps, and compared the final bin sets.

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
	Pooled gut	11.89 M	7.4 kb	88.3 Gb	Q41
Sheep	Sheep gut	11.84 M	11.2 kb	206.5 Gb	Q35
Environmental	Activated sludge	0.99 M	15.4 kb	15.3 Gb	Q35

Table 1. HiFi datasets. Summary of HiFi metagenomic datasets used for analyses. A larger list of available datasets can be found on github: **PacificBiosciences/pb-metagenomics-tools**

HiFi metagenomic assemblies produce many high-quality MAGs

- Regardless of binning strategy, we assembled 50–250 HQ MAGs per sample
- Between 22–171 (35–68%) of the total MAGs are single-contig, highlighting HiFi metagenome assembly routinely produces complete bacterial genomes (Fig. 4)

Circular-aware binning consistently yields more total MAGs

- Found 4–22% increase in total high-quality MAGs (avg 13% increase; Fig. 4)
- Gain of 2–46 total MAGs per sample (avg gain = 16)

Circular-aware binning effectively rescues single-contig, complete MAGs

- Found 5–66% increase in single-contig MAGs (avg 35% increase; Fig. 4)
- Gain of 1–43 single-contig MAGs per sample (avg gain = 15)

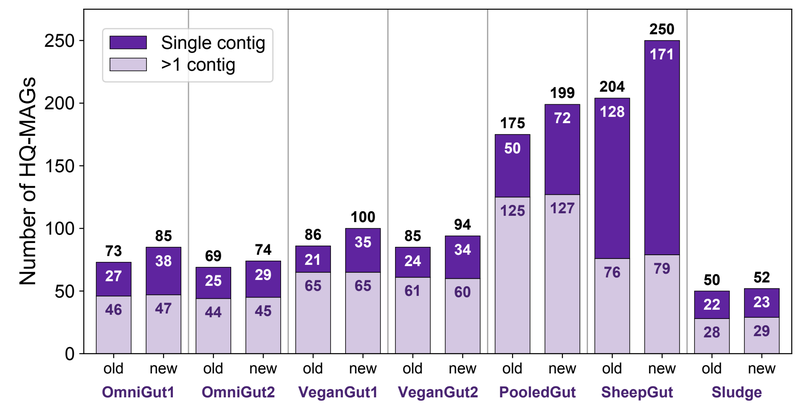
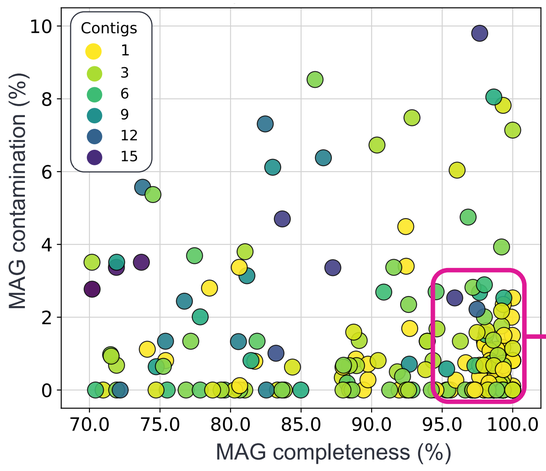


Figure 4. Binning results. MAG yield from MetaBAT2-only (old) vs. circular-aware binning (new) strategies implemented in HiFi-MAG-Pipeline. Dark purple represents single-contig MAGs and light purple represents MAGs containing >1 contigs. Numbers in the stacked bars represent number of MAGs in each category, whereas numbers above represent the total HQ MAGs per sample.

Visualizations and metadata



The HiFi-MAG-Pipeline produces informative figures summarizing the quality characteristics for all MAGs recovered (Fig. 5). Metadata are based on results from CheckM and GTDB-Tk, facilitating downstream analyses.

Exceptionally high-quality MAGs
- 102 MAGs in this zone
- 54 single contig

Figure 5. MAG characteristics. Completeness versus contamination scores for **199 HQ MAGs** found in the human pooled gut sample, as evaluated by CheckM. Each dot represents a MAG, and colors indicate the number of contigs contained in the MAG. We found 102 HQ MAGs (51%) were exceptionally high quality, displaying >95% completeness. Of these MAGS, 54 (27%) were composed of a single contig and are reference-quality.

Conclusions

- **PacBio HiFi sequencing offers clear advantages for metagenome assembly.**
- Recent studies have demonstrated it is possible to produce many single-contig, complete MAGs using HiFi reads and appropriate assembly methods[1,3,4,5].
- Many binning methods assume all genomes are fragmented. HiFi metagenome assemblies can violate this assumption leading to unexpected behavior.
- We developed **HiFi-MAG-Pipeline** to automate binning, quality control, and taxonomy steps to obtain HQ MAGs from long-read metagenome assemblies.
- Up to 22% more HQ MAGs are recovered using the circular-aware binning strategy implemented in the PacBio HiFi-MAG-Pipeline (versus MetaBAT2 alone).
- Up to 66% more single-contig MAGs are recovered using circular-aware binning. This indicates that the mis-binning of single-contig MAGs is a pervasive problem for some standard binning approaches (including MetaBAT2).
- PacBio metagenomics pipelines, documentation, and datasets are freely available on github:

  PacificBiosciences / **pb-metagenomics-tools**

References

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. Sieber, C.M.K., et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3: 836–843.
9. 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.
10. Chaumeil, P.-A., et al. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 35: 1925–1927.