MG7: A SCALABLE FAST TOOL BASED ON CLOUD COMPUTING AND GRAPH DATABASES FOR MICROBIAL COMMUNITY PROFILING

M. Manrique[1], P. Pareja-Tobes[1], E. Kovach[1], A. Alekhin[1], E. Pareja[1], E. Pareja-Tobes[1], **R. Tobes**[1]

[1]Oh no sequences! Research group, Era7 Bioinformatics, Granada, Spain

Objectives
To provide a scalable fast tool for community profiling of 16S metagenomics data.

Methods

MG7 is an open source tool implemented in Java and Scala and based on cloud computing (Amazon Web Services) that uses the graph database Bio4j (www.bio4j.com) for retrieving taxonomy related information.

Results

We provide with an open-source scalable fast tool for community profiling based on the analysis of 16S metagenomics data.

MG7 is a tool that allows fast 16s metagenomics analysis. It benefits from the use of cloud computing so it is inherently scalable and gives the user the possibility of analysing multiple samples at the same time. MG7 performs the community profiling of a sample starting from raw Illumina reads in approximately 1 hour. MG7 approach carry out the analysis applying parallelization strategies and adjusting automatically the computation capacity to the computational tasks of each project. The taxonomic assignment can be done using a Best BLAST hit paradigm or a Lowest Common ancestor Paradigm. The user can choose between both assignment algorithms and setting the similarity parameters required for the assignment.

As an output, MG7 generates the frequencies of all the identified taxa in any of the samples in tab-separated values text files as well as in the standard BIOM format compliant with other metagenomics tools. MG7 provide direct assignment frequencies and cumulative frequencies based on the hierarchical structure of the taxonomy tree. It also provides with output files for heat-map representations.

Conclusions

MG7 is an open-source tool based on cloud computing and on the use of Bio4j graph database that allows to analyse 16S metagenomics data. MG7 is fast and scalable tool with computational capabilities for huge amount of data management allowing the characterization of the microbial diversity of multiple samples.