DOCTORAL THESIS

Advanced Computational Methods to Decipher Microbial Dark Matter Using Metagenomic Sequencing

ZHANG, Zhenmiao

Date of Award:
2024

Advanced Computational Methods to Decipher Microbial Dark Matter Using Metagenomic Sequencing

Zhenmiao Zhang

Abstract

Microbial communities are ubiquitous in diverse environments, such as soil, wastewater, and the gastrointestinal tracts of humans and animals. Studying the collective genomes in microbial communities, known as metagenomes, provides invaluable insights for understanding the microbial communities and their association with the environment and humans. However, the vast majority of microbes are difficult to isolate and culture in the laboratory, rendering the genomes of most microorganisms within the microbial communities as "microbial dark matter." Therefore, metagenomic sequencing, a technique that sequences the entire metagenome without the need for isolating individual microbes, has become indispensable for studying microbial communities and deciphering the microbial dark matter.

To reconstruct microbial genomes, metagenome assembly is a critical step that considers metagenomic sequencing reads as the input and produces longer genome sequences of the microbes, known as contigs. Short-read sequencing has been widely used for metagenome assembly; however, it only generates highly fragmented contigs owing to the limitations of short read length in resolving complex genomic regions, such as inter- and intra-species repeats. Emerging long-read sequencing technologies produce continuous long-reads that markedly improve the generation of complete genomes from microbial communities. However, the high cost associated with long-read sequencing hinders its widespread application, especially in large cohorts.

A microbial genome is commonly assembled into a number of contigs in the metagenome assembly. Contig binning is a subsequent step in reconstructing microbial genomes, where the contigs are grouped based on their sequence features into bins representing different microbial genomes, referred to as metagenome-assembled genomes (MAGs). Metagenome assembly produces a significant number of contigs that are shorter than 2 Kb, accounting for a substantial proportion of the total assembly length. However, the existing contig binning tools commonly exclude these short contigs as these cannot provide stable sequence features.

We performed a series of research works to advance the generation of MAGs in the areas of metagenome assembly and contig binning. We conducted a comprehensive benchmarking study involving 19 assembly tools commonly applied to metagenomic sequencing of datasets obtained from simulations, mock communities, or human gut microbiomes. We found that long-read assemblies resulted in the most contiguous genomes, but they could often miss many MAGs owing to insufficient sequencing depth. To explore a cost-effective approach for generating high-quality microbial genomes, we proposed a metagenome assembler called Pangaea, which

improved metagenome assembly using short-read data with long-range conceitedness. It employed a deep-learning-based co-barcoded read binning algorithm to assemble co-barcoded reads with similar sequence contexts and abundances; it also leveraged a multi-threshold reassembly strategy to refine assembly for low-abundance microbes. To address the challenge of binning short contigs, we developed DeepMetaBin, a contig binning method that utilizes a graph-based Gaussian mixture variational autoencoder to group short (1–2 Kb) and long (longer than 2 Kb) contigs. To further improve contig binning for ultra-short contigs (<1 Kb), we created METAMVGL, a multi-view graph-based contig binning algorithm that uses label propagation for integrating both assembly and paired-end graph information. We performed extensive experiments to evaluate the tools described in this thesis. The results showed that our proposed methods generally outperformed state-of-the-art tools, leading to an overall improvement in generating MAGs.

Keywords: Metagenome Assembly, Contig Binning, Variational Autoencoder, Label Propagation, Metagenome-assembled Genome