MG7: A fast horizontally scalable tool based on cloud computing and graph databases for microbial community profiling

Evdokim Kovach, Alexey Alekhin, Marina Manrique, Pablo Pareja-Tobes, Eduardo Pareja, Raquel Tobes and Eduardo Pareja-Tobes*

Oh no sequences! Research Group. Era7 bioinformatics

*eparejatobes@ohnosequences.com

Abstract. *Methods:* MG7 is an open source tool implemented in Java and Scala, based on cloud computing (Amazon Web Services). The graph data platform Bio4j (www.bio4j.com) is used for retrieving taxonomy related information, while Nispero (http://ohnosequences.com/nispero) is used for distributing and coordinating compute tasks.

Results: MG7 is an open-source, fast and horizontally scalable tool for community profiling based on the analysis of 16S metagenomics data. It is entirely cloud-based and specifically designed to take advantage of it: it performs the community profiling of a sample starting from raw Illumina reads in approximately 1 hour, needing approximately the same time for doing the same on hundreds of samples, adjusting automatically the computation capacity to the resources needed in 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 value text files as well as in the standard BIOM format compliant with other metagenomics tools. This output includes direct assignment frequencies and cumulative frequencies based on the hierarchical structure of the taxonomy tree. It also provides with output files suitable for generating heat-map representations.

MG7 is an open-source tool available under the AGPLv3 license

This project is funded in part by the ITN FP7 project INTERCROSSING (Grant 289974) and the Spanish CDTI (Centro para el Desarrollo Tecnológico Industrial) grant NEXTMICRO, ref. IDI-20120242.

Keywords: Metagenomics; 16S; microbiome; microbial diversity; cloud computing; high performance; bio4j; distributed systems.