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

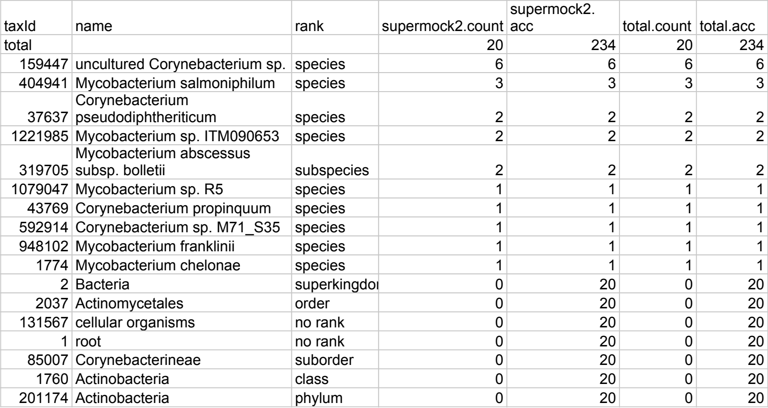
April 8, 2014

What is Metapasta?

put something about 16S
thing that produce taxonomical assigment to taxonomy tree tree:

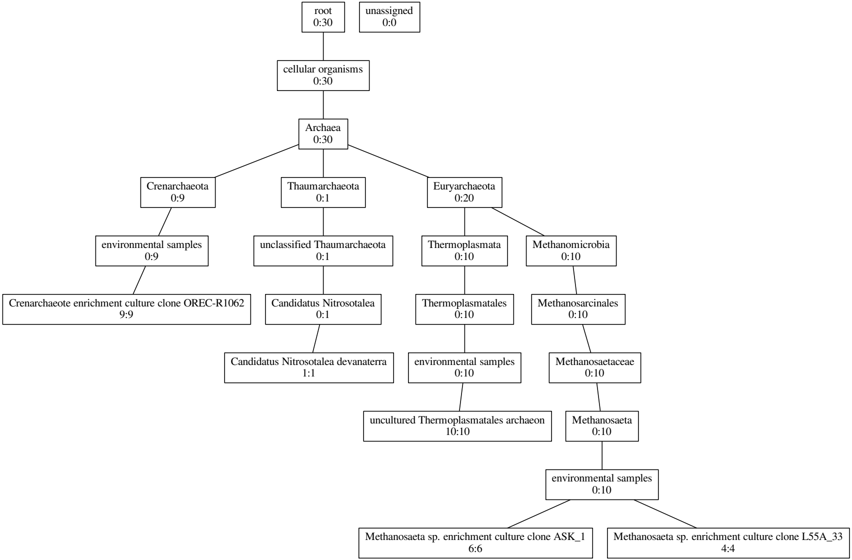
- Best BLAST/LAST hit
- a Lowest Common ancestor.

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	superkingdor	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



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

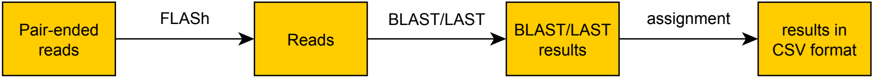
Methanosaeta 0:10

environmental samples 0:10

Methanosaeta sp. enrichment culture clone ASK_1 6:6

Methanosaeta sp. enrichment culture clone L55A_33 4:4

Pipeline



- FLASh merging paired-end reads into big reads
- BLAST/LAST mapping to 16S database
- Assignmnet to the taxonomy tree using Bio4j

Mapping problem

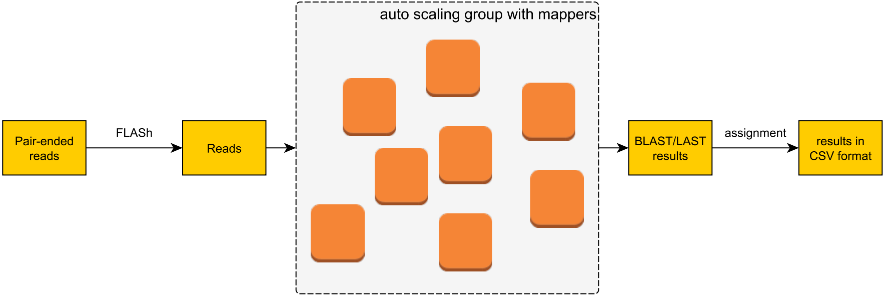
Mapping NGS reads to 16S takes requires a lot of computational resources. For example even on fast computers with SSD and size of RAM mapping of one read with BLAST takes more than 0.2 seconds.

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

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

Cloud solution

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



Cloud solution. Data management

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

Nispero

- Scala library for building distributed systems
- Dedicated to provide maximal level of scalability and availability

Nispero

- Amazon auto scaling groups
- SQS queues

Monoids

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

Bio4j

NCBI taxonomy

INTERCROSSING

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

Thank you for your attention!