MBP-666
6/18/2023

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

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

Introduction

There are many challenges to **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

Highly accurate long reads can overcome many of the obstacles associated with metagenome assembly. **PacBio HiFi sequencing** of metagenomic samples with the Sequel IIe or Revio systems 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, including hifiasm-meta[1], it is now possible to reconstruct full metagenome-assembled genomes (MAGs) for high abundance species (Fig. 1).

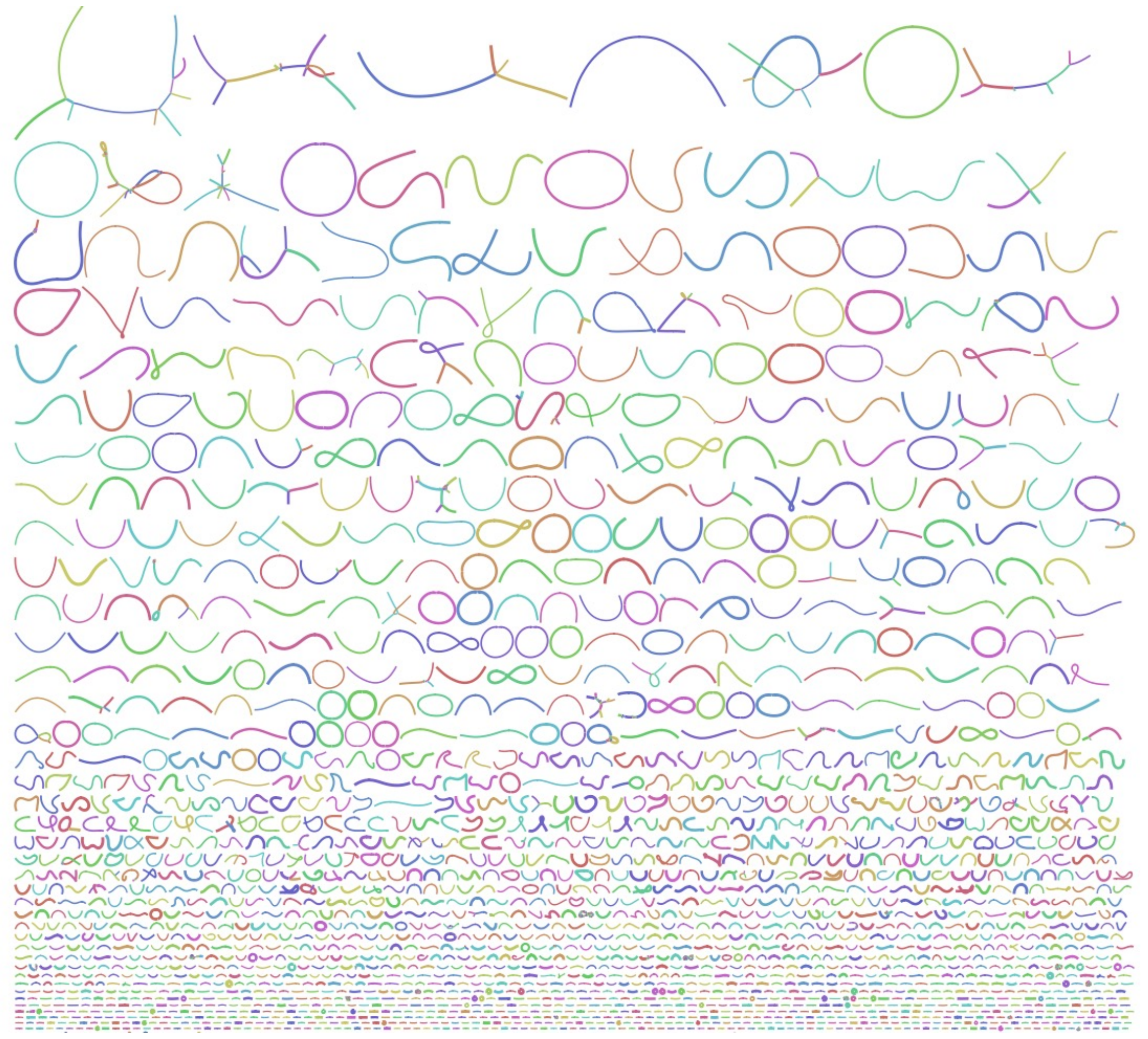


Figure 1. A partial hifiasm-meta assembly graph for a pooled human gut microbiome dataset. The graph reveals many large circular contigs (1-6 Mb) produced directly from assembly. However, many large linear contigs are also produced in the assembly. These represent fragmented genomes and postprocessing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

However, discontiguous assemblies (e.g., fragmented MAGs) will occur for lower abundance taxa. Post-assembly tools incorporating binning methods are therefore required to identify and extract additional MAGs.

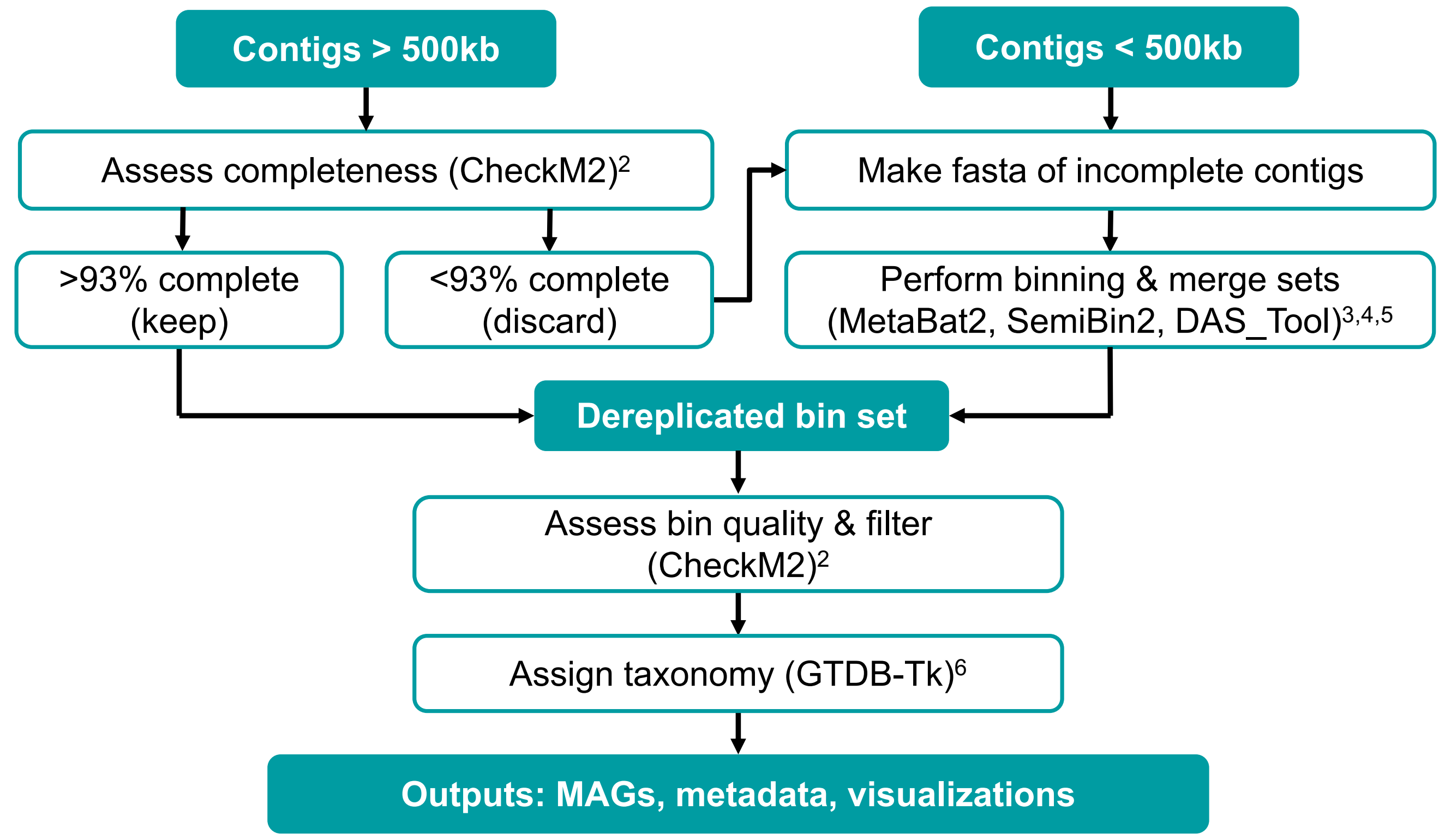
Here, we present the newest version of the HiFi-MAG-Pipeline (v2.0), a comprehensive workflow that automates major steps including binning, quality filtering, and taxonomic identification.

HiFi-MAG-Pipeline

Completeness-aware binning strategy

- Standard binning assumes genomes (MAGs) are fragmented and occur as multiple contigs.
- This causes unexpected behavior – long, complete contigs can be mis-binned with additional contigs, inflating the contamination score and causing removal during filtering steps.
- The completeness-aware strategy begins by extracting long, complete contigs – all contigs >500kb are assessed using CheckM2 for completeness.
- For all remaining contigs, a multi-binning strategy is used, and the bin sets are de-replicated and merged.
- The binned contigs are added to the set of long complete contigs, and the combined set is filtered to extract high-quality MAGs.

HiFi-MAG-Pipeline (v2.0)



Contigs > 500kb
Assess completeness (CheckM2)[2]
>93% complete (keep)
<93% complete (discard)
Contigs < 500kb
Make fasta of incomplete contigs
Perform binning & merge sets (MetaBat2, SemiBin2, DAS_Tool)[3,4,5]
Dereplicated bin set
Assess bin quality & filter (CheckM2)[2]
Assign taxonomy (GTDB-Tk)[6]
Outputs: MAGs, metadata, visualizations

Figure 2. Overview of the completeness-aware binning strategy in HiFi-MAG-Pipeline v2.0.

Benchmarking

We assembled 10 publicly available HiFi metagenomic datasets[7] with hifiasm-meta and performed binning using a standard tool (MetaBat2) or HiFi-MAG-Pipeline v2.0.

Organism	Dataset	HiFi Reads	Avg Read Length	Total Data	Median QV
Environmental	Photobioreactor	1.41 M	3.2 kb	4.6 Gb	Q40
	Hot spring sediment	2.69 M	10.3 kb	27.9 Gb	Q31
	Activated sludge	0.99 M	15.4 kb	15.3 Gb	Q35
Sheep	Sheep gut	11.84 M	11.2 kb	206.5 Gb	Q35
Human	French gut	1.64 M	7.9 kb	13.0 Gb	Q35
	Korean gut	2.01 M	14.6 kb	29.6 Gb	Q34
	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

Results

HiFi assemblies produce many high-quality MAGs

- Recovered 60–325 MAGs per sample
- Found 33–193 MAGs (up to 65%) are single-contig (Figs. 3, 4)

HiFi-MAG-Pipeline yields more total MAGs than other standard methods

- Found 14–67% increase in total MAGs (Fig. 3)
- Gain of 12–120 total MAGs per sample

Completeness-aware binning rescues single-contig, complete MAGs

- Found 10–142% increase in single-contig, complete MAGs (Figs. 3, 4)
- Incomplete long contigs are successfully binned

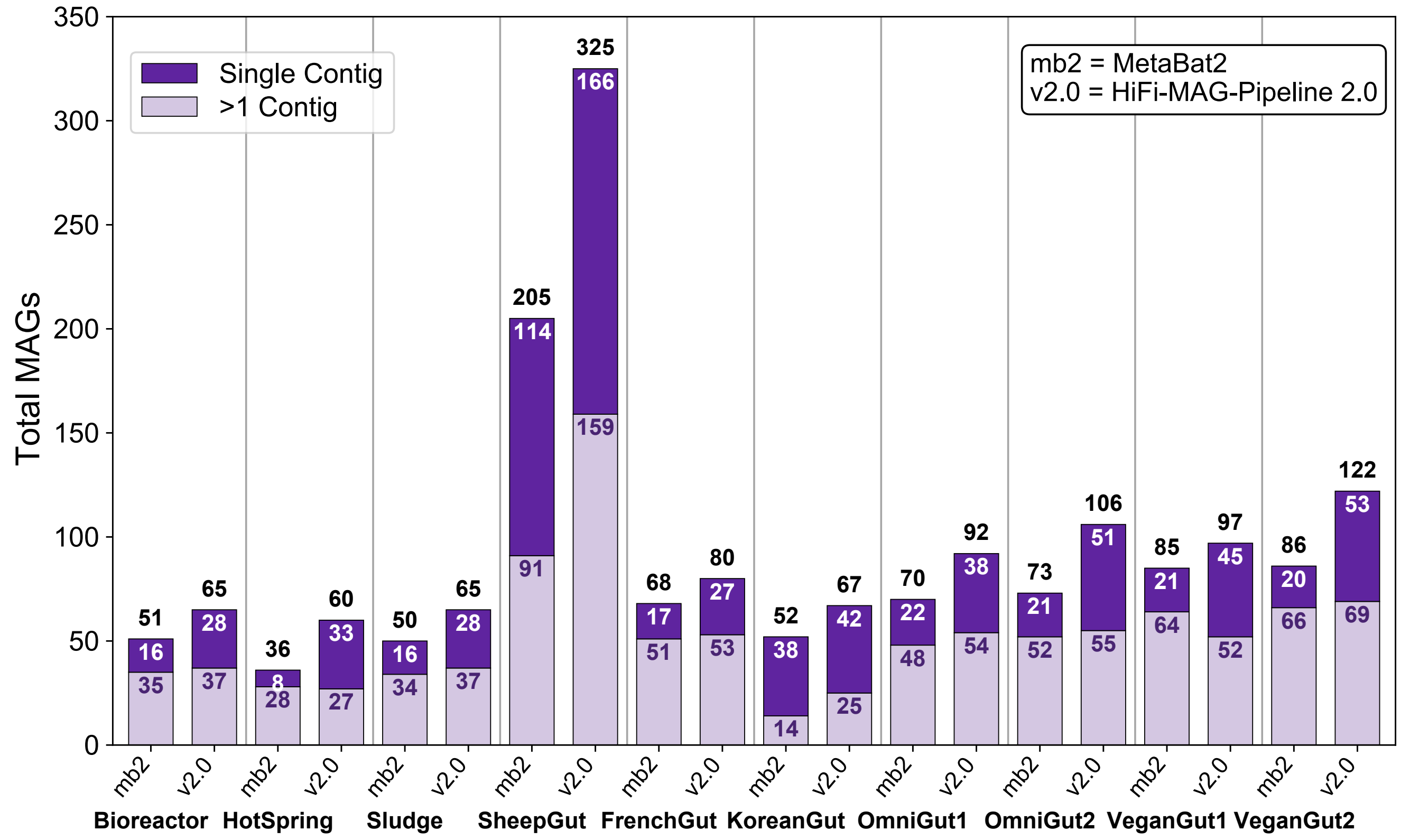


Figure 3. MAG yields from standard binning with MetaBAT2 (mb2) vs. HiFi-MAG-Pipeline (v2.0). Dark purple represents single-contig circular MAGs and light purple represents MAGs containing >1 contigs (all with >70% completeness, <10% contamination). Numbers in the stacked bars represent each category, and numbers above represent total MAGs.

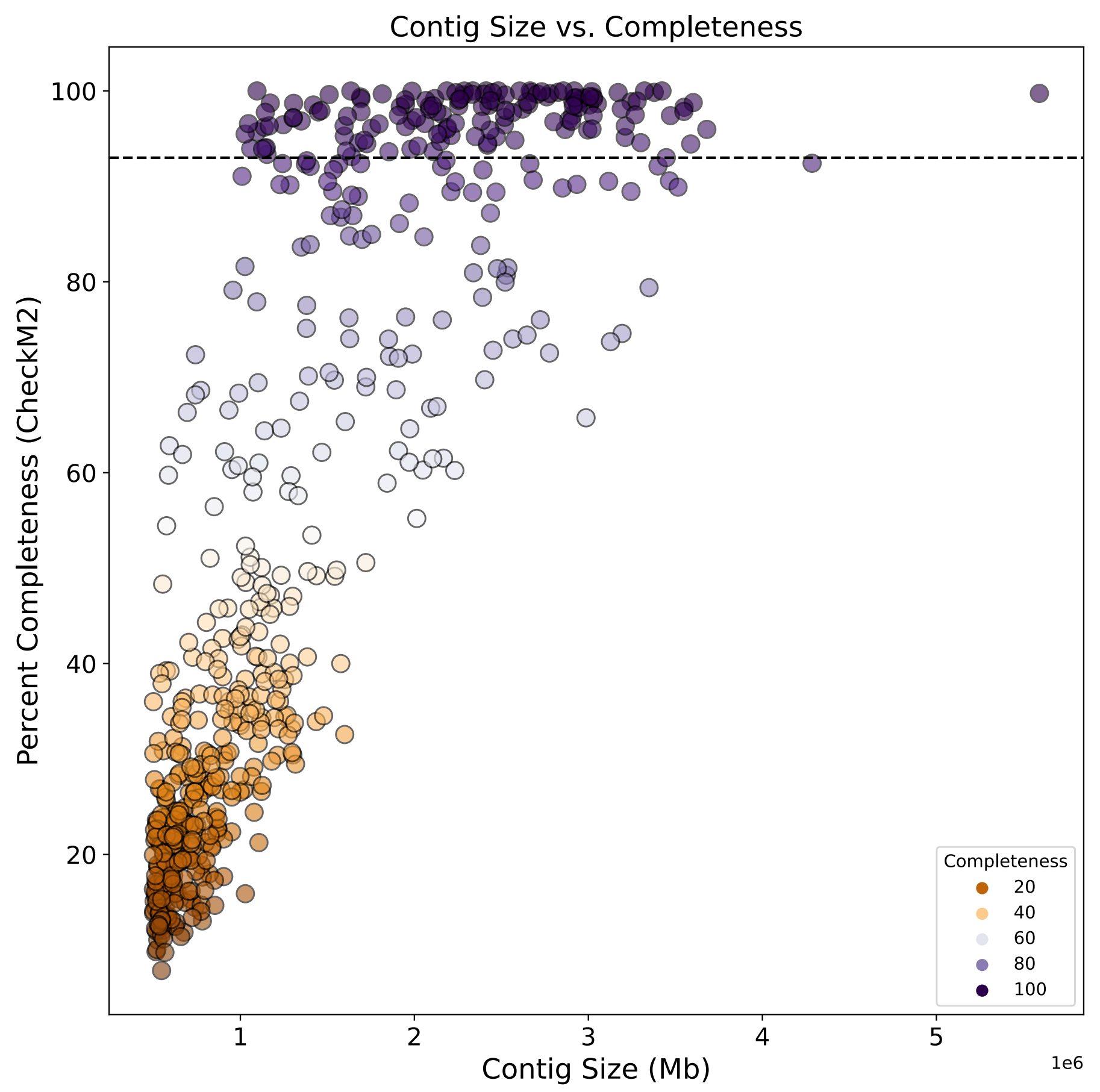


Figure 4. Relationship between contig size and completeness scores (using CheckM2) for the sheep gut assembly. Identifying long, complete contigs is the first step of the completeness-aware binning strategy. In this dataset, there are 147 long contigs with high completeness (93–100%). After their initial identification, the long complete contigs are moved to the final bin set and forego binning. Binning is then performed on all remaining contigs.

Outputs

HiFi-MAG-Pipeline produces several informative figures displaying quality characteristics for MAGs recovered (Fig. 5). It also provides metadata from CheckM2 and GTDB-Tk, and all MAG sequences are provided as individual fasta files for downstream analysis.

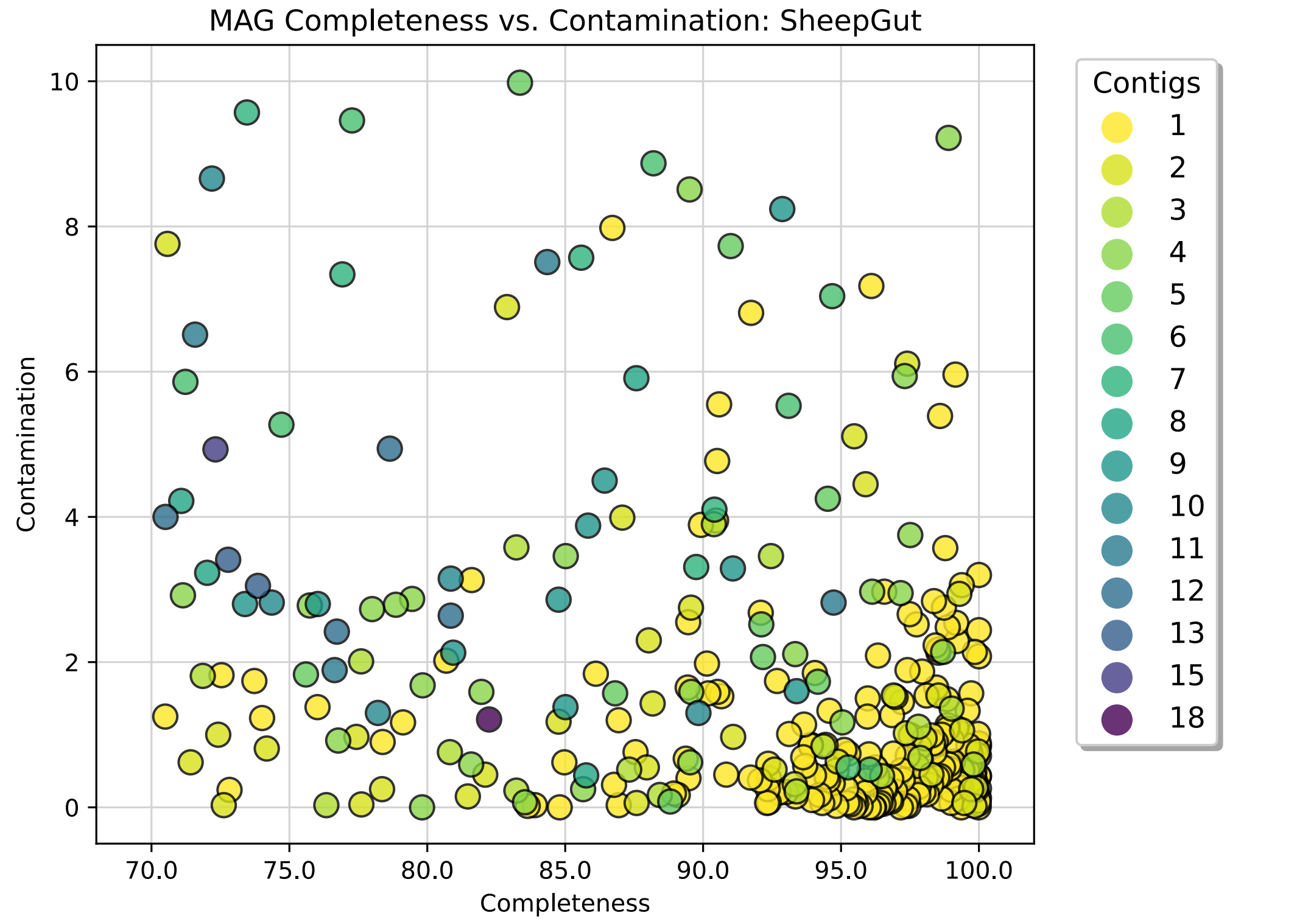


Figure 5. Completeness versus contamination scores for 325 high-quality MAGs found by HiFi-MAG-Pipeline for the sheep gut assembly. Each dot represents a MAG, and colors indicate the number of contigs contained in the MAG. We found 156 MAGs (48%) displayed >95% completeness, with 126 being single contig.

Accessibility

- All PacBio metagenomics pipelines are open-source and publicly available on github:

PacificBiosciences / **pb-metagenomics-tools**

Conclusions

- PacBio HiFi sequencing offers major advantages for metagenome assembly.
- Complete, single-contig MAGs can be routinely assembled from HiFi reads (33-62% of total MAGs).
- The HiFi-MAG-Pipeline automates all key steps required to obtain high-quality MAGs from long-read metagenome assemblies.
- Completeness-aware binning recovers substantially more MAGs than other methods (67% increase in total MAGs, 142% increase in single-contig MAGs).
- HiFi sequencing is an effective strategy for obtaining large numbers of high-quality MAGs, particularly for uncultured and uncharacterized species.

References

1. Feng, X., et al. 2022. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, 19: 671–674.
2. 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
3. Kang, D.D., et al. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7: e7359.
4. Pan et al. 2023. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *bioRxiv*, https://doi.org/10.1101/2023.01.09.523201
5. Sieber, C.M.K., et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3: 836–843.
6. Chaumeil, P.-A., et al. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 35: 1925–1927.
7. https://github.com/PacificBiosciences/pb-metagenomics-tools/blob/master/docs/PacBio-Data.md