King Abdullah University of Science and Technology

جامعة الملك عبدالله للعلوم والتقنية

Understanding Bioinformatic Pipelines

Final Project: Metagenomics Whole Genome Sequencing Data Analysis

Group 2:

Alejandra Velazquez | Rund Tawfiq | Jiayi Huang | Turki Sobahy

May 7th, 2024

1. INTRODUCTION

Metagenomics represents a transformative approach in microbiology that enables the analysis of microbial communities directly from environmental samples without the need for culture. The word metagenomics was first introduced in 1998 and is defined as 'evaluation of all the genetic materials isolated directly from the environmental samples' [1].

Traditional microbiological studies are often limited by the necessity of culturing organisms; it can overlook the vast majority of microbes that are unculturable under laboratory conditions [2]. Metagenomics, however, offers a powerful alternative by employing both targeted and shotgun sequencing techniques to explore microbial diversity, evolutionary relationships, functional activities, and interactions within the environment.

The rapid development of sequencing technologies in the last decades, such as Next-Generation Sequencing (NGS) and Third-Generation Sequencing (TGS), has significantly enhanced the scope of metagenomic studies. These technologies facilitate the rapid detection of pathogenic microorganisms, and with the aid of novel algorithms, improve the accuracy of taxonomic profiling and gene prediction [3]. Functional metagenomics, a key aspect of this field, enables the discovery of novel bioactive compounds and genes by screening microbial genes and metabolites extracted from environmental samples [4].

In this project, we implemented Nextflow, a workflow management tool, to streamline the execution of our metagenomics pipeline. Nextflow enables the integration of different analysis tools and manages dependencies, making it possible to perform complex computational workflows reliably and efficiently. To further enhance reproducibility and portability across different computing environments, we utilized Docker and Singularity. These container technologies package software and its dependencies into standardized units, ensuring that our pipeline functions identically regardless of the underlying infrastructure. This approach not only simplifies the deployment of our tools but also significantly reduces conflicts between software versions and dependencies, thus facilitating a more robust and scalable research framework.

This report details the development and application of a custom bioinformatic pipeline designed to analyze Whole Genome Sequencing data from metagenomic samples. Utilizing Nextflow for workflow management and Docker/Singularity for software consistency, our pipeline incorporates the principles of reproducibility and scalability essential for bioinformatics in metagenomics.

2. METHODS

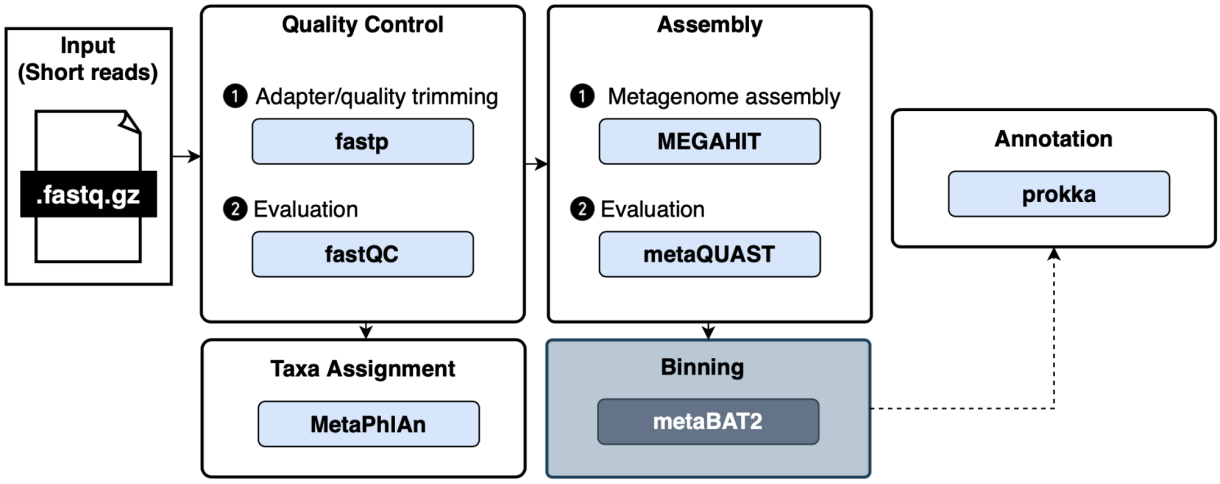

Figure 1. Pipeline overview

2.1 Sample preparation and quality control

In this pipeline, Fastp is utilized for preprocessing high-throughput sequencing raw data, which is important for the accuracy of subsequent analyses. Fastp filters out low-quality bases, cuts adapters, corrects mismatched base pairs, and provides a visualization of quality control and filtering results on a single HTML page. This step is essential in metagenomics WGS analysis as it ensures the reliability of data by removing artifacts.

Following preprocessing, FastQC performs quality control checks on the data. FastQC provides a quick overview of the data quality, highlights areas that may pose problems, and generates summary graphs and tables for a rapid assessment. The tool exports results to an HTML-based permanent report, allowing for easy verification and documentation of data quality, which is fundamental for robust metagenomic analysis.

2.2 Taxonomy assignment

Taxonomic composition of the microbial communities was determined using Metaphlan2, which profiles metagenomic shotgun sequencing data at the species level. It provides unambiguous taxonomic assignments and accurately estimates organism relative abundance, which is relevant for understanding microbial diversity and function. Moreover, the tool offers species-level genome bins (SGB) resolution, which enhances strain identification and tracking, crucial for metagenomic strain-level population genomics.

2.3 Assembly

MEGAHIT, was employed to assemble high-quality reads into contigs, minimizing computational resources while maximizing assembly quality. This tool produces high-quality assemblies with longer contig N50 and average contig length. These features are important for reconstructing genomes from complex microbial communities, providing a comprehensive view of genetic content and potential functionalities within the sample. The completeness and accuracy of the assembled genomes were evaluated using metaQUAST, which assesses assembly quality by analyzing the number of large contigs, their total length, the largest contig, and the N50 metric. It also estimates the number of predicted genes, enabling an assessment of the assembly's completeness and the potential discovery rate of functional elements, which is essential for accurate downstream analysis.

2.4 Binning and annotation

Binning in metagenomics involves grouping reads or contigs and assigning them to individual genomes. Metabat2 uses sequence composition, coverage, and abundance data, including tetranucleotide frequencies (TNFs) and a clustering algorithm, to effectively group contigs. Its statistical models and iterative clustering steps refine bin assignments, ensuring accurate identification of individual microbial genomes within complex samples.

Prokka rapidly annotates genomes, identifying and labeling features such as protein-coding genes, non-coding RNAs, tRNAs, and rRNAs. Prokka's speed, due to parallel processing techniques, and its ability to produce standards-compliant output files, make it particularly suitable for high-throughput metagenomic analysis, ensuring that functional insights are both comprehensive and promptly available.

3. OVERVIEW OF RESULTS

This pipeline is designed to be applicable to any metagenomics sequencing data. The github guides the user to process the raw sequencing data through four simple steps: cloning repository, installing nextflow, installing singularity/docker, and creating the folder of the input data.

The results of this pipeline, as illustrated in *Figure 2*, include detailed outputs from each stage such as trimmed reads, assembled genomes, annotated genomic features, and taxonomic profiles. Utilizing Nextflow, this pipeline benefits from robust data handling, scalable workflow management, and reproducible analyses across different computational environments.

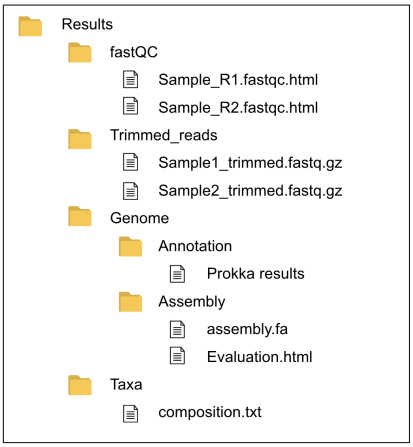

Figure 2. Pipeline results per sample processed

4. DISCUSSION

We replicated the described pipeline in terms of known species quantification, and metagenome biodiversity analysis. The taxonomic composition was quite similar, however species abundance proportions per sample were visibly different between the two pipelines (**Figure 3**). Further, two samples "failed" taxonomic analysis using MetaPhlAn due to the availability of many very short contigs (**Figure 3-** L7/S7). The beta biodiversity analysis was done using the R vegan package. Our findings confirmed the results from the study (**Figure 4, 5**). A PERMANOVA of the beta biodiversity was significantly different between the four different flight groups on the International Space Station (ISS), while comparing samples from the different locations showed no significant biodiversity.

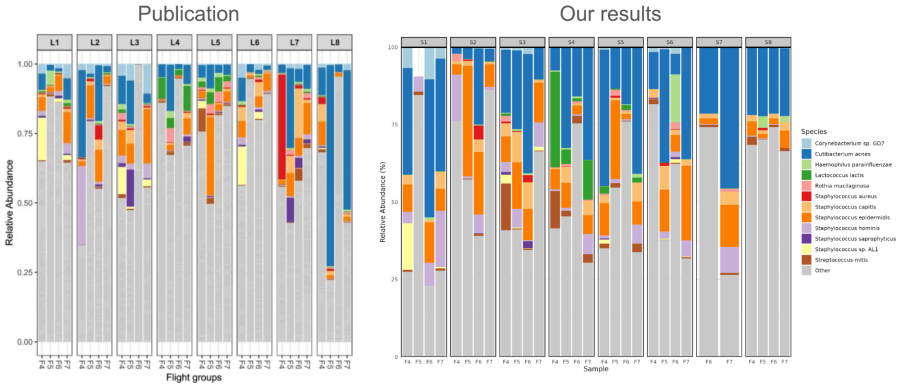

Figure 3. Comparison of taxonomic assignments from published data and pipeline results

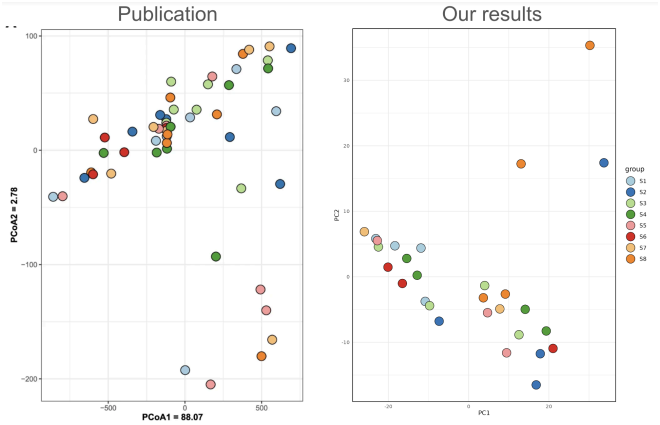

Figure 4. Comparison of beta diversity between locations

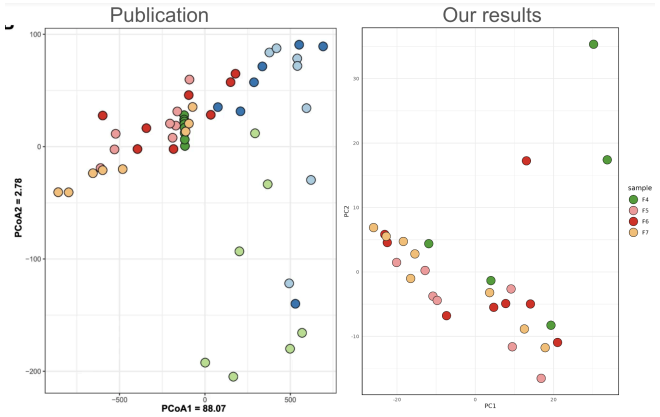

Figure 5. Comparison of beta diversity between flight groups

5. REFERENCES

[1] Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews, 68(4), 669-685.

[2] Kellenberger, E. (2001). Exploring the unknown. EMBO reports.

[3] Fadiji, A. E., & Babalola, O. O. (2020). Metagenomics methods for the study of plant-associated microbial communities: a review. Journal of microbiological methods, 170, 105860.

[4] Chen, G., Bai, R., Zhang, Y., Zhao, B., & Xiao, Y. (2022). Application of metagenomics to biological wastewater treatment. Science of The Total Environment, 807, 150737.