http://dx.doi.org/10.7124/bc.000AFF

Realistic assortment of novel metagenomics benchmarks with diverse biological and technological characteristics

N.K. Sharma[1], K. Chhugani[1], V. Munteanu[2], P. Skums[3,4], A. Zelikovsky[3], S. Mangul[1]

[1] University of Southern California
Los Angeles, CA, USA, 90007
[2] Technical University of Moldova
168, Stefan cel Mare Blvd., Chisinau, Republic of Moldova, MD-2004
[3] Georgia State University
Atlanta, GA, USA, 30302,
[4] University of Connecticut
352, Mansfield Rd., Storrs, CT, USA, 06269
niteshku@usc.edu, mangul@usc.edu

Aim. To address the scarcity of high-quality experimental benchmarks in metagenomics by developing a cost-effective, semi-real benchmark generation tool, MetaWiz, which can modify existing experimental data to create diverse and realistic metagenomic benchmarks. **Methods.** In this study, we employed a series of steps to modify mock data representing known microbial genomes or strains. These steps involved selection, downloading, preprocessing, mapping, calculating mapping statistics, selecting uniquely mapped reads, and making modifications to create diverse benchmarks. The data was downloaded in fasta/fastq formats from the open source repositories like NCBI, MG-RAST, and Github. We preserved the original data quality and length without filtering or trimming to maintain uniformity. For mapping, we used bwa-mem and Minimap2, running both tools with default parameters. We discarded unmapped reads and selected uniquely mapped reads, we filtered out supplementary and alternative alignment. To modify the benchmark dataset, we implemented several functionalities, including the adjustment of strain frequencies, converting long Pacbio reads into short reads, correcting erroneous bases in Nanopore ONT reads using genomes, followed by read length adjustment similar to Pacbio reads. **Results.** MetaWiz successfully generated diverse benchmarks from real metagenomic data, encompassing various biological and technological characteristics. MetaWiz effectively reduced raw data error rates through mapping and error correction. Specifically, we were able to modify the following datasets into multiple benchmarks:

Dataset #1: 9 strains of HIV and 2 strains of HIV (Illumina). We altered the frequency of viral strains and removed strain-specific reads.

Dataset #2: 11 strains of bacterial species (Illumina). We altered the frequency of strains and species and removed species-specific reads.

Dataset #3: 26 strains of Bacteria and Archaea (Illumina). We selected some of the species/strains from this dataset to mix with Dataset #2, to create multiple benchmarks with different combinations.

Dataset #4: 12 strains of bacterial strains (Pacbio and Nanopore). For Pacbio reads, we shortened the read length to 100bp, generating up to 40 million short reads. For Nanopore ONT reads, we corrected the errors using the genomes and shortened the read length to 100bp, generating up to 60 million short reads.

Conclusions. MetaWiz offers a cost-effective solution to the challenge of generating high-quality benchmarks in metagenomics. The tool's ability to create semi-real benchmarks from existing experimental data maintains crucial technological and biological aspects, providing advantages over purely computational simulations. This approach can be extended beyond viral sequences to various sequencing data, enhancing the comprehensiveness of metagenomics benchmarking studies across different biological contexts.

Keywords: metagenomics, benchmark generation, MetaWiz, microbial genomes, error correction.