Mapler: a pipeline for assessing assembly quality in taxonomically rich metagenomes sequenced with HiFi reads

Nicolas Maurice, Claire Lemaitre, Riccardo Vicedomini, Clémence Frioux

Bioinformatics, 2025, **41(6)**, btaf334
https://doi.org/10.1093/bioinformatics/btaf334
Advance Access Publication Date: 6 June 2025
Applications Note

OXFORD

Sequence analysis

Mapler: a pipeline for assessing assembly quality in taxonomically rich metagenomes sequenced with HiFi reads

Nicolas Maurice[1,2,*], Claire Lemaitre[1], Riccardo Vicedomini[1], Clémence Frioux[2,*]

[1]Univ Rennes, Inria, CNRS, IRISA – UMR 6074, F-35000 Rennes, France
[2]Inria, University of Bordeaux, INRAE, F-33400 Talence, France

*Corresponding authors. Nicolas Maurice, Inria, Université de Rennes, CNRS, IRISA, UMR 6074, 263 avenue du Général Leclerc, Rennes F 35000, France. E-mail: nicolas.maurice@inria.fr; Clémence Frioux, Inria, Univ. Bordeaux, INRAE, 200 avenue de la Vieille Tour, Talence F 33400, France. E-mail: clemence.frioux@inria.fr.

Associate Editor: Can Alkan

Abstract

Summary: Metagenome assembly seeks to reconstruct the most high-quality genomes from sequencing data of microbial ecosystems. Despite technological advancements that facilitate assembly, such as Hi-Fi long reads, the process remains challenging in complex environmental samples consisting of hundreds to thousands of populations. Mapler is a metagenome assembly and evaluation pipeline with a focus on evaluating the quality of Hi-Fi long read metagenome assemblies. It incorporates several state-of-the-art metrics, as well as novel metrics assessing the diversity that remains uncaptured by the assembly process. Mapler facilitates the comparison of assembly strategies and helps identify methodological bottlenecks that hinder genome reconstruction.

Availability and implementation: Mapler is open source and publicly available under the AGPL-3.0 licence at https://github.com/Nimauric/Mapler. Source code is implemented in Python and Bash as a Snakemake pipeline. A snapshot of the code is available on Software Heritage at swh:1:snp:df4f5f02e22ebbab285ec14af58d4d88436ee5d6. Raw data and results are available at https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/2SA8AB.

1 Introduction

Evaluating the quality of metagenome assemblies can be a challenging task, especially when no reference genome is available and when comparing samples with varying taxonomic richness and sequencing depths. Taxonomic richness refers to the number of distinct populations within the sample: microbial communities may consist of only a handful of populations, as in acid mine drainage communities (Tyson *et al.* 2004), or of up thousands of distinct populations, as observed in soil ecosystems (Roesch *et al.* 2007). Assembly of metagenomic reads leads, in the best case scenario, to the reconstruction of genomes, but in most cases, to the generation of genome fragments of varying lengths called *contigs*. Those are then grouped into *bins*, presumed to originate from the same microbial populations; bins of sufficient quality are referred as *Metagenome-Assembled Genomes* (MAGs) (Cerk *et al.* 2024). A high-quality metagenome assembly is not only expected to yield high-quality bins, but also to be representative of the majority of the read sequences. Recent studies showed significant improvements in both the number and quality of bins obtained using highly accurate PacBio HiFi long reads (Benoit *et al.* 2024). However, in highly taxonomically rich ecosystems, assembly methods still struggle to reconstruct the numerous low-abundance genomes (Xu *et al.* 2021, Benoit *et al.* 2024), and it remains unclear how much of the sample these bins are representative of, resulting in a need for comparison and development of dedicated evaluation methods.

Several tools and pipelines exist to evaluate metagenomes. CheckM2 (Chklovski *et al.* 2023) assesses binned contigs based on the presence of marker genes, allowing the identification of MAGs from bins with low contamination and high completeness scores. The PacBio HiFi-MAG-Pipeline (Portik *et al.* 2024) is a pipeline developed to identify high-quality MAGs from previously generated metagenome assemblies. It follows a "completeness-aware" strategy based on CheckM2 and several state-of-the-art binning tools, incorporating stringent filtering criteria to exclude low-quality bins, which are common in taxonomically rich ecosystems. MetaQUAST (Mikheenko *et al.* 2016) performs a reference-based evaluation, either using user-defined references or retrieving references via taxonomic assignment. However, in complex ecosystems, many species are absent from databases, thus limiting its effectiveness. Finally, custom metrics or visualizations have also been used for method validation (Benoit *et al.* 2024, Feng and Li 2024). For example, the percentage of reads aligned to the assembly has been used to validate metagenome assembly in (Benoit *et al.* 2024). Nevertheless, these approaches are rarely documented nor provided in an easy-to-use implementation that allows for replication on new datasets.

Received: 7 March 2025; **Revised:** 26 May 2025; **Editorial Decision:** 1 June 2025; **Accepted:** 4 June 2025

In this work, we present Mapler, a metagenomic assembly and evaluation pipeline. It avoids filtering out any sequences, it does not rely on the availability of reference sequences, and it considers both unassembled reads and unbinned contigs. Mapler integrates several state-of-the-art tools as well as novel metrics and visualizations based on read-to-contig alignments. It provides a broad view of the sequence characteristics after assembly and binning, in order to identify the bottlenecks faced during bioinformatic processes. Mapler is therefore an effective way to examine assembly in taxonomically rich ecosystems, where high-quality bins and references are scarce.

2 Software description

2.1 Pipeline

Mapler is a Snakemake (Mölder *et al.* 2021) pipeline dedicated to the evaluation of taxonomically rich metagenome assemblies of HiFi long reads. It can be run either locally or on Slurm-based computing environments. Its modular design allows for easy integration of additional custom steps in the pipeline or modification of existing ones. The pipeline can run multiple steps in parallel, including analyzing multiple samples at once. The structure of the pipeline is illustrated in Fig. 1A.

2.2 Integrated tools

While its focus being on evaluation, Mapler integrates state-of-the-art tools for assembly and binning suitable for HiFi sequencing data: metaMDBG (Benoit *et al.* 2024), hifiasm-meta (Feng *et al.* 2022), metaFlye (Kolmogorov *et al.* 2020), OPERA-MS (Bertrand *et al.* 2019), and MetaBAT 2 (Kang *et al.* 2019). Users may alternatively skip the assembly and/or binning steps by providing their own input contigs and/or bins. Each bin can be taxonomically classified using either GTDB-Tk (Chaumeil *et al.* 2022), or Kraken 2 (Wood *et al.* 2019) in order to facilitate the comparison with the taxonomic assignment of the reads. By default, bins are qualitatively assessed with CheckM2 (Chklovski *et al.* 2023) and categorized according to the following levels of completeness (comp.) and contamination (cont.): near complete (single contig, ≥99% comp., ≤1% cont.), high quality (≥90%

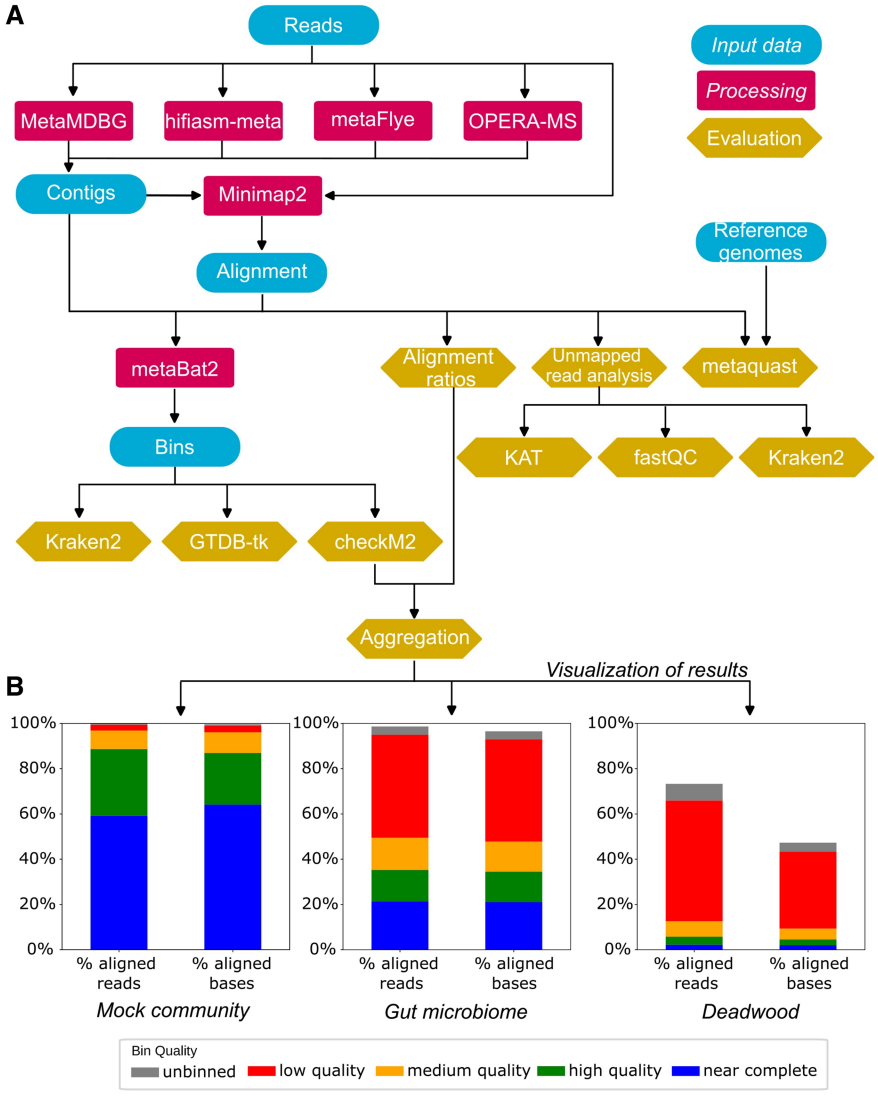

Figure 1. (A) Overview of the Mapler pipeline. Contigs and bins can either be generated by the pipeline or given as input. Several long-read assemblers are integrated in Mapler. (B) Example of Mapler's output. Histograms show the aligned read/base percentages for bins of different quality and reveal the increasing complexity of different ecosystems, from the mock community to the gut microbiome sample and the highly diverse deadwood sample, all three assembled with metaMDBG.

comp., ≤5% cont.), medium quality (≥50% comp., ≤10% cont.), and low quality for the remaining bins. These criteria match the completeness and contamination estimates used by the Genomic Standards Consortium (Bowers *et al.* 2017) for defining low-quality to high-quality MAGs. MetaQUAST (Mikheenko *et al.* 2016) is also integrated to compare contigs with reference genomes, if available and provided as input by the user.

2.3 Novel metrics

Mapler aligns the reads on the contigs with Minimap2 (Li 2018), and uses these alignments to calculate various metrics. The *aligned read percentage* is the number of reads aligned to at least one contig divided by the total number of reads, while the *aligned base percentage* is the number of read bases aligned to at least one contig, divided by the total length of the reads. These metrics can be computed on all Minimap2 alignments (default behavior) or only on alignments whose length is above a user-defined threshold. These metrics can be computed with or without the binning information. In the former case, the percentage is separately calculated for reads or bases that align to contigs belonging to bins of near complete, high, medium or low quality, or to contigs that were assembled but not binned. A text report is produced for both the binning-aware and binning-unaware versions, and a summarizing plot is generated for the binning-aware version (Fig. 1B). In cases where a read is aligned to multiple contigs, it is only taken into account for the highest bin quality level.

Another analysis proposed by Mapler is the comparison of the sets of reads aligned or unaligned to the contigs, in order to gain insight into the characteristics of reads that participate in, or have been excluded from the assembly. Both sets are analyzed separately with the following tools:

- FastQC (https://github.com/s-andrews/FastQC), used to assess read quality and generate a comprehensive report. It can be used to check whether the assembly process is more effective on higher quality reads, longer reads, or reads with a certain GC ratio.
- K-mer Analysis Toolkit (KAT) (Mapleson *et al.* 2017), which computes the abundance of assembled and unassembled reads. The abundance of a given read is estimated by its median k-mer abundance, with k-mer abundances being computed from the full read dataset. Mapler integrates these results to visualize both distributions with two overlapping histograms.
- Kraken 2 (Wood *et al.* 2019), alongside Krona (Ondov *et al.* 2011), is used to analyze the taxonomic composition and abundance of both sets of reads, providing insight on over- or under-represented clades in the assembly.

3 Application

We demonstrated Mapler's ability to evaluate assemblies of diverse samples on three datasets of increasing taxonomic complexity, sequenced with PacBio Sequel II SMRT.

- *Mock community*: the ZymoBIOMICS Gut Microbiome Standard D6331 (SRR13128014) consists of 21 populations, including 17 species and 5 strains of *Escherichia coli*. The sample contains 18.0 Gbp spread over 1 978 852 reads.
- *Gut microbiome*: a pooled extraction of four stool samples from adult humans following a vegan diet (SRR15275211). Human digestive microbiomes generally host a few hundreds of species. The sample contains 18.8 Gbp spread over 1 904 159 reads.
- *Deadwood*: four separately sequenced samples of deadwood that were co-assembled as in Richy *et al.* (2024). The samples (SRR28211698 to SRR28211701) contain a total of 16.1 Gbp spread over 866 007 reads.

Each dataset was processed by Mapler with metaMDBG, hifiasm-meta, and metaFlye.

Mapler first summarizes in scatter plots the bins obtained in each sample (Fig. 1, available as supplementary data at *Bioinformatics* online), highlighting that the number and quality of bins vary across the datasets. Compared to the *Mock community*, the number of low-quality bins is much higher in the *Gut community* and *Deadwood* samples, due to either low completeness or high contamination scores.

Mapler then generates, after mapping reads to contigs and bins, plots that highlight a decreasing proportion of reads assembled and binned at each quality level as dataset complexity increases (Fig. 1B). More precisely, on the metaMDBG assemblies, 96.9% of reads and 96.1% of bases map to bins of at least medium quality in the *Mock community*, while in the *Gut microbiome* these values drop to 49.5% and 47.8%, respectively. Furthermore, *Deadwood*'s high diversity and lower sequencing depth result in a lower-quality assembly with only 12.6% of reads and 9.4% of bases aligned with bins of medium quality or higher.

Because a significant proportion of reads of the *Deadwood* sample did not participate in the assembly (26.7% of reads and 52.7% of bases did not align with any contig), we compared the aligned and unaligned reads in this sample. Despite the read length variation in the original sample, assembled and unassembled reads are of similar length (18 437 and 18 583 base pairs on average, respectively, see Fig. 2, available as supplementary data at *Bioinformatics* online). Taxonomic assignment of reads with Mapler illustrates that some microbial populations were only detected in unassembled reads, such as several species of *Legionella* (Fig. 3, available as supplementary data at *Bioinformatics* online). Sequences were also generally assigned with less precision in the unassembled reads (56% of bacteria are assigned at the phylum level in the unassembled reads, compared to 76% in the assembled reads), meaning that the unassembled populations were less well represented in the taxonomic database. Gaps in the reference database can be explained by rare or otherwise hard to assemble taxa, or by missing representatives in the standard Kraken2 database (e.g. protists, viruses, plasmids, other mobile genetics elements). Additionally, some unaligned reads share the same taxonomic assignment as aligned reads, possibly representing partially assembled genomes, strain variants, or genomic islands from low-abundance strains. As expected, unassembled reads were mostly made up of rare k-mers: nearly all unassembled reads have a median k-mer abundance as low as 1 (Fig. 4, available as supplementary data at *Bioinformatics* online). These results suggest that assemblers and binners used in the metagenome analysis could not improve the results by much, and that a deeper sequencing would rather be needed to enhance the quality of the assembly. We nonetheless compared the *Deadwood* assemblies performed with different metagenome assembly tools. MetaMDBG outperformed the other

assemblers in terms of total captured diversity: 52.7% of bases do not align with any contig, compared to 76.7% for metaFlye and 66.1% for hifiasm-meta (Fig. 5, available as supplementary data at *Bioinformatics* online). Conversely, hifiasm-meta outperformed metaMDBG in terms of bases aligned to at least medium-quality bins (14.2% versus 12.6%).

We recorded the execution time of the pipeline on the three datasets. For each dataset, we executed the pipeline on a Intel (R) Xeon(R) CPU E5-2670 v3 @ 2.30 GHz node, allocating a total of 48 CPUs and 200G of memory. The detailed breakdown of how much memory was allocated to each substep of the pipeline is described in Table 1, available as supplementary data at *Bioinformatics* online. We performed the evaluations on the three samples separately. For each sample, we evaluated the time it took to evaluate the assembly and binning quality of three assemblers (the assembly and binning was performed separately). The analysis of the *Mock community* took 2 h and 4 min in wall-clock time to run, followed by the *Gut microbiome* with 3 h and 46 min and, finally, the *Deadwood* with 5 h and 3 min.

4 Conclusion

Mapler is a metagenome evaluation pipeline that allows a thorough examination of assembly and binning results, implemented in an easy-to-use and customizable workflow. Mapler is specifically implemented to analyze HiFi long-read datasets that are currently the most suitable to characterize taxonomically rich microbial ecosystems. On top of integrating multiple state-of-the-art evaluation methods, Mapler incorporates new evaluation metrics such as the aligned read and aligned base percentages. When combined with the bin quality information, these metrics provide a way to measure how much of the sample's original diversity was assembled at each level of quality, and highlight potential assembly issues. In cases where a significant proportion of reads cannot be aligned back to the assembly, comparing the assembled and unassembled reads provides further insight into the reasons why the assembly may not be sufficiently representative of the sample, and whether the contigs are missing key taxa that are only present in the reads.

Acknowledgements

We acknowledge the GenOuest bioinformatics core facility (https://www.genouest.org) for providing the computing infrastructure. A CC-BY public copyright license (https://creativecommons.org/licenses/by/4.0/) has been applied by the authors to the present document, in accordance with the grant's open access conditions.

Author contributions

Nicolas Maurice (Conceptualization [equal], Investigation [lead], Methodology [lead], Software [lead], Visualization [lead], Writing—original draft [equal], Writing—review & editing [equal]), Claire Lemaitre (Conceptualization [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Riccardo Vicedomini (Conceptualization [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Clémence Frioux (Conceptualization [equal], Supervision [equal], Visualization [supporting], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the French National Research Agency (ANR) France 2030 PEPR Agroécologie et Numérique MISTIC ANR-22-PEAE-0011.

References

Benoit G, Raguideau S, James R *et al.* High-quality metagenome assembly from long accurate reads with metaMDBG. *Nat Biotechnol* 2024;**42**:1378–83. https://doi.org/10.1038/s41587-023-01983-6

Bertrand D, Shaw J, Kalathiyappan M *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* 2019;**37**:937–44. https://doi.org/10.1038/s41587-019-0191-2

Bowers RM, Kyrpides NC, Stepanauskas R *et al.*; Genome Standards Consortium. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;**35**:725–31. https://doi.org/10.1038/nbt.3893

Cerk K, Ugalde-Salas P, Nedjad CG *et al.* Community-scale models of microbiomes: articulating metabolic modelling and metagenome sequencing. *Microb Biotechnol* 2024;**17**:e14396. https://doi.org/10.1111/1751-7915.14396

Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;**38**:5315–6. https://doi.org/10.1093/bioinformatics/btac672

Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;**20**:1203–12. https://doi.org/10.1038/s41592-023-01940-w

Feng X, Cheng H, Portik D *et al.* Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat Methods* 2022;**19**:671–4. https://doi.org/10.1038/s41592-022-01478-3

Feng X, Li H. Evaluating and improving the representation of bacterial contents in long-read metagenome assemblies. *Genome Biol* 2024;**25**:92. https://doi.org/10.1186/s13059-024-03234-6

Kang DD, Li F, Kirton E *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359. https://doi.org/10.7717/peerj.7359

Kolmogorov M, Bickhart DM, Behsaz B *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–10. https://doi.org/10.1038/s41592-020-00971-x

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191

Mapleson D, Garcia Accinelli G, Kettleborough G *et al.* KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 2017;**33**:574–6. https://doi.org/10.1093/bioinformatics/btw663

Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016;**32**:1088–90. https://doi.org/10.1093/bioinformatics/btv697

Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with Snakemake. *F1000Res* 2021;**10**:33. https://doi.org/10.12688/f1000research.29032.2

Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* 2011;**12**:385. https://doi.org/10.1186/1471-2105-12-385

Portik DM, Feng X, Benoit G *et al.* Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods. BioRxiv, https://doi.org/10.1101/2024.05.10.593587, 2024.

Richy E, Thiago Dobbler P, Tláskal V *et al.* Long-read sequencing sheds light on key bacteria contributing to deadwood decomposition processes. *Environ Microbiome* 2024;**19**:99. https://doi.org/10.1186/s40793-024-00639-5

Roesch LFW, Fulthorpe RR, Riva A *et al.* Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* 2007;**1**:283–90. https://doi.org/10.1038/ismej.2007.53

Tyson GW, Chapman J, Hugenholtz P *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004;**428**:37–43. https://doi.org/10.1038/nature02340

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**:257. https://doi.org/10.1186/s13059-019-1891-0

Xu L, Dong Z, Chiniquy D *et al.* Genome-resolved metagenomics reveals role of iron metabolism in drought-induced rhizosphere microbiome dynamics. *Nat Commun* 2021;**12**:3209. https://doi.org/10.1038/s41467-021-23553-7

Bioinformatics, 2025, 41, 1–5
https://doi.org/10.1093/bioinformatics/btaf334
Applications Note