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
Granada, Spain

April 8, 2014

What is metagenomics?

Metagenomics is the the study about collecting of genetic material from mixed community of organisms (usually bacteria):

- Soil samples
- Marine samples
- Clinical samples.

The problem of metagenomics

The typical problem of metagenomics is to obtain information about species composition in the sample:

- Which species are presented in the sample?
- How many different species are presented in the sample?
- How many species from the given genus are presented in the sample?

16S databases

16S rRNA gene is widely used to identify bacteria. There are several publicly available databases:

- NCBI 16S
- Greengenes
- SILVA.

What is MG7?

MG7 is a cloud tool that performs assignment to the taxonomy tree based on mapping of reads from samples against 16S database.

Metapasta is a version of MG7 written in Scala that designed to process a really big amounts of metagenomics data.

Input

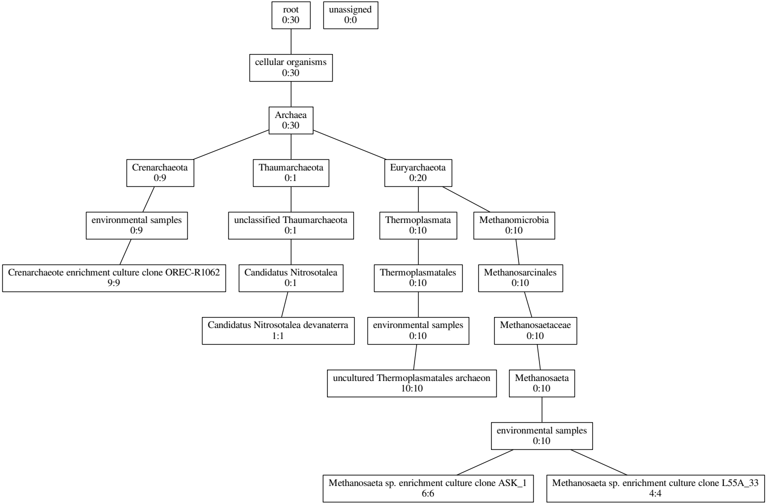
- Sequenced metagenomics samples – reads in FASTQ format
- Taxonomy database (we are using NCBI taxonomy database that is integrated to Bio4j)
- 16S database (by default Metapasta uses NCBI 16S with some filtering).

Output

As result Metapasta produces

- For every taxonomy 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. Trees



root
0:30

unassigned
0:0

cellular organisms
0:30

Archaea
0:30

Crenarchaeota
0:9

Thaumarchaeota
0:1

Euryarchaeota
0:20

environmental samples
0:9

unclassified Thaumarchaeota
0:1

Thermoplasmata
0:10

Methanomicrobia
0:10

Crenarchaeote enrichment culture clone OREC-R1062
9:9

Candidatus Nitrosotalea
0:1

Thermoplasmatales
0:10

Methanosarcinales
0:10

Candidatus Nitrosotalea devanaterra
1:1

environmental samples
0:10

Methanosaetaceae
0:10

uncultured Thermoplasmatales archaeon
10:10

Methanosaeta
0:10

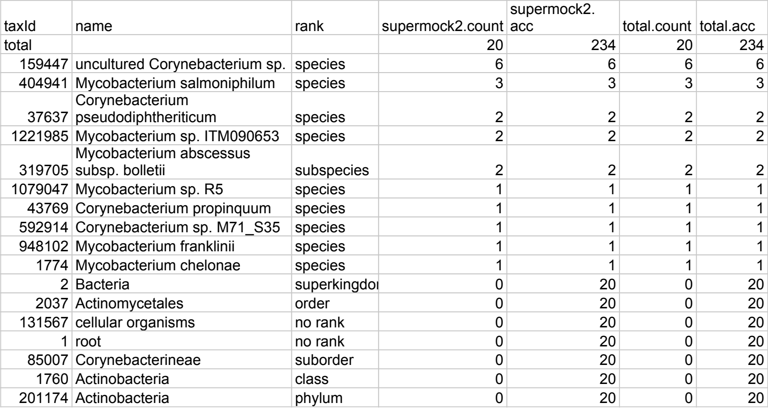
environmental samples
0:10

Methanosaeta sp. enrichment culture clone ASK_1
6:6

Methanosaeta sp. enrichment culture clone L55A_33
4:4

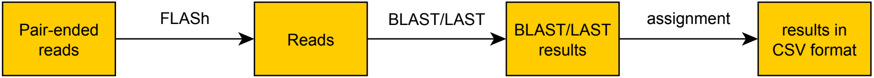
Era7 bioinformatics

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	superkingdoi	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

Metapasta pipeline



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

Mapping problem

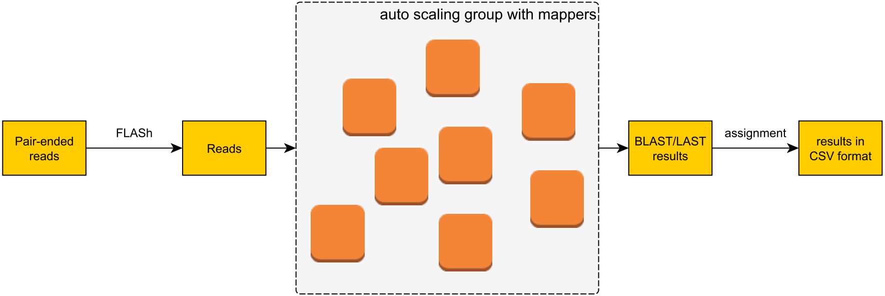
Mapping NGS reads against the 16S database requires really a lot of computational resources.
For example even on fast computers with SSD and big size of RAM mapping of one read with BLAST takes more than 0.2 seconds.

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

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

Cloud solution

Metapasta uses AWS (Amazon Web Services) to perform all computations (EC2 instances):



Cloud solution. Data management

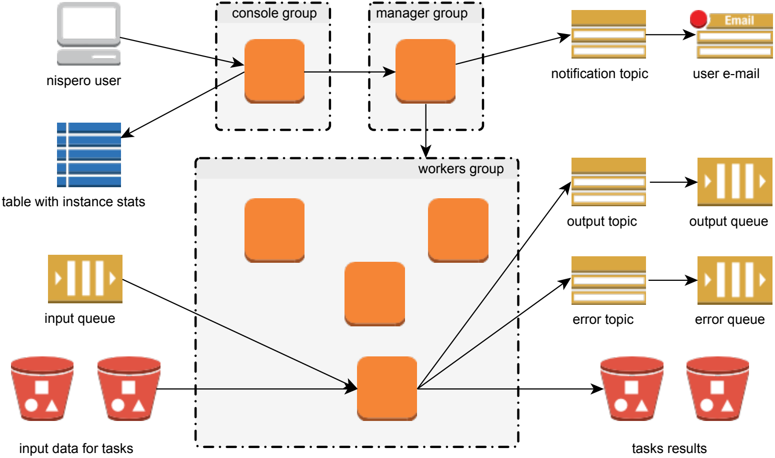
Besides computations Metapasta uses AWS for all data management:

- Reads from samples (S3)
- Taxonomy assignment tables and trees in PDF (S3)
- Metapasta can upload all reads (with assignments) to DynamoDB table.

Nispero

- Scala library for building distributed using AWS
- Easy to use (only AWS account is needed)
- Scalable and robust.

Nispero. Architecture



Monoids

We are using idempotent commutative monoids to describe distributed systems in Nispero.

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

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

Nispero

Do not lose opportunity to try this amazing Spanish fruit!

Thank you for your attention!