Microbial communities through the lens of high throughput sequencing, data integration and metabolic networks analysis

developing computational approaches to better understand microbial assemblages

Haris Zafeiropoulos
PhD candidate

Microbial ecology & biogeochemical cycles

a corner-stone for life on earth

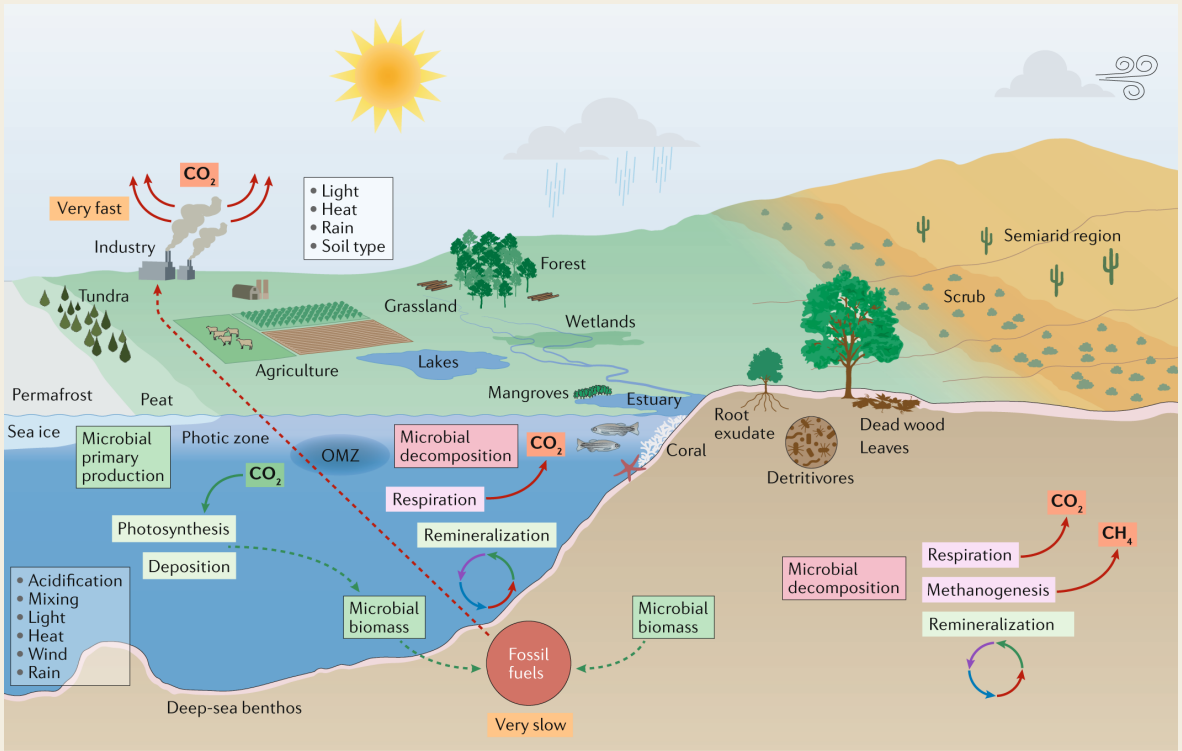


Figure from: Nature Reviews Microbiology 17.9 (2019): 569-586.

- composition
- functions
- interactions

→ power biogeo-chemical cycling

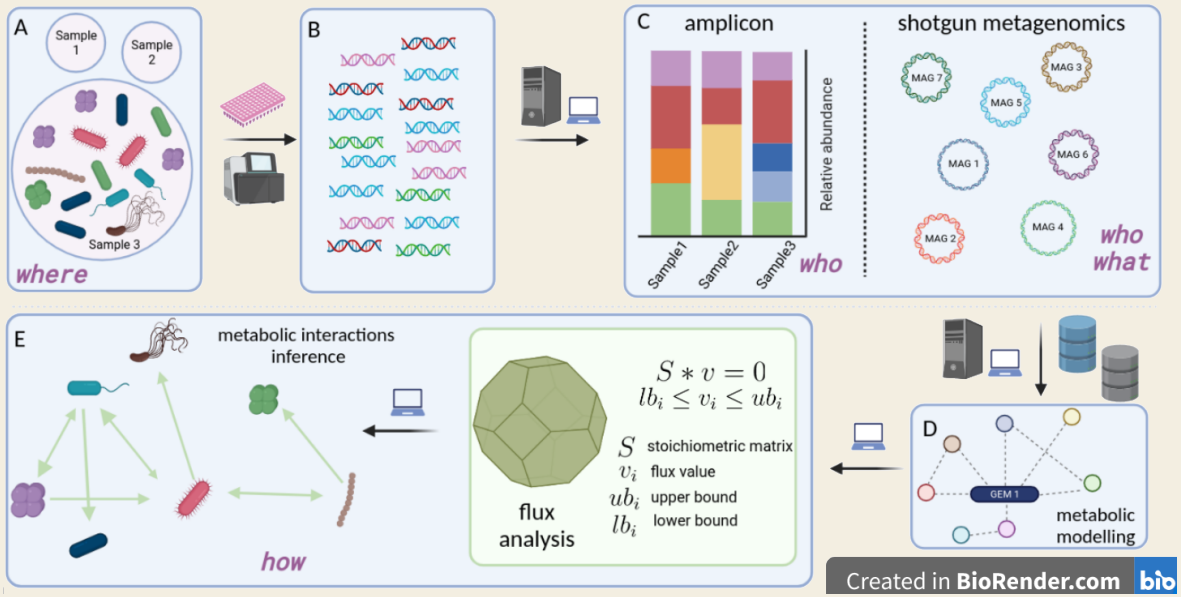
Main questions regarding a microbial community

for a deeper understanding of such assemblages

Community structure ***who***	Functional potential ***what***	Microbial interactions ***why / how***
eveyrone is everywhere	*zero-sum game*	*the entagnled bank*

Reverse ecology

transforming ecology into a high-throughput field



community ecology studies with no *a priori* assumptions about the organisms under consideration by exploiting advances in systems biology and genomic metabolic modeling

High Throughput technologies

a new era bringing its own challenges

- biology-oriented issues
- techonology-oriented issues
- computing requirements
- multiple channels of information

Aims and objectives

- to enhance the analysis of microbiome data by building algorithms and software that address limitations and on-going computational challenges

- to exploit state-of-the-art methods to identify taxa and functions that play a key part in microbial community assemblages in hypersaline sediments

1. **Chapter 1**: Introduction

2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment

    2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

    2.2 `darn`: known unknowns in COI amplicon data

3. PREG0: a knowledge-base for organisms - environments - processes associations

eDNA metabarcoding

for biodiversity assessment

Marker genes

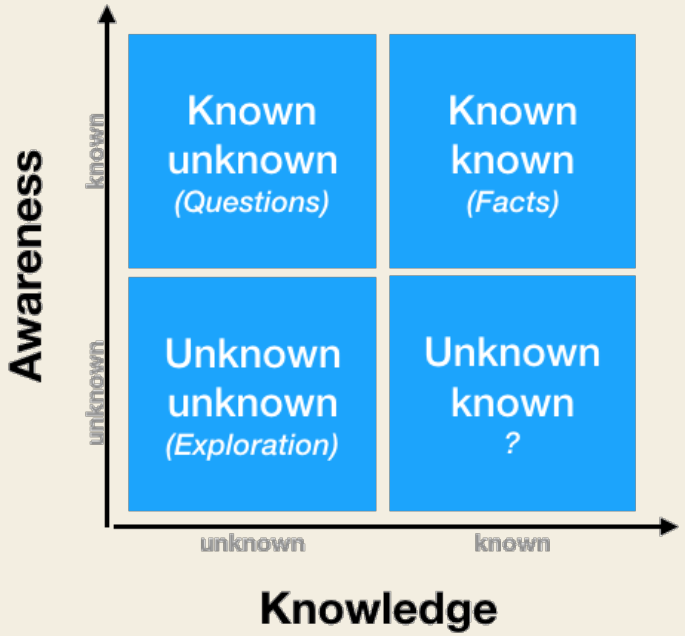
1. **16S rRNA:** Bacteria, Archaea
2. **12S rRNA:** Vertebrates
3. **18S rRNA:** Small eukaryotes, Metazoa
4. **ITS:** Fungi
5. **COI:** Eukaryotes
6. ***rbcl*:** Plants
7. ***dsrb*:** Bacteria, Archaea
8. ...

Methodology

- Sampling
- Extraction
- Bioinformatics
- Biodiversity analysis

Bioinformatics challenges

for the analysis and the interpretation of amplicon data



Chapter 2.1: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

Aim of the study and contribution

To build an open source pipeline that bundles state-of-the-art bioinformatics tools for all necessary steps of amplicon analysis and aims to address:

- one-stop-shop for several marker genes & approaches
- easey-to-set & easy-to-use
- scalable
- flexible

Methods / Implementation

PEMA coding insights

```
for(int i : range(1,
    in  := "in_$i.tx
    sys date > $in

    out := "out_$i.t
    task( out <- in
        sys echo Tas
    }
}
```

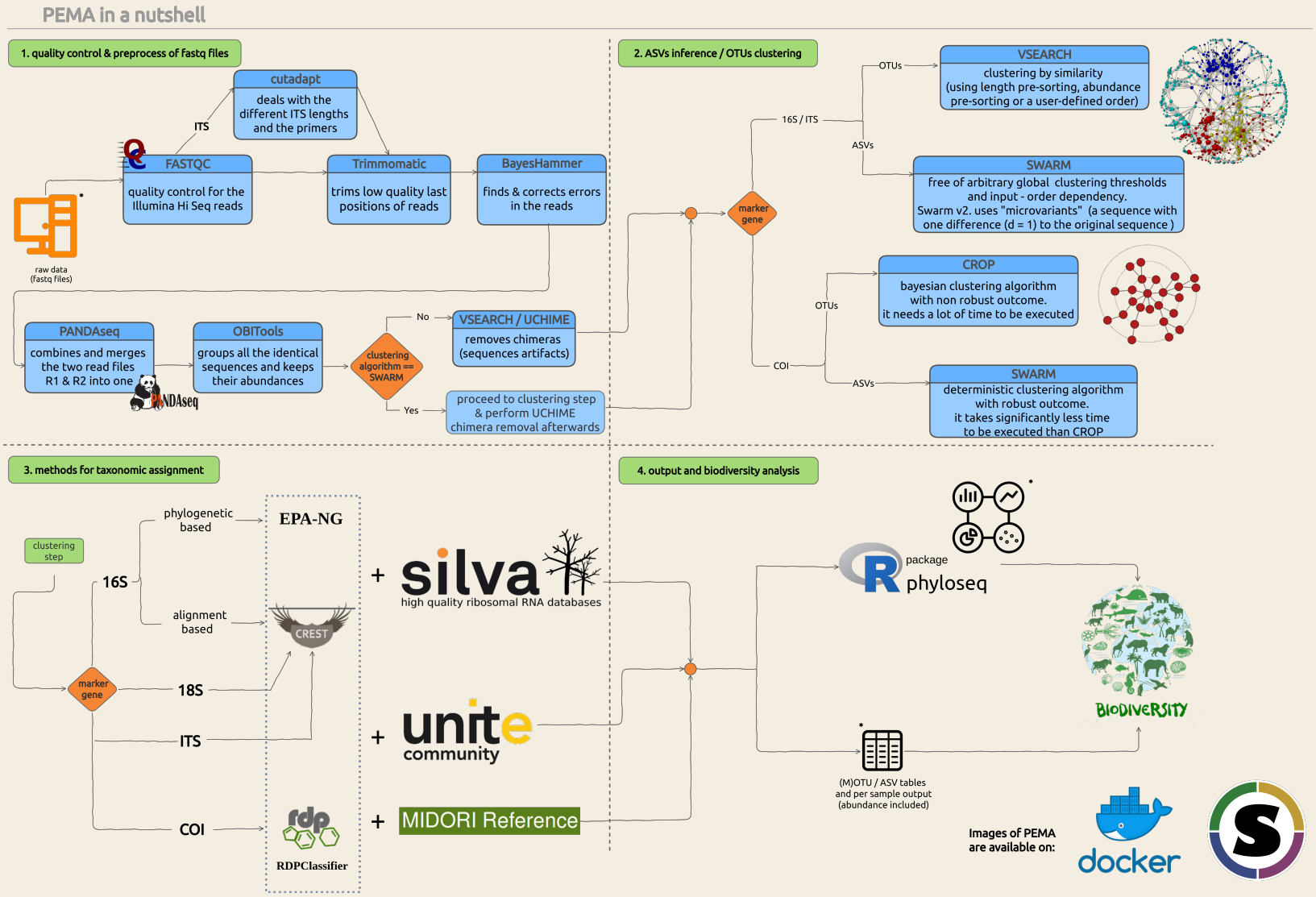
Big-DataScript programming language

Containerization

High performance computing

Results: PEMA features

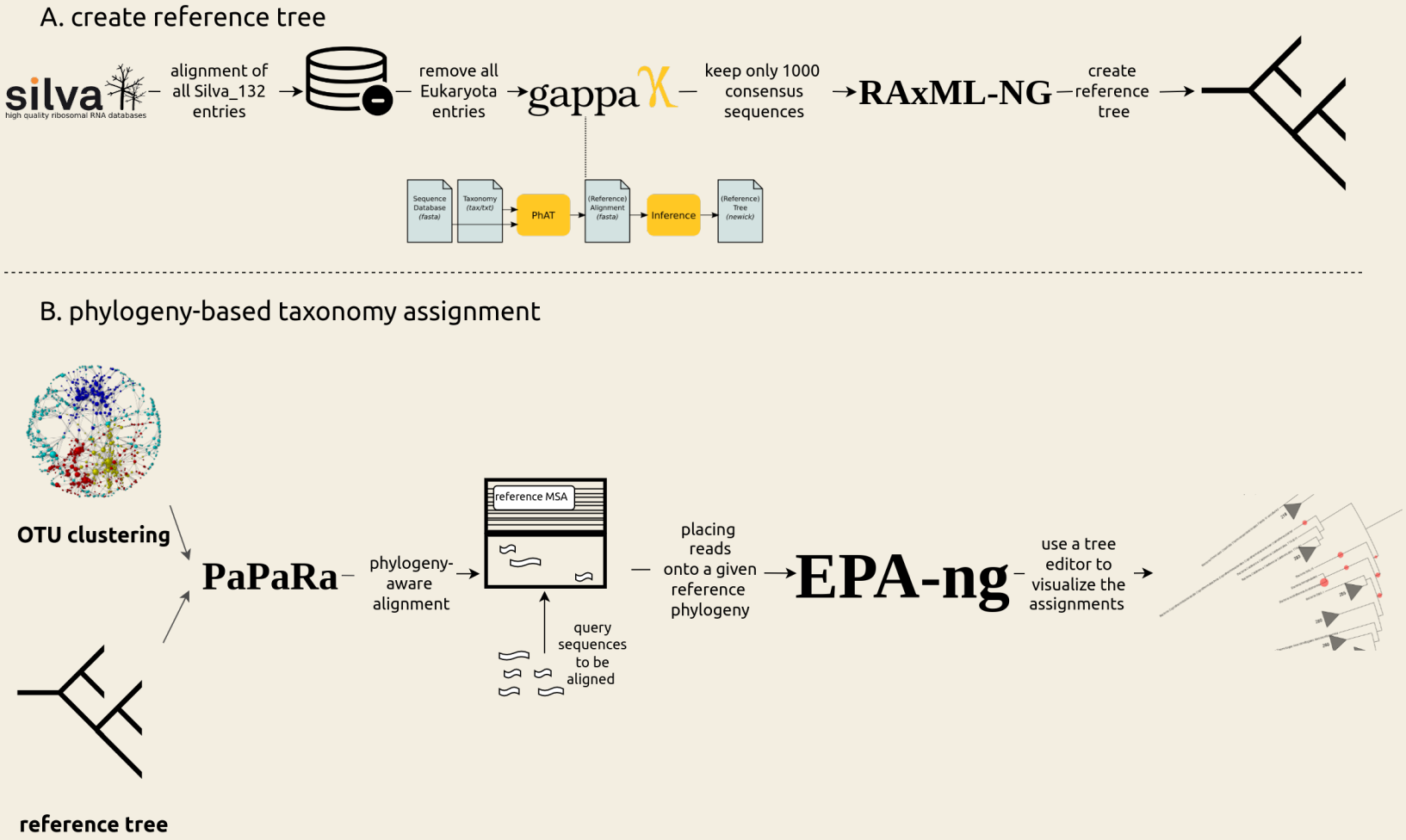
an overview



* Illustrations with asterisk are from the Noun Project

Results: PEMA features

phylogeny-based taxonomy assigment for the case of 16S rRNA gene



Results: tuning effetcts

in mock communities

Mock communities using the 16S rRNA gene and multiple parameter sets:

mock community of Gohl et al. (2016) KAPA protocol	Swarm (d = 1 strict = 0.8 no singletons)	Swarm (d = 1 strict = 0.8 with singletons)	Swarm (d = 3 strictness = 0.6 no singletons)	Swarm (d = 3 strictness = 0.8 with singletons)	Swarm (d = 10 strictness = 0.8 with singletons)	Swarm (d = 10 strictness = 0.8 no singletons)	Swarm (d = 25 strictness = 0.6)	Swarm (d = 25 strictntess = 0.8)	Swarm (d = 30 strictness = 0.6)	Swarm (d = 30 strictness = 0.8)
TP	12		15	18	18	15	17	17	17	17
FP	2		2	21	11	6	5	5	4	5
FN	8		5	2	2	5	3	3	3	3
PREC (TP / TP+FP)	0.86		0.88	0.46	0.62	0.71	0.77	0.77	0.81	0.77
REC (TP / TP+FN)	0.6		0.75	0.9	0.9	0.75	0.85	0.85	0.85	0.85
F1 (2 * (PREC * REC) / (PREC+REC)	0.71		**0.81**	0.61	0.73	0.73	0.81	0.81	0.83	0.81

mock community of Gohl et al. (2016) KAPA protocol	vsearch (id =0.95 strict = 0.8)	vsearch (id =0.97 strictness = 0.8)	vsearch (id =0.99 strictness = 0.8)
TP	11	12	12
FP	3	3	3
FN	9	8	8
PREC (TP / TP+FP)	0.79	0.80	0.80
REC (TP / TP+FN)	0.55	0.6	0.6
F1 (2 * (PREC * REC) / (PREC+REC)	0.65	0.69	0.69

Results: Tuning effects

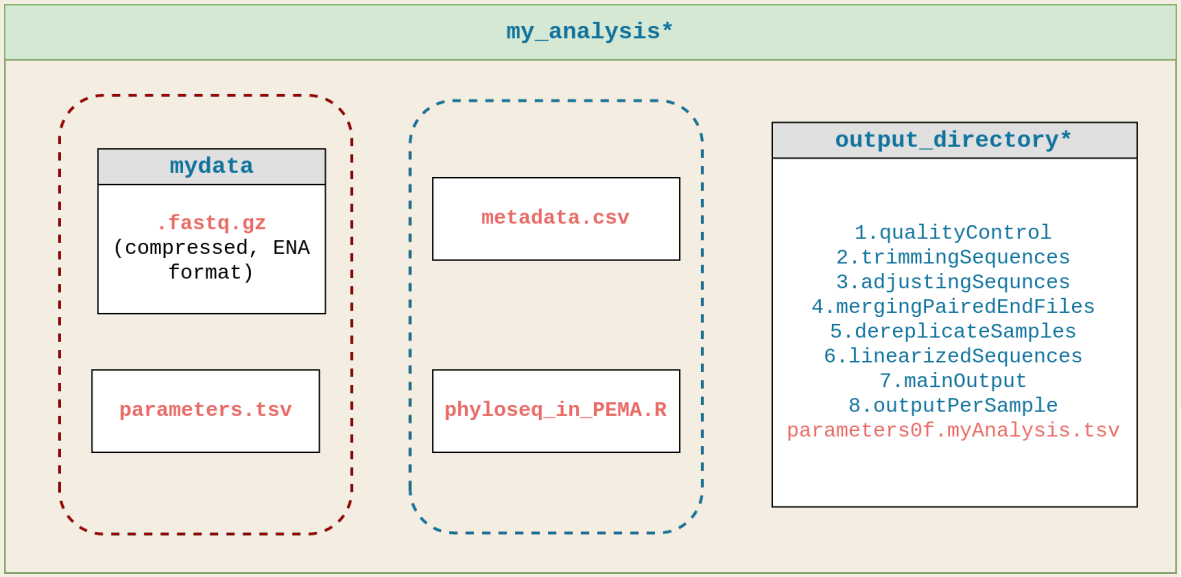
in real-world data

Using the *Bista* et al. dataset (COI) and multiple *d* values of the Swarm algorithm

Parameter	***d* = 1**	***d* = 2**	***d* = 3**	***d* = 10**	***d* = 13**
MOTUs after pre-process and clustering steps	83,791	59,833	33,227	7,384	4,829
MOTUs after chimera removal	80,347	57,863	32,539	7,339	4,796
Non-singleton MOTUs	6,381	4,947	2,658	1,914	1,634
Assigned species	62	83	86	86	84
Execution time (h)	2:01:35	2:09:49	1:51:44	2:17:26	2:31:15

Mount your I/O

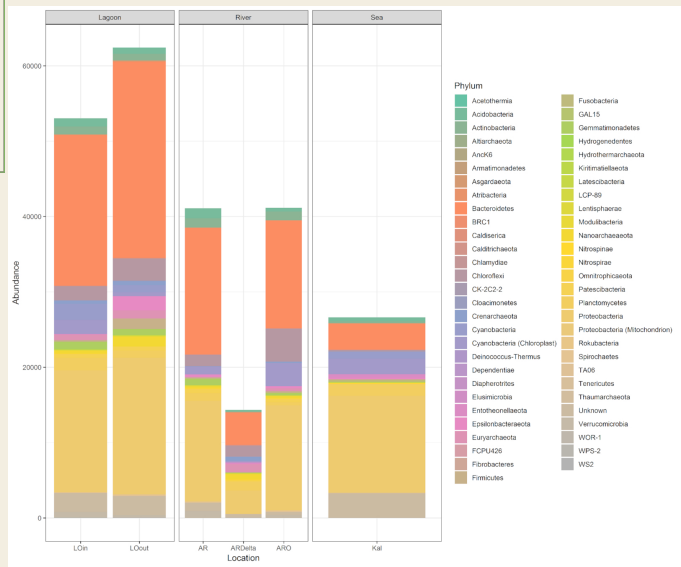
give & take



text file
directory
***user can edit the name**
(the rest **need** or will have the exact names shown)

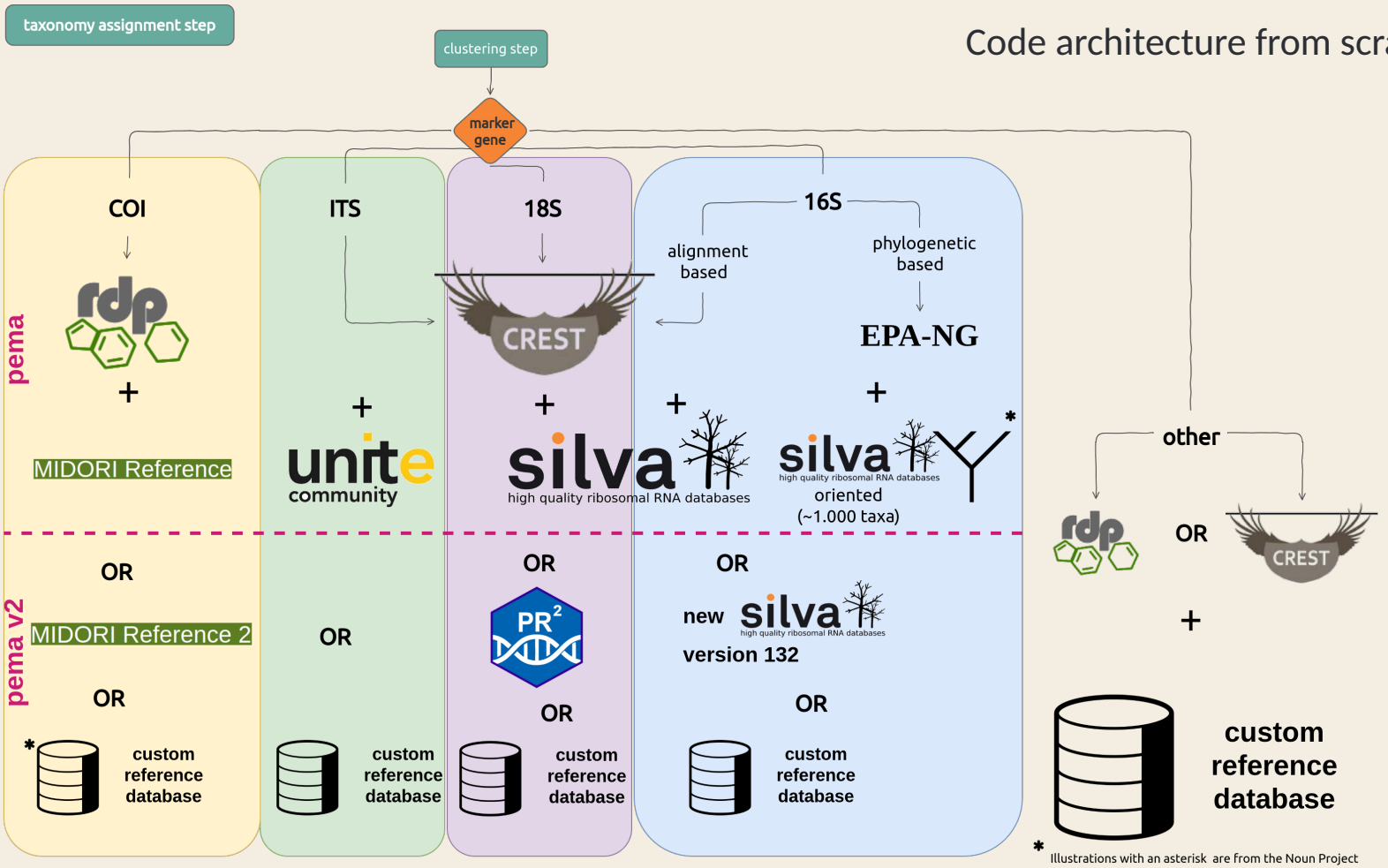
- - - - - mandatory input files
- - - - - optional input files

	Sample 1	Sample 2	Sample 3	Sample 4
Taxon 1	1	0	1	2
Taxon 2	0	1	0	2
Taxon 3	1	1	0	4



PEMA v.2

addressing some of the challenges



Code architecture from scratch!

Latest PEMA version

addressing the challenges of the community

ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

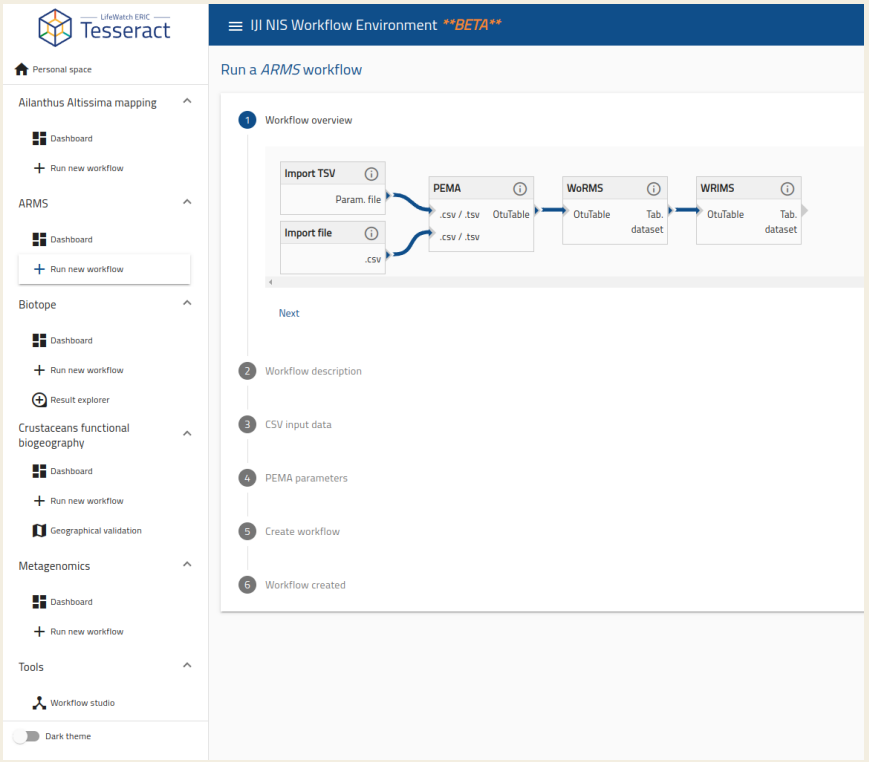
MBON
Marine Biodiversity Observation Network

`pema:v.2.1.4` includes:

1. analysis of 12S rRNA data now supported (12S Vertebrate Classifier v2.0.0-ref database)
2. PR2 as an alternative reference database for the case of 18S rRNA
3. the `ncbi-taxonomist` tool was added to return the NCBI Taxonomy Id of the taxonomies found

Moving at the large scale

PEMA @ infrustuctures



1. Web - interface make analysis even easier
2. researchers without access to HPC/clouds etc are now able to run big scale analyses

darn

https://github.com/hariszaf/darn

1. **Chapter 1**: Introduction

2. **Chapter 2**: Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
    2.1 **Chapter 2.1**: PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes
    2.2 darn: known unknowns in COI amplicon data

3. PREG0: a knowledge-base for organisms - environments - processes associations

related repositories under

https://github.com/orgs/lab42open-team

web-app under

http://prego.hcmr.gr/