

# 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](mailto:eparejatobes@ohnosequences.com)*

April 8, 2014

MG7 is cloud tool for community profiling based on the analysis of 16S metagenomics data. More specifically it performs assignments to the taxonomy tree based on mapping reads from samples to a 16S database.

Metapasta is a version of MG7 written in Scala programming language.

- Metagenomic samples with reads in FASTQ format
- 16S database (by default Metapasta uses NCBI 16S with some filtering)
- Taxonomy database (we are NCBI taxonomy database that integrated to Bio4j).

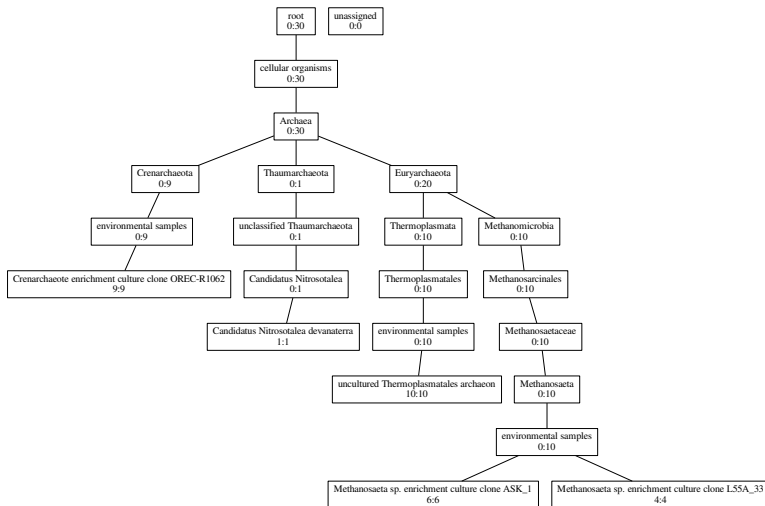
As results metapasta produce

- For every tax id and sample:
  - number of reads from the sample that were assigned to this tax id
  - number of reads from the sample that were assigned to all successors of this tax id
- Aggregated data by samples.

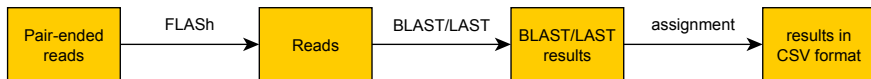
# Results. Tables

taxId	name	rank	supermock2.count	supermock2. acc	total.count	total.acc
total			20	234	20	234
159447	uncultured Corynebacterium sp.	species	6	6	6	6
404941	Mycobacterium salmoniphilum	species	3	3	3	3
37637	Corynebacterium pseudodiphtheriticum	species	2	2	2	2
1221985	Mycobacterium sp. ITM090653	species	2	2	2	2
319705	Mycobacterium abscessus subsp. bolletii	subspecies	2	2	2	2
1079047	Mycobacterium sp. R5	species	1	1	1	1
43769	Corynebacterium propinquum	species	1	1	1	1
592914	Corynebacterium sp. M71_S35	species	1	1	1	1
948102	Mycobacterium franklinii	species	1	1	1	1
1774	Mycobacterium chelonae	species	1	1	1	1
2	Bacteria	superkingdom	0	20	0	20
2037	Actinomycetales	order	0	20	0	20
131567	cellular organisms	no rank	0	20	0	20
1	root	no rank	0	20	0	20
85007	Corynebacterineae	suborder	0	20	0	20
1760	Actinobacteria	class	0	20	0	20
201174	Actinobacteria	phylum	0	20	0	20

# Results. Trees



# Pipeline



- FLASH merging paired-end reads into big reads
- Mapping reads against the 16S database (with BLAST or LAST)
- Assignment to the taxonomy tree using Bio4j.

# Mapping problem

Mapping NGS reads against the 16S database requires really a lot of computational resources.

For example even on fast computers with SSD and big size of RAM mapping of one read with BLAST takes more than 0.2 seconds.

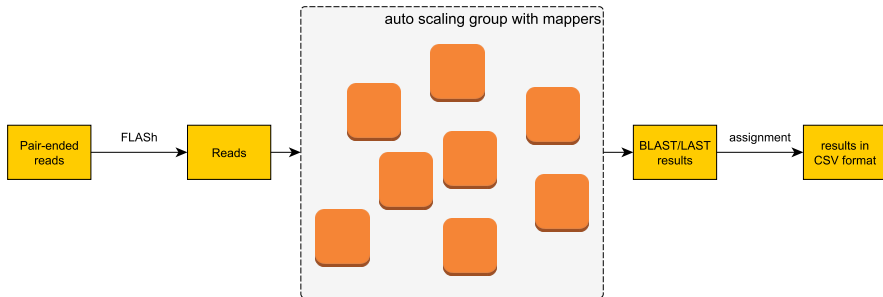
$$1000000 \times 0.2s \approx 56h$$

The mapping time can be improved by using more efficient mapping tool (by default Metapasta uses LAST that is 100 times faster than BLAST in this case).



# Cloud solution

Metapasta uses AWS (Amazon Web Services) for performing all computations (EC2 instances):



Besides computations Metapasta uses AWS for all data management:

- input data – samples are stored in S3
- output assignment tables and trees in PDF
- Metapasta can upload all reads (with assignment) to DynamoDB table.



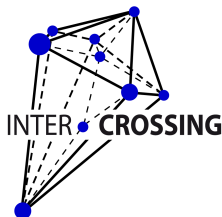
- Scala library for building distributed using AWS
- Easy to use (only AWS account is needed)
- Designed to provide maximal level of scalability and availability (in case of Metapasta it means that it can process **with a good performance** as many samples as needed, and never will lose any of reads).

$$Reads \otimes Reads \xrightarrow{merge} Reads \xrightarrow{BLAST} AssignTable \otimes Reads$$

- They are powerful enough to build complex system
- It is easy to work with them abstractly (for example to create graphical workflow editor that produces distributed systems).

# INTERCROSSING

This project is funded in part by the ITN FP7 project INTERCROSSING (Grant 289974).



Thank you for your attention!