[BIOINFO 2024]

Comprehensive benchmarking study of metagenomic binning tools across diverse datasets

Jungyeon Kim[1] and Insuk Lee[1,*]

[1] *Department of Biotechnology, College of Life Science & Biotechnology, Yonsei University, Seoul 03722, Republic of Korea*
**Corresponding author: insuklee@yonsei.ac.kr*

Metagenomic binning is a crucial step in reconstructing individual microbial genomes from complex communities, enabling researchers to better understand microbial diversity and function. However, most established catalogs continue to rely on traditional, standardized pipelines even though a wide range of binning tools, from probabilistic models to neural network-based algorithms, have been developed. In this study, we perform a comprehensive benchmarking of several leading metagenomic hybrid binning tools, evaluating their performances across a diverse range of datasets, including samples from soil, ocean, and human gut environments. In addition, to explore the impact of dataset characteristics on binning performance, simulated microbial datasets with varying sequencing depths and complexities were created and utilized, along with the first CAMI challenge datasets. By evaluating each tool's precision, recall, ARI, and the number of high-quality genomes, this study offers insights into the strengths and limitations of current binning approaches. The results show that certain neural network-based methods, particularly those incorporating multi-split binning modes or data augmentation steps, tend to outperform others across most datasets. Moreover, increased sequencing depth is shown to contribute to improved binning accuracy to a certain degree. Based on these findings, the new proposed optimized pipeline demonstrates the ability to reconstruct a greater number of sample-specific MAGs while also significantly improving the taxonomic diversity of the recovered genomes. As a result, this study provides a valuable guide in selecting the most appropriate binning tool for their metagenomic analyses and highlights which computational algorithms are most effective for binning.