Metagenomic strain-level classification based on reference genome database reduction

Josipa Lipovac, MSc
Supervisor: Asst. Prof. Krešimir Križanović, PhD
University of Zagreb Faculty of Electrical Engineering and Computing

UNIVERSITY OF ZAGREB
Faculty of Electrical Engineering and Computing

1. Introduction

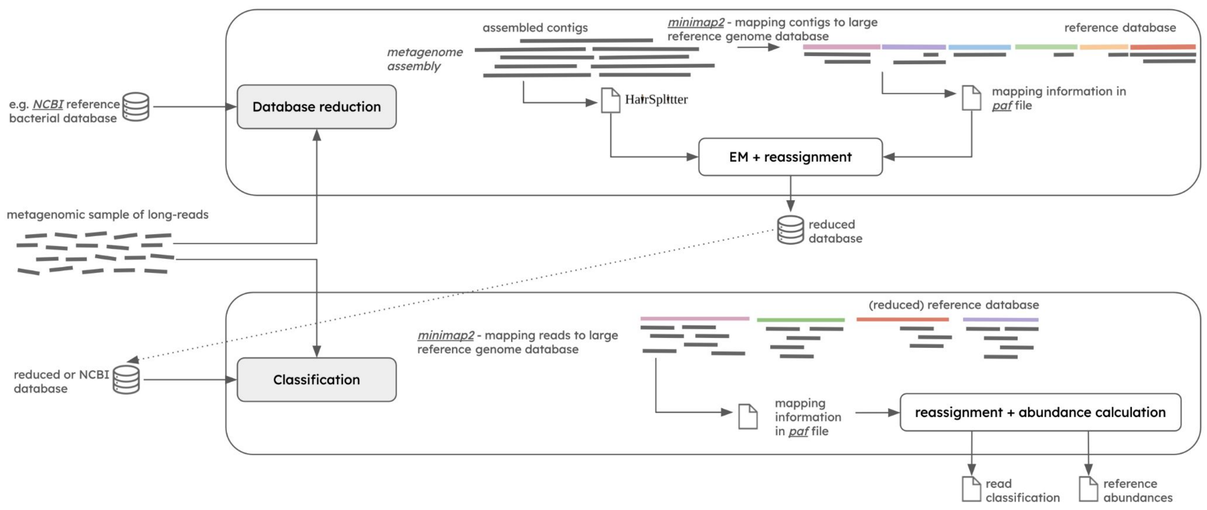
Metagenomics allows the direct study of genetic material from complex microbial communities, providing insights into their diversity, functions, and interactions. Long-read sequencing technologies further enhance metagenomic analysis by enabling better genome assembly and strain-level resolution. Strain-level identification is critical, as different strains can exhibit distinct metabolic activities and pathogenic potential, influencing both ecological dynamics and clinical outcomes.

2. Problem Description

Many metagenomic classification tools perform well at the species level, but strain-level classification remains challenging. Strains are highly similar, making them difficult to distinguish, especially when using large reference databases. Existing tools often require prior species identification, struggle with scalability, or cannot resolve closely related strains accurately. Mapping-based methods improve precision but are computationally intensive on large databases. There is a need for scalable, precise, and efficient approaches that enable direct strain-level classification without prior assumptions.

3. Methodology

MADRe is a two-step pipeline for long-read strain-level metagenomic classification enhanced with **M**etagenome **A**ssembly-Driven **D**atabase **Re**duction.



1) *database reduction step* - uses long-read assemblies and an EM-based contig-to-reference mapping strategy to identify relevant strains and build a reduced reference set

mapping score:

$$\mathrm{H} = 2 * \frac{\mathrm{N} \times ml}{\mathrm{N} + ml} \qquad \mathrm{S}(c_i, r_j) = \sum_{m \in M(c_i, r_j)} \mathrm{H}(p)$$

E step: $P(r_i \mid c_i) = \frac{P(c_i \mid r_i) \cdot P(r_i)}{\sum_{j=1}^{G} P(c_i \mid r_j) \cdot P(r_j)}$

M step: $P(y_i) = \frac{\sum_{j=1}^{X} P(r_i \mid c_j)}{|M|}$

2) *reads classification step* - reads are mapped to the reduced database, and ambiguous mappings are resolved based on mapping scores and probabilistic support.

mapping score:

$$\frac{N}{ml}$$

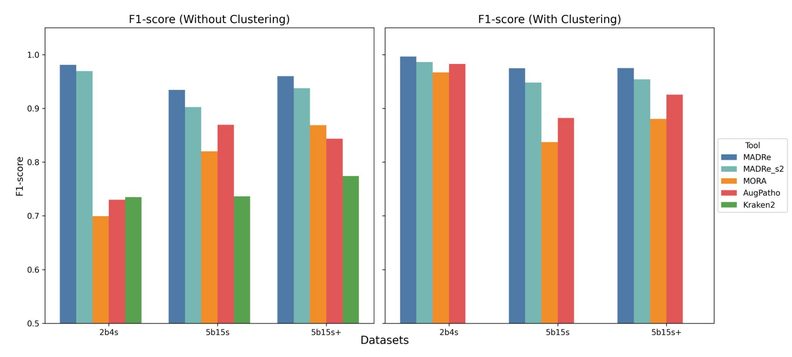
Both steps can also be run independently, allowing flexible use depending on available sample information.

4. Results

We evaluated MADRe on simulated datasets, Zymo mock communities and a real anaerobic digester sludge metagenome.

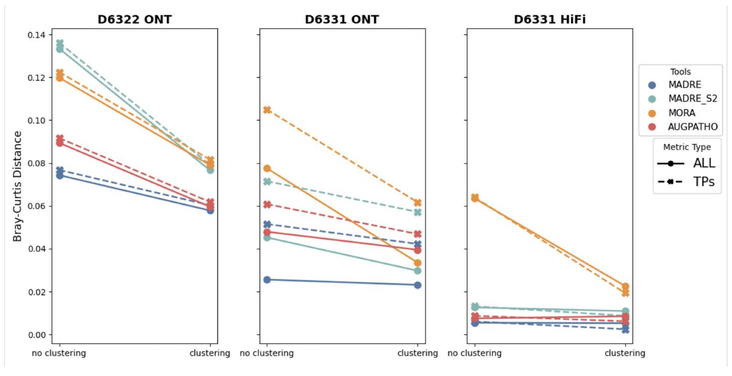
Simulated datasets

MADRe achieved higher F1 scores and lower false positive rates compared to state-of-the-art tools.



Zymo mock communities

MADRe achieved the lowest Bray–Curtis distance when comparing obtained read count abundances with expected ones.

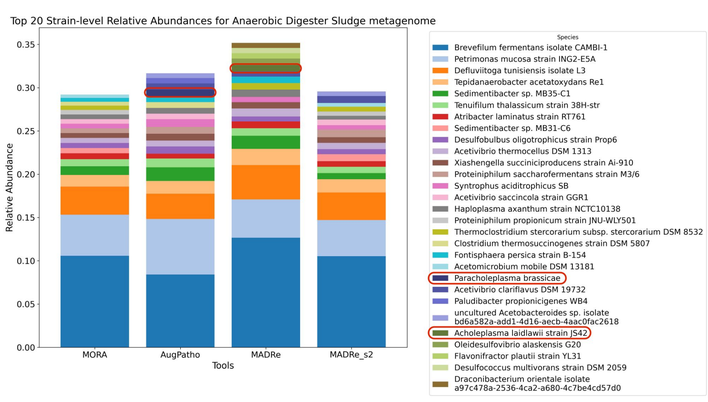


$$\mathrm{BC}(x, y) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)}$$

Applying post-classification clustering of highly similar strains further improved the performance of all tools, particularly highlighting MADRe's ability to distinguish highly similar strains.

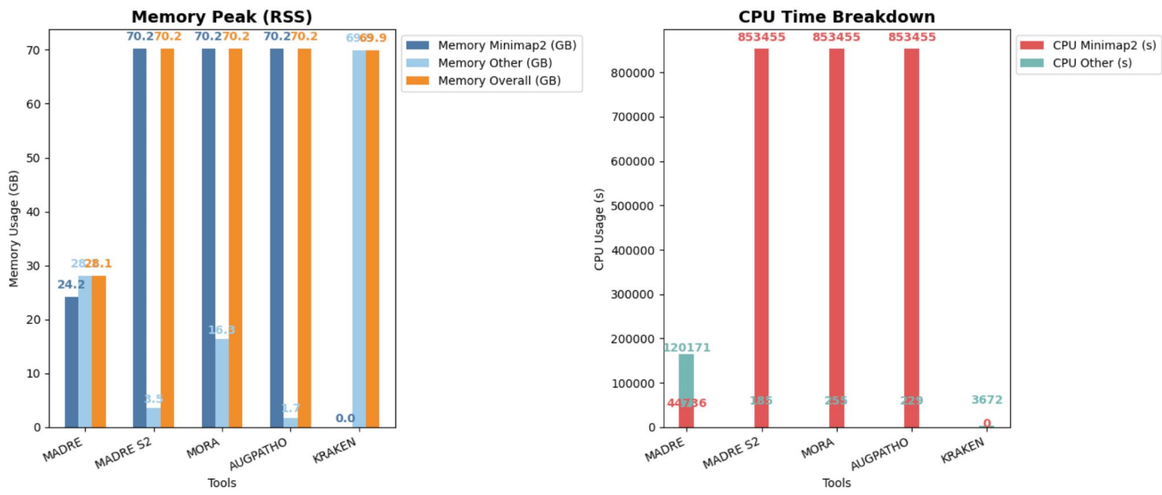
Real metagenome data

MADRe classifies fewer strains and focuses on a confident subset of dominant organisms



Two strains are highlighted in red to illustrate cases where different tools classified reads originating from an unrepresented reference to distinct false positives that share similar genomic regions.

Time and memory performance on Zymo mock community



MADRe is resource efficient:
Lower memory usage and faster runtimes compared to other strain-level classification tools.

5. Conclusion

MADRe is designed to operate with large, diverse databases spanning multiple taxonomic levels, enabling high-resolution strain-level classification while minimizing false positives.

MADRe is open source and publicly available:

Acknowledgment

The funding was provided by the Croatian Science Foundation under grant IP-2018–01-5886 (SIGMA).

HrZZ Croatian Science Foundation

References

[1] Zheng, A., Shaw, J., & Yu, Y. W. (2024). Mora: abundance aware metagenomic read re-assignment for disentangling similar strains. *BMC bioinformatics*, *25*(1), 161.
[2] Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., ... & Johnson, W. E. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, *2*, 1-15.
[3] Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, *20*, 1-13.

PhD Day, June 5, 2025

Contact

Josipa Lipovac, MSc
josipa.lipovac@fer.hr