A Snakemake WGS pipeline for Environmental Sciences

Francisco Campuzano Jiménez[1], *Thanassis Zervas*[1]

Abstract

Whole genome sequencing (WGS) of bacterial isolates is a widely-used technique in research.

At the High-Throughput Sequencing Center of Environmental Sciences (Aarhus University), we developed a Snakemake pipeline for our standard workflow for investigating plant and Greenland ice sheet microbiome.

Our pipeline is designed to be portable, reproducible, and efficient, utilizing parallel computing and automatically generating reports.

Background

WGS of culturable microorganisms from environmental samples is a powerful tool for studying biodiversity. By analyzing the genomes of multiple isolates, researchers gain insights into genetic diversity and ecological roles.

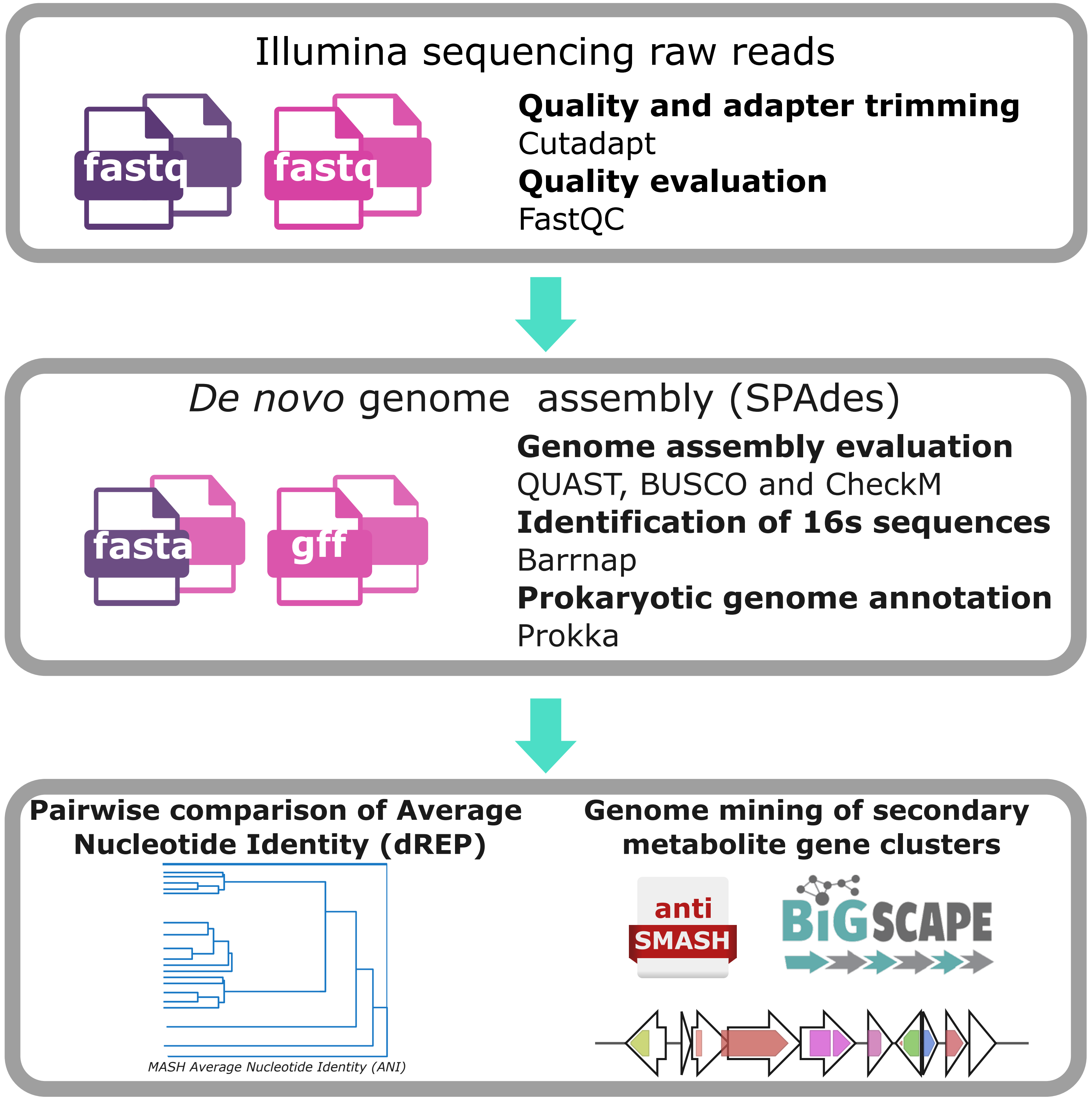
A standard workflow involves several steps: read trimming, *de novo* assembly, evaluation, annotation, and identifying the closest relative by blasting the predicted 16S rRNA sequence.

We also compare each genome based on Average Nucleotide Identity (ANI) and perform genome mining to identify natural product biosynthetic pathways.

Utilizing Open-Source workflow engines like Snakemake or NextFlow facilitates reproducibility, optimizes resource utilization, and speeds up large-scale bioinformatics analysis.

Methods

All the code for the pipeline is publicly available on GitHub.



Illumina sequencing raw reads

Quality and adapter trimming
Cutadapt
Quality evaluation
FastQC

De novo genome assembly (SPAdes)

Genome assembly evaluation
QUAST, BUSCO and CheckM
Identification of 16s sequences
Barrnap
Prokaryotic genome annotation
Prokka

Pairwise comparison of Average Nucleotide Identity (dREP)

Genome mining of secondary metabolite gene clusters

Results

The implemented pipeline has a modular organization, and is easy to extend. We have followed the best practice recommendations and implemented GitHub Action workflows to ensure the quality of our code.

Each program runs in its dedicated environment or Singularity container (when a database is needed). Likewise, all code is under version control. Thus, we ensure the reproducibility and portability of the pipeline.

The pipeline scales easily to compute clusters. It can process 19 samples in less than two hours on an HPC workstation with 75 assigned threads. Genome mining, a separate and more computationally intensive task, takes approximately 13 hours to complete under the same conditions.

The pipeline has proven successful in several publications regarding plants and the Greenland ice sheet microbiome.

Conclusions

In conclusion, we have successfully developed a portable, reproducible, and efficient pipeline for whole genome sequencing of prokaryotes by using Snakemake.

The pipeline has been employed in multiple ecology research projects and might be a valuable resource for the scientific community.

References

Campuzano, Francisco. (2023). AU-ENVS-BioinformaticsIlluminaSnakemake:V1.1. Zenodo.https://doi.org/10.5281/zenodo.7648481

Affiliations

[1]Department of Environmental Science, Aarhus University, Roskilde, Denmark

AARHUS UNIVERSITY