Metapasta: scalable tool 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

September 22, 2014

Era7 bioinformatics

What is Metapasta?

Metapasta is a tool for microbial community profiling.

It's designed to answer questions like:

- 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 metagenomics

To identify presented organism in the sample gene markers are used.

The most widely used gene marker for bacteria is 16S rRNA gene. There are several publicly available databases with 16S sequences:

- NCBI 16S
- Greengenes
- SILVA.

16S metagenomics II

16S databases give us a pretty straightforward approach to analyse the species composition:

- 16S rRNA amplicon sequencing
- mapping reads against an 16S database.

16S metagenomics challenges

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 0.25 seconds.

$$1\ 000\ 000 \text{ reads} \times 0.25 \text{ seconds each} \approx 70 \text{ hours.}$$

Metapasta: cloud based and scalable

To solve these challenges we developed Metapasta:

- cloud based
- horizontally scalable, makes possible to use as many instances as possible.

Cloud solution

Metapasta uses AWS (Amazon Web Services) for

- all computations (EC2 instances)
- storing samples and results (S3 buckets)
- communications between computational nodes (SQS queues).

Era7 bioinformatics

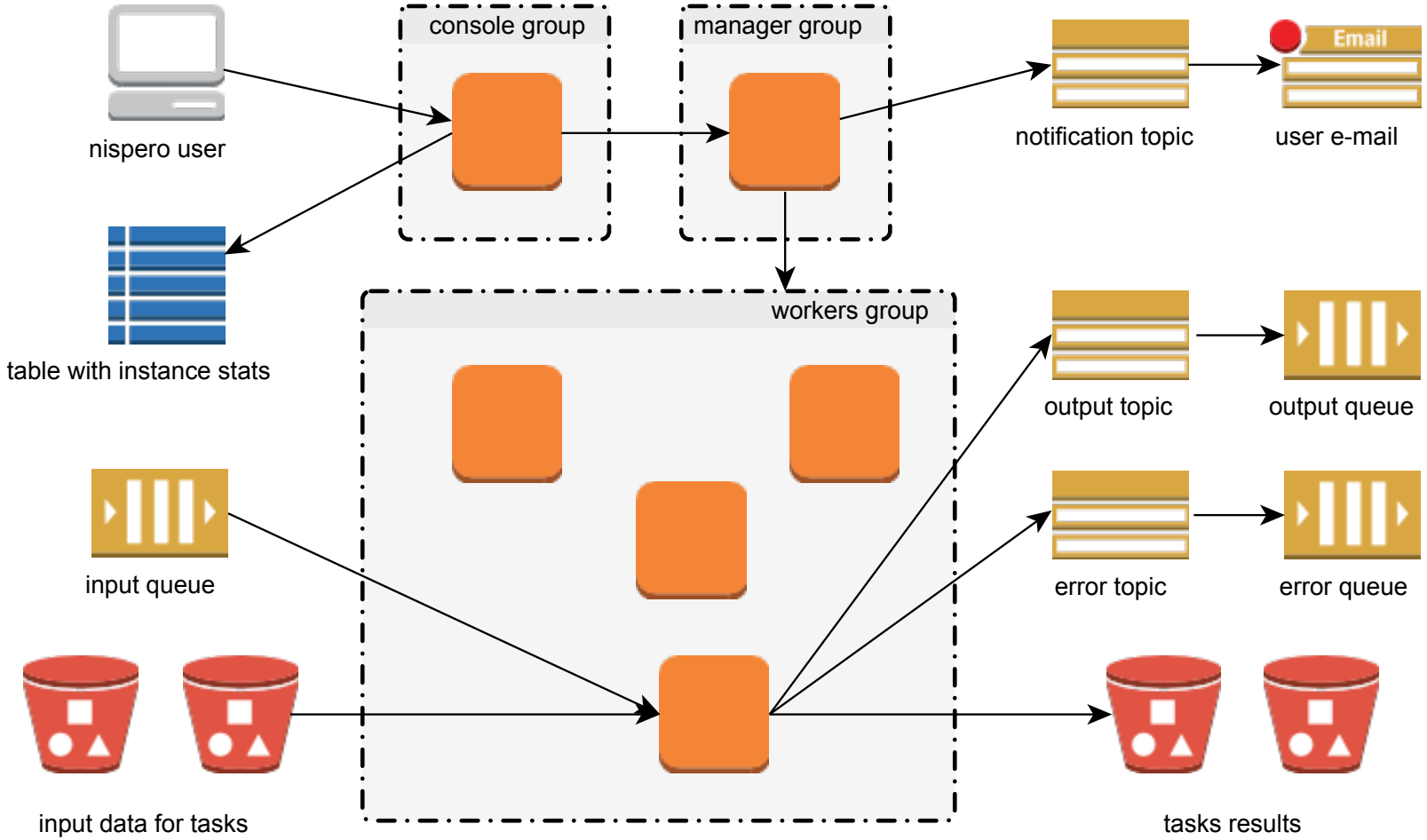
Compota

Compota is a Scala library for performing computations in Amazon cloud:

- easy to use (only AWS account is needed)
- designed to be *fault tolerant* – failures in particular nodes never affect the result
- *scalable* – unlimited number of computational nodes can be used with minimal communication overhead.

Era7 bioinformatics

Compota. Architecture



Era7 bioinformatics

Compota. Monoids

We are using monoid morphisms to describe computations in Compota:

- express distributiveness
  $f(a) = merge(f(split(a))) = merge(f(a_1), f(a_2), \ldots)$
- can be composed in pipelines:

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

Era7 bioinformatics

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.

Era7 bioinformatics

Assignment

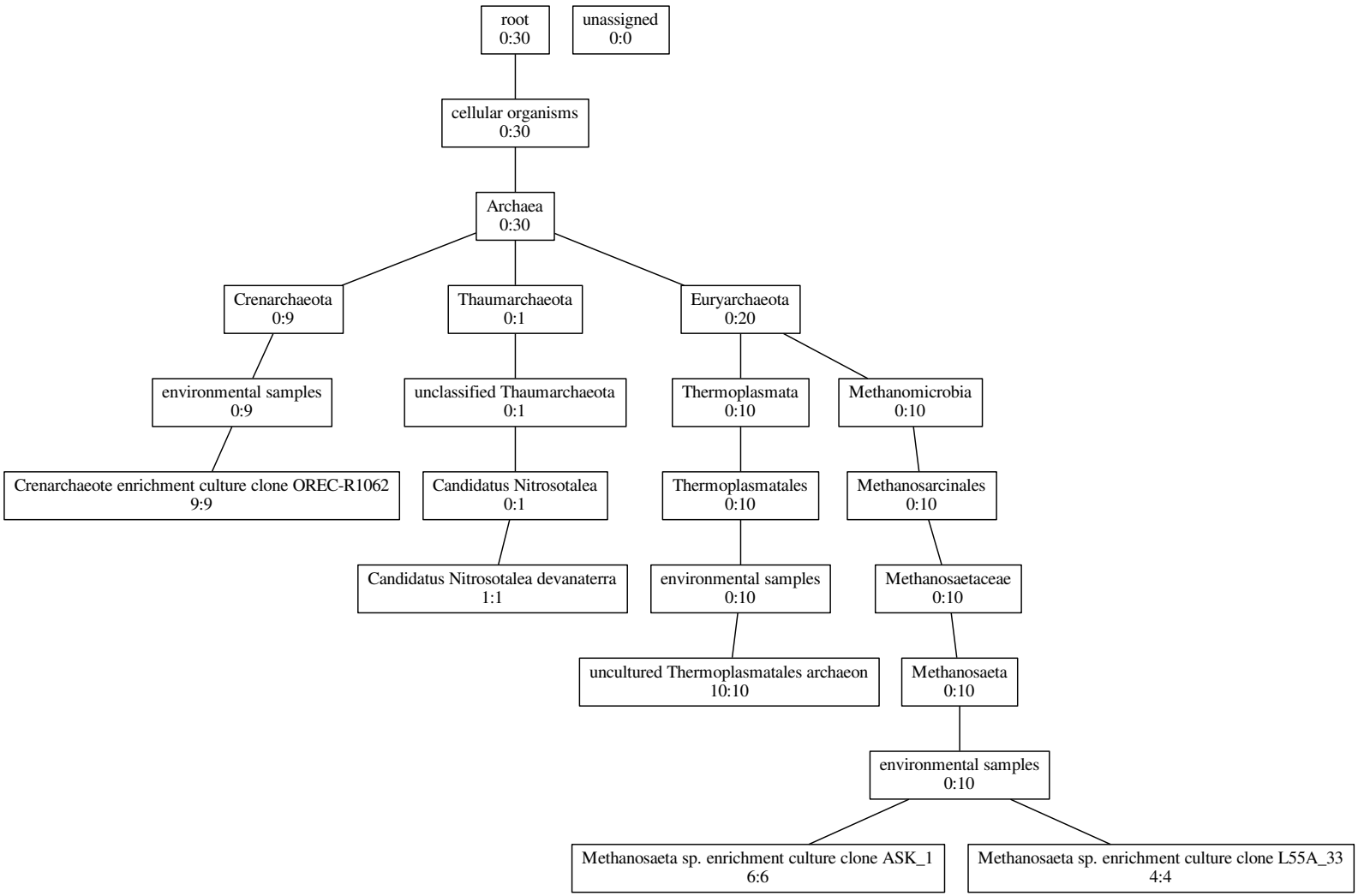
Results of *mapping* step:

- $read_1 \mapsto \{(taxon_1^1, score_1^1), (taxon_2^1, score_2^1)\}$
- $read_2 \mapsto \{(taxon_1^2, score_1^2), (taxon_2^2, score_2^2), (taxon_3^1, score_3^1)\}$
- $read_3 \mapsto \{\}$
- $read_4 \mapsto \{taxon_1^4\}$

We assign reads to taxonomy tree using to independent algorithms:

- BBH – Best blast hit
- LCA – Lowest common ancestor.

Results. Trees



Era7 bioinformatics

Results. CSV 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	superkingdo	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

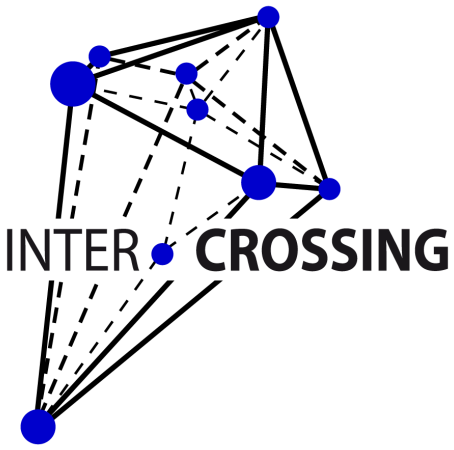
Availability

Compota is an open-source project released under AGPLv3 license. The source code is available at `github.com/ohnosequences/metapasta`.

Era7 bioinformatics

INTERCROSSING

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



Era bioinformatics

Thank you for your attention!

Era7 bioinformatics