Standardized and Validated Workflow for Comparative Metatranscriptomics

Muhammad Zohaib Anwar[1], Anders Lanzen[2], Toke Bang-Andreasen[3], Carsten Suhr Jacobsen[1]

[1]Aarhus University, Denmark, [2]AZTI-Tecnalia, Herrera Kaia, Pasaia, Spain and IKERBASQUE, Basque Foundation for Science, Spain, [3]Copenhagen University, Denmark

mzanwar@envs.au.dk

@Xohaib_Anwar

INTRODUCTION

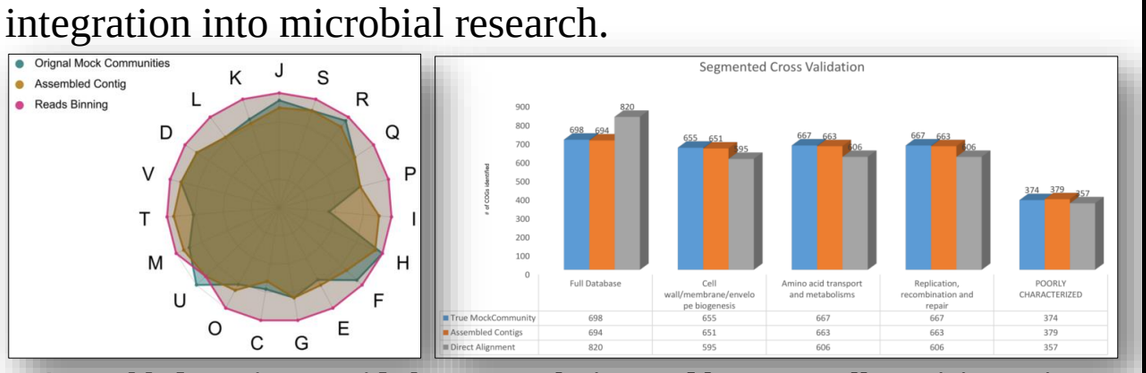
There has been a large emphasis on microbial responses, how microorganisms' response can be quantitatively analyzed remains debated. Metatranscriptomics is the study of rRNA and mRNA of a microbial community in an environment. It allows the simultaneous investigation of the gene expression (mRNA) and abundance (rRNA) of the active microorganisms. It holds great potential to uncover biological information that may be otherwise obscured by other genomic methodologies. It provides an accurate snapshot, at a given moment in time and under specific conditions, of the actual gene expression profile rather than its potential, as inferred from DNA-based shotgun metagenomic sequencing. As metatranscriptomes experiments are consistently increasing in size and number, automated, efficient, high-throughput analyses are essential to infer the biological meaning from these datasets.

OBJECTIVE

- Independent and direct benchmarking between two alternative approaches of metatranscriptomics analyses using simulated datasets and real metatranscriptomes.
- Evaluation of accuracy in precision and recall using the Md5nr and eggNOG hierarchical databases.
- Standardized and validated workflow.

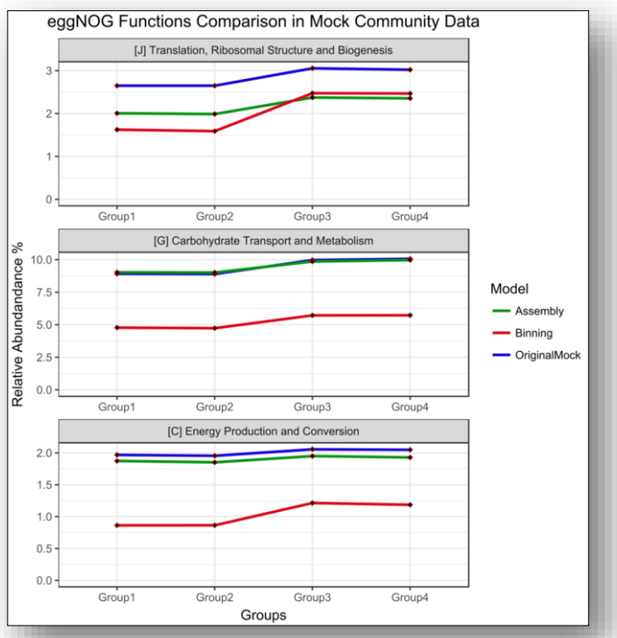
STANDARDIZATION & BENCHMARKING

At present, multiple methods may at times produce variable results, even if identical databases are used in the analysis. Thus, standardization of analysis is warranted to enable further dissemination of metatranscriptomics methods and their integration into microbial research.



Assembled contigs provide better resolution and better recall precision using Md5nr database and customized databases in synthetic mock communities

Using segmented cross validation, we also show that the assembly-based method is much more robust with customized databases or environment specific databases. At a lenient confidence threshold assembly-based identification resulted in a maximum FDR of 3.5% (<1% using an optimized threshold) whereas direct alignment of reads resulted in FDRs up to 15% (>3% even at very strict confidence).



Quantification of transcripts is more realistic and accurate using assembled contigs.

VALIDATED WORKFLOW

A few analytical steps are vital in this process and are, therefore, present uniformly in all metatranscriptomic analysis. The common overall goal of inferring the gene expression levels and variations in gene expression levels, from the raw sequenced mRNA reads. The standardized work flow we present here is based on the results and findings from comparison of approaches, however the work flow has multiple optional steps such as abundance based and non-coding RNA filtering which can be different in data sets from a different environment. The scripts are designed to cater more than one assembler output to enable diverse range of environments to be studied.

A complete standardized workflow to sort, pre-process, assemble, identify, and quantify metatranscriptomes

CONCLUSION

Our findings support the argument of assembling short reads into contigs before alignment to a reference database, since this provides higher resolution and minimizes False Discovery Rate. By virtue of the comparative analysis we also present an open source metatranscriptomic analysis workflow written in python.

This work is supported by a grant from the European Commission's Marie Sklowdowska Curie Actions program under project number 675546.