CENTRE OF BIOLOGICAL ENGINEERING

University of Minho
School of Engineering

Development of an automated pipeline for metagenomics and metatranscriptomics data analyses

Sequeira J. C.[1]; Rocha M.[1]; Alves M. M.[1]; Salvador A. F.[1]

[1]Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

Introduction

Metagenomics (MG) and Metatranscriptomics (MT) are useful approaches to study complex microbial communities in their natural environment, without the need for cultivation. However, MG and MT data analysis and interpretation is still challenging. None of the existing pipelines (e.g., MG-RAST[1], IMP[2], FMAP[3] and SAMSA2[4]) perform a complete and integrated MG/MT data analysis including, raw files preprocessing, reconstruction of metagenomes, annotation and differential gene expression. Therefore, an automated pipeline including all these functionalities is lacking.

Here we present the Meta-Omics Software for Community Analysis (MOSCA), a new pipeline that integrates major steps of MG and MT analysis, including preprocessing, assembly, annotation, differential gene expression and multi-sample comparison, with emphasis on automation and independence from web access.

Methods

Development and description of the pipeline

- MOSCA was developed in **Python 3** for **Unix systems**
- Available at **GitHub**: https://github.com/iquasere/MOSCA

- Input files are FastQ reads obtained from MG and MT Illumina sequencing
- Preprocessing removes artificial and low quality sequences, and ribosomal RNA
- Quality trimming parameters are automatically determined using FastQC reports information
- Assembly can be performed either with MetaSPAdes or Megahit
- MT reads are aligned to MG contigs in the differential expression step
- Comparison between multiple samples is performed
- Quality control tools are integrated through the entire pipeline

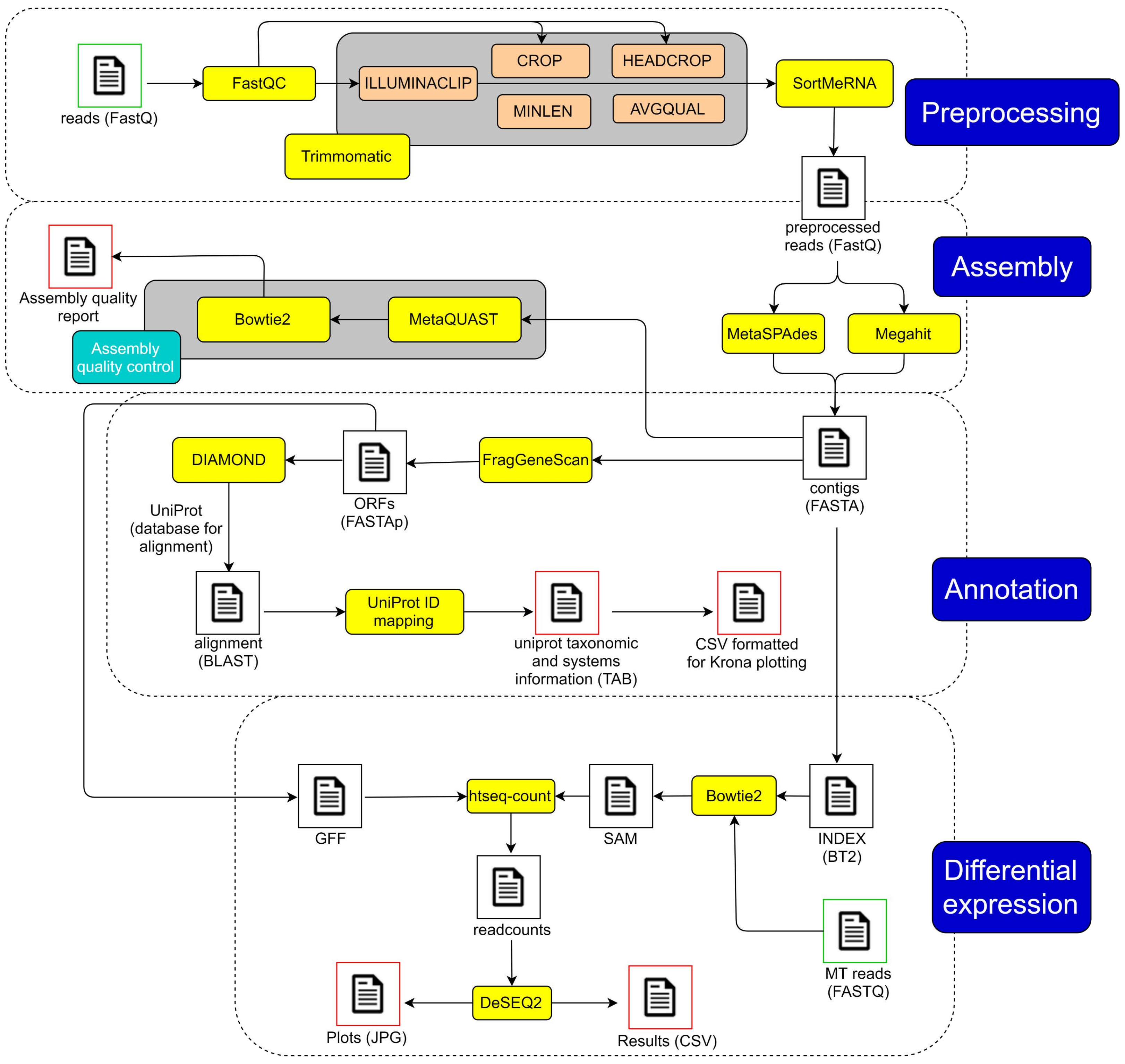

Fig. 1 – Workflow of MOSCA. Main steps (dark blue boxes), sub-steps (light blue boxes), tools integrated (yellow boxes), functionalities of tools (orange boxes), input files (green squares), intermediate files (black squares), and final output files (red squares).

Simulated datasets

Grinder was used for simulating metagenomic (MG) datasets in FastQ format, considering different relative taxonomic abundances, and **polyester** to simulate metatranscriptomics (MT) datasets in FASTA format considering differential gene expression for three different conditions: control, over- and underexpression.

Results

Taxonomic analysis of metagenomes and metatranscriptomes

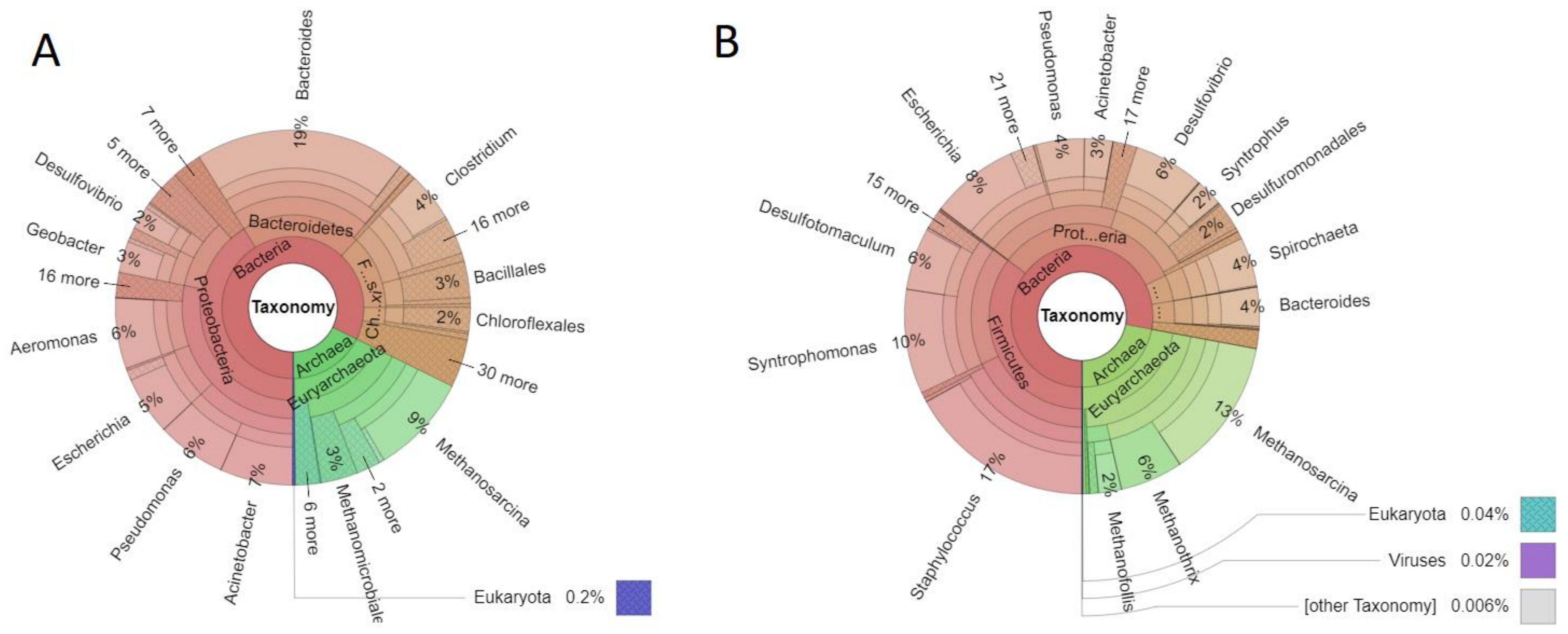

Fig. 2 – Krona plots showing the relative abundance of each genus in MG (A) and MT (B) samples.

- All microorganisms included in the simulated datasets could be identified
- More genes were assigned to the microorganisms set as more abundant in the simulated datasets

Functional analysis and differential gene expression

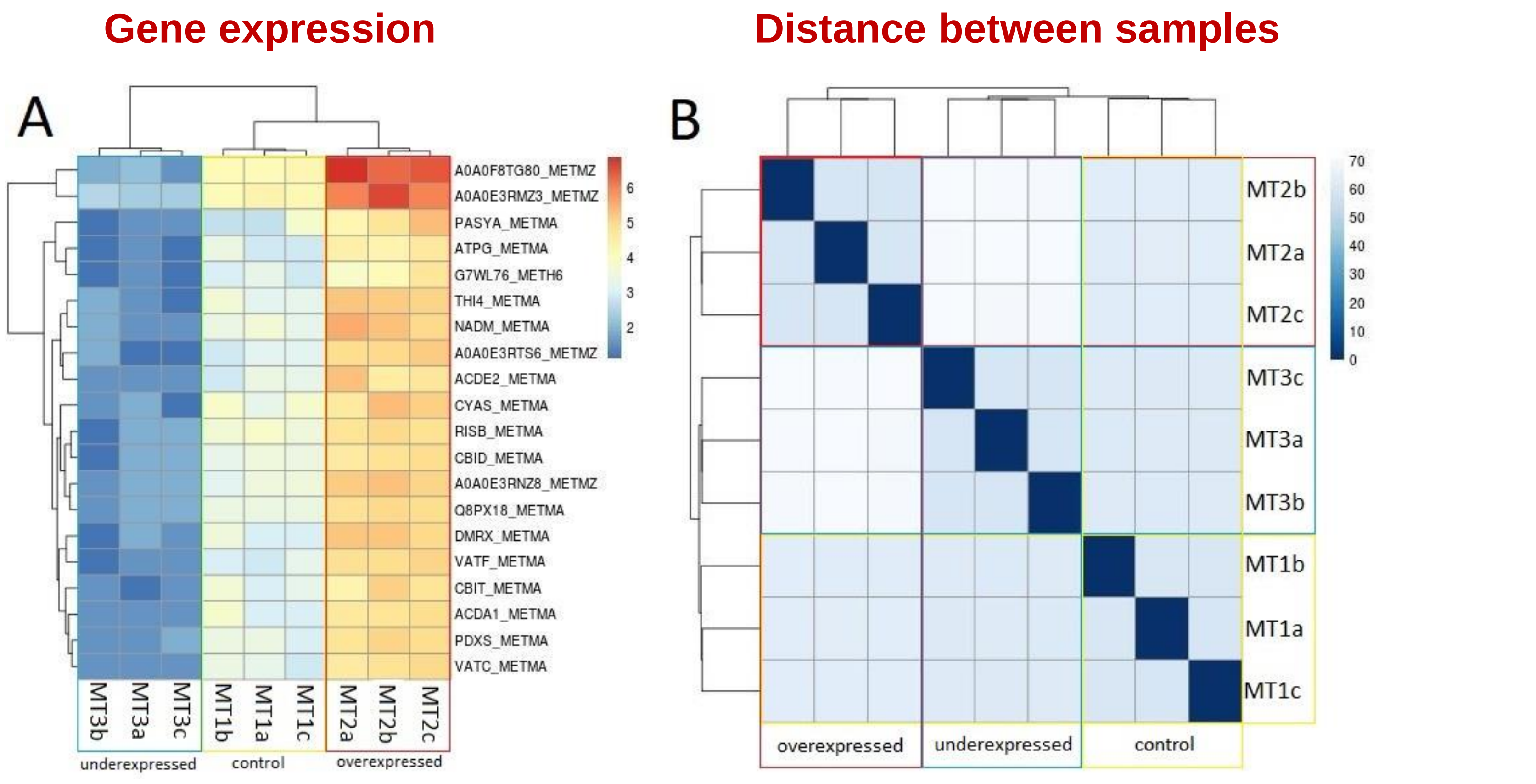

Fig. 3 – Visualization of the differential gene expression results obtained with the simulated datasets. The heatmaps represent the most expressed genes (A) and the distance between samples (B) for MT1 (control condition), MT2 (overexpressed condition) and MT3 (underexpressed condition). The letters a, b and c refer to replicates. Red color represents the most expressed genes and blue the least expressed genes (A); white color represents the highest distance between samples and blue the highest proximity between samples (B).

- The MT triplicates clustered together at the gene expression level (A) and at the sample level (B)
- MOSCA clearly distinguished the datasets corresponding to the three different conditions simulated (MT1, MT2, MT3)
- Most expressed genes were obtained for the microorganisms and pathways that were set as more abundant in the simulated MT datasets

Conclusions

- MOSCA was developed as a command-line pipeline that integrates all major steps of metagenomics and metatranscriptomics data analysis
- MOSCA performs an automated preprocessing, by adjusting quality trimming arguments based in the quality control output obtained with FastQC
- MOSCA is different from other existing pipelines because it includes the assembly of the reads, integrates MG and MT data, and performs multi-sample gene expression analysis.

References

[1] Wilke, et al (2015) Nucleic acids research 44(D1):D590{D594
[2] Narayanasamy, et al (2016) bioRxiv (7):039,263
[3] Kim, et al (2016) BMC bioinformatics 17(1):420
[4] Westreich, et al (2017) bioRxiv p 195826

Acknowledgements

This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte, and by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 323009.

Bioinformatics Open Days 2018