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

1. Introduction

2. A regional High Performance Computing perspective

3. Software development to support HTS-oriented methods for microbial diversity assessment
    - 3.1 PEMA: a Pipeline for Environmental DNA Metabarcoding Analysis
    - 3.2 DARN: investigating the known unknowns in COI amplicon data

Microbial ecology & biogeochemical cycles

a corner-stone for life on earth

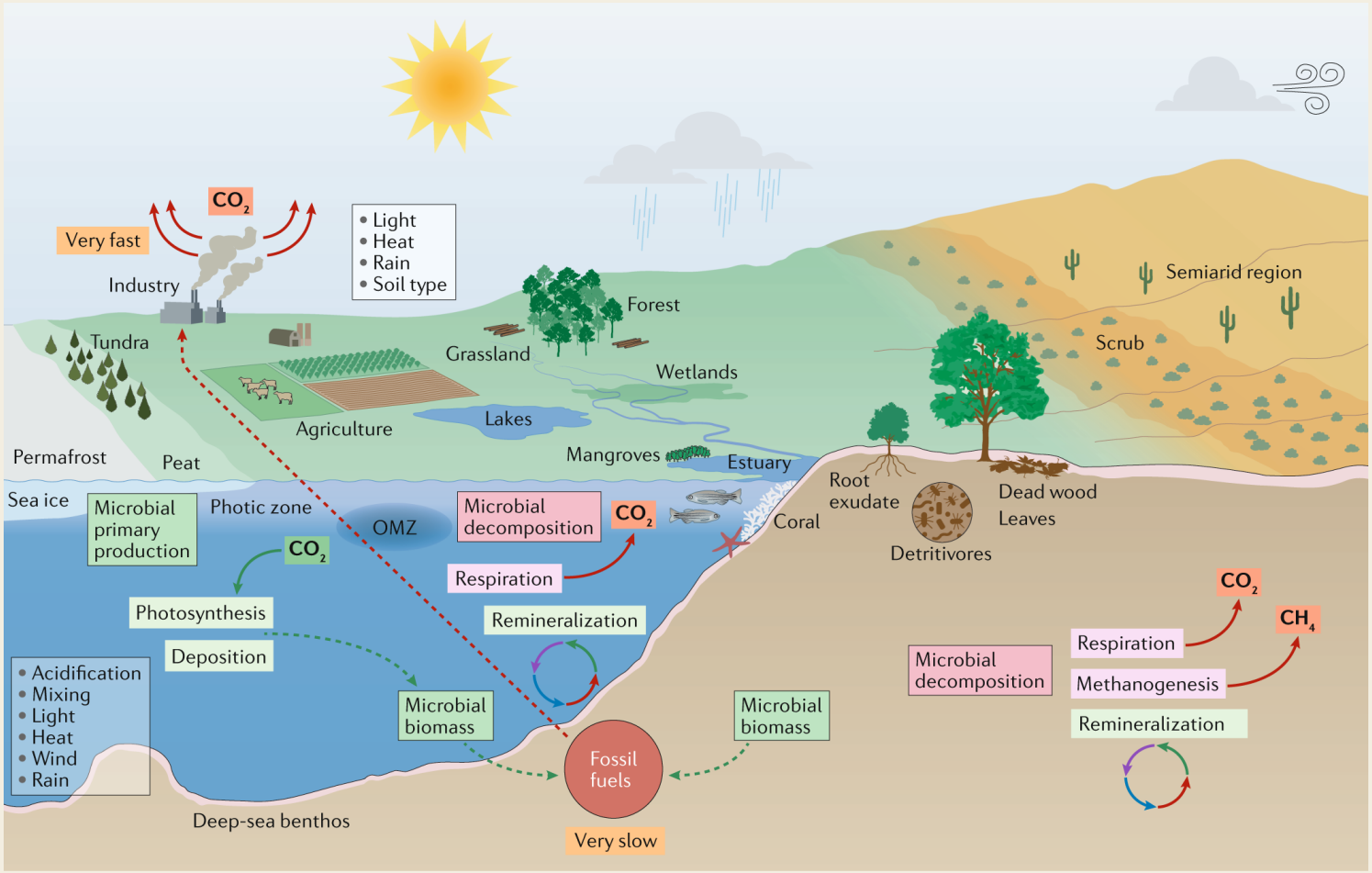


Figure from: Nature Reviews Microbiology 17.9 (2019): 569-586.

Main questions regarding a microbial community

for a deeper understanding of such assemblages

community structure
who

taxa, abundance

ecosystem type
where

habitats

functional potential
what

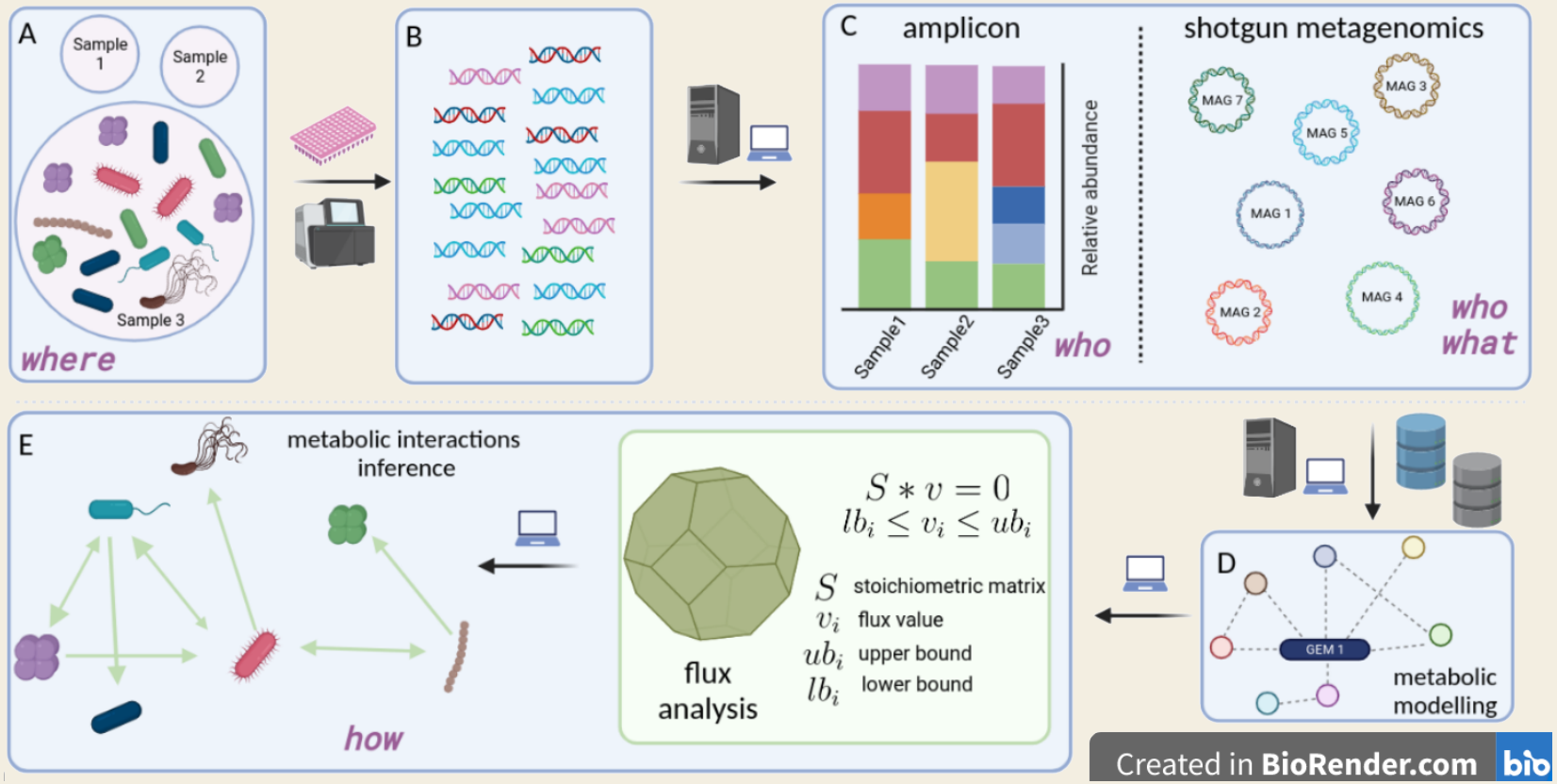
processes

underlying mechanisms
how

intreactions, fluxes

Reverse ecology

transforming ecology into a high-throughput field



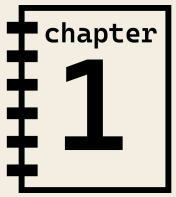
the study with no *a priori* assumptions about the organisms under consideration, by exploiting HTS and systems biology approaches

From raw reads to community analysis

not a straight-forward way

- biology-oriented issues
- techonology-oriented issues
- computing requirements
- data and metadata accessiblity

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

Graphical abstract of this PhD thesis

1. Introduction

2. A regional High Performance Computing perspective

3. Software development to support HTS-oriented methods for microbial diversity assessment
    - 3.1 PEMA: a Pipeline for Environmental DNA Metabarcoding Analysis
    - 3.2 DARN: investigating the known unknowns in COI amplicon data

0s and 1s in marine molecular research: a regional HPC perspective

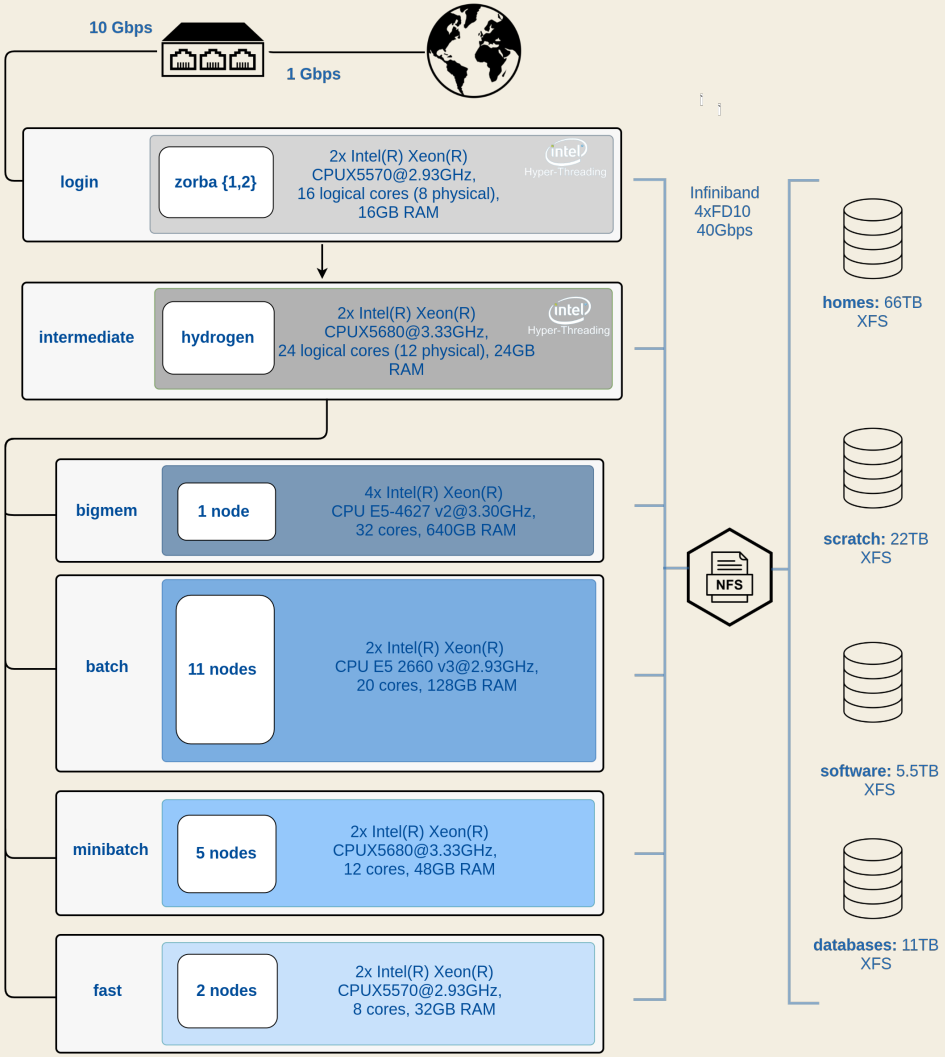
Aim of the study and contribution

To present insights from a thorough analysis of the research supported by the IMBBC HPC facility and some of its latest usage statistics in terms of resource requirements, computational methods, and data types as well as how the latter contributed in shaping the facility along its lifespan

chapter 6

Zorbas: the HPC facility of IMBBC

a Tier 2 (regional) HPC facility

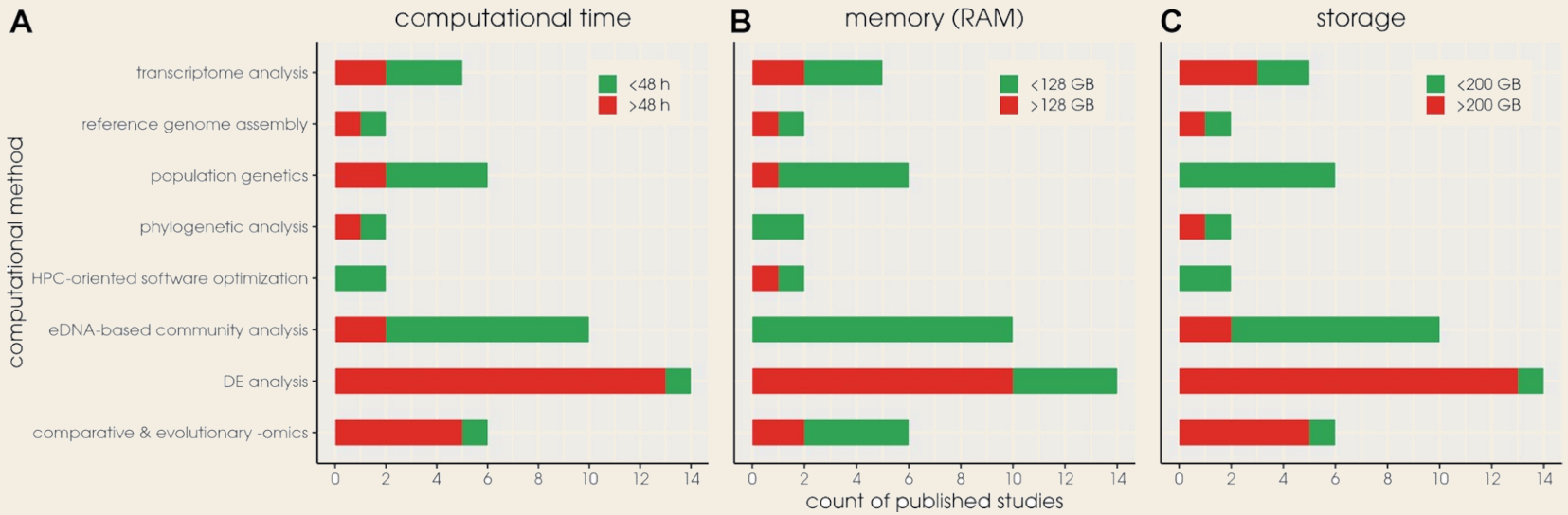


Block diagram of the Zorba architecture

chapter 6

Computational requirements

for trivial bioinformatic tasks



Red bars denote published research with high resource requirements
of the various computational methods employed at the IMBBC HPC facility

1. Introduction

2. A regional High Performance Computing perspective

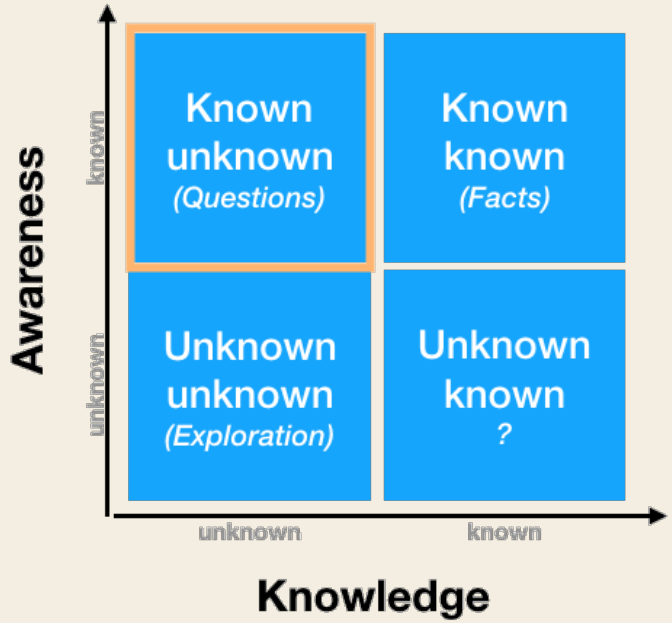
3. Software development to support HTS-oriented methods for microbial diversity assessment
    - 3.1 PEMA: a Pipeline for Environmental DNA Metabarcoding Analysis
    - 3.2 DARN: investigating the known unknowns in COI amplicon data

chapter 2

Bioinformatics challenges

for the analysis and the interpretation of amplicon data

- multiple steps
- several tools & databases
- computing power
- scalability & reproducibilty



OTUs/ASVs with no taxonomy assignment:
novel or non-target taxa

PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis

Aim of the study and contribution

To build an open source pipeline that bundles state-of-the-art bioinformatics tools for amplicon analysis that is:

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

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

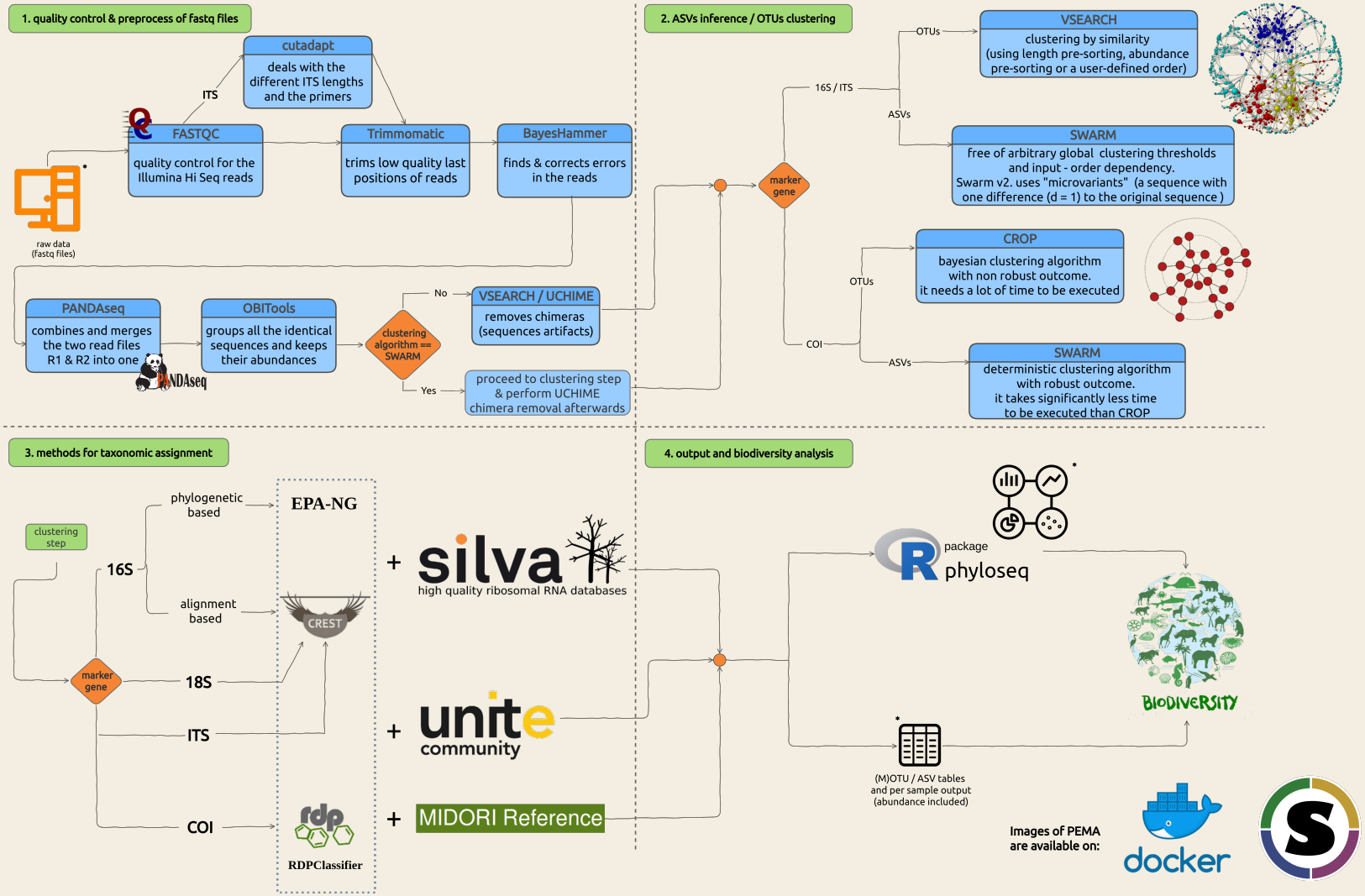
High performance computing

PEMA features

an overview

PEMA

PEMA in a nutshell



* Illustrations with asterisk are from the Noun Project

Results: tuning effetcts

in mock communities

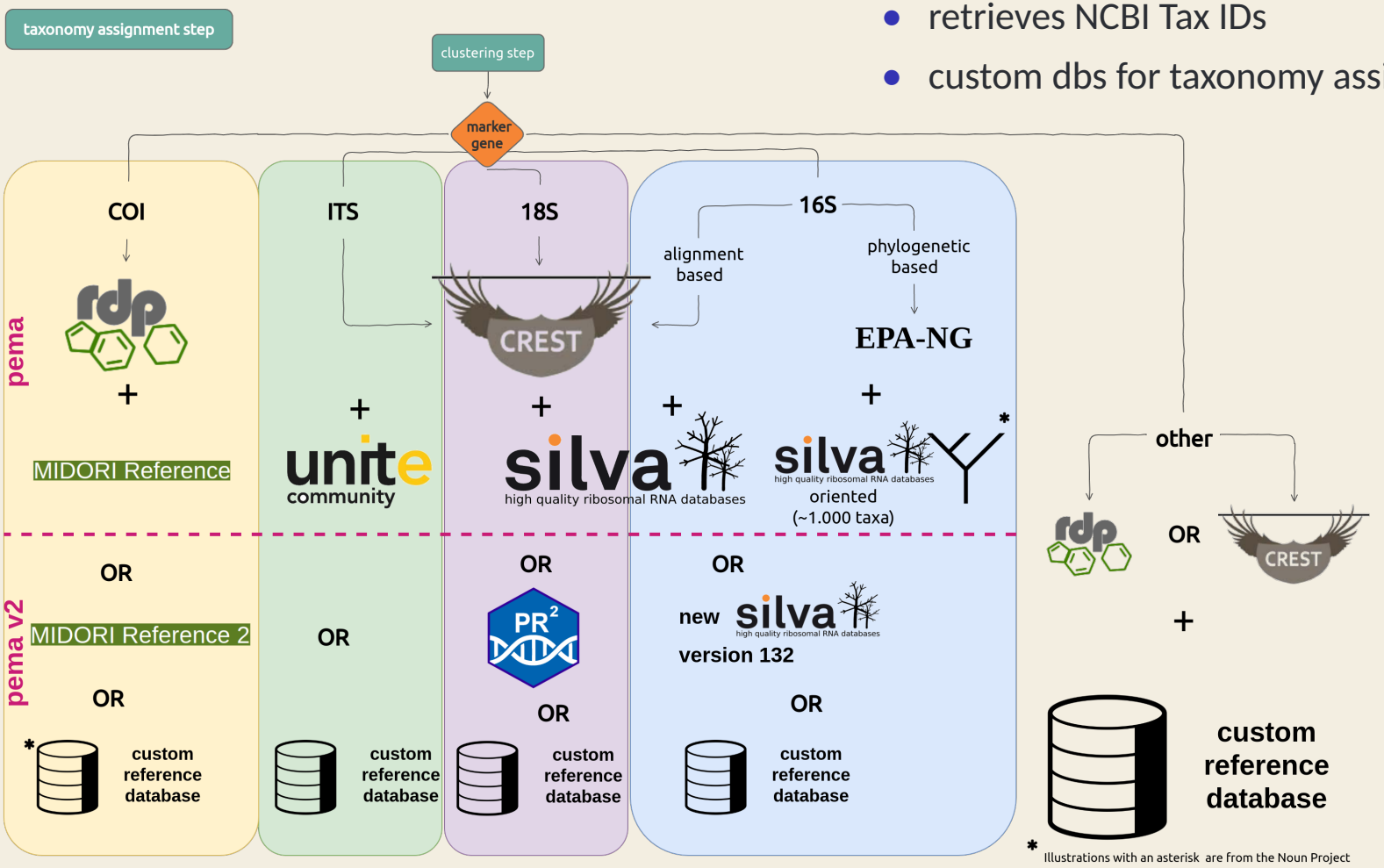
Mock communities using the 16S rRNA gene and multiple parameter sets (identification at the genus level; part of the initial table):

mock community Gohl et al. (2016)	**Swarm (d = 1 strict = 0.8 no 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)**
TP	12	15	18	18	15	17	17	17
FP	2	2	21	11	6	5	5	4
FN	8	5	2	2	5	3	3	3
PREC (TP / TP+FP)	0.86	0.88	0.46	0.62	0.71	0.77	0.77	0.81
REC (TP / TP+FN)	0.6	0.75	0.9	0.9	0.75	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

PEMA v.2

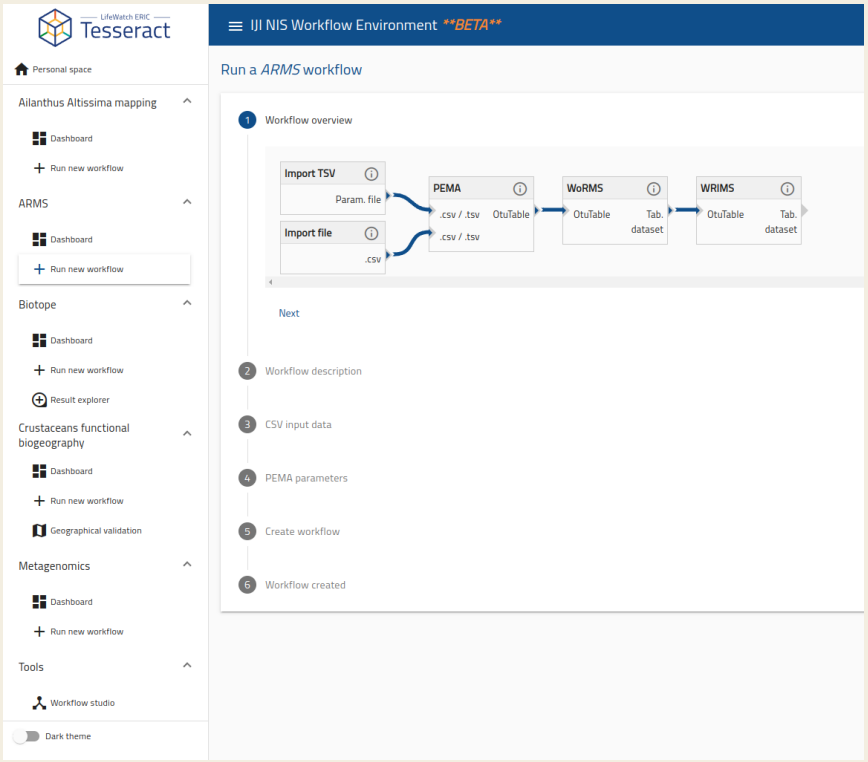
addressing some of the challenges

- new code architecture
- 12S rRNA now supported
- retrieves NCBI Tax IDs
- custom dbs for taxonomy assignment



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
3. Combine with other tools

LifeWatch ERIC:
www.lifewatch.eu/
Tesseract VRE Development Portal:
www.lifewatch.dev/dashboard

Conclusions

on PEMA and eDNA metabarcoding

- tuning is essential in metabarcoding analyses
- sequencing a mock community along with your samples can be of great help in parameter tuning
- difference between well studied and unexplored environments
- infrustuctures benefit studies with great number of samples and CLI non-familiar users

The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data

Aim of the study and contribution

Extract "dark matter" from COI amplicon data
i.e. non-target, unassigned or assigned with low confidence sequences.

Methodology / Implementation

sequences retrieved

Resources	bacteria		archaea		eukaryotes	
	# of sequences	# of strains	# of sequences	# of strains	# of sequences	# of species
BOLD	3,917	2,267	117	117		
PFam-oriented	9,154	4,532	217	115		
Midori 2					1,315,378	183,330
Total unique entries	**11,421**	**6,798**	**334**	**201**	**1,315,378**	**183,330**

↓

	bacteria	archaea	eukaryotes
consensus sequences (tree branches)	463	25	1,109

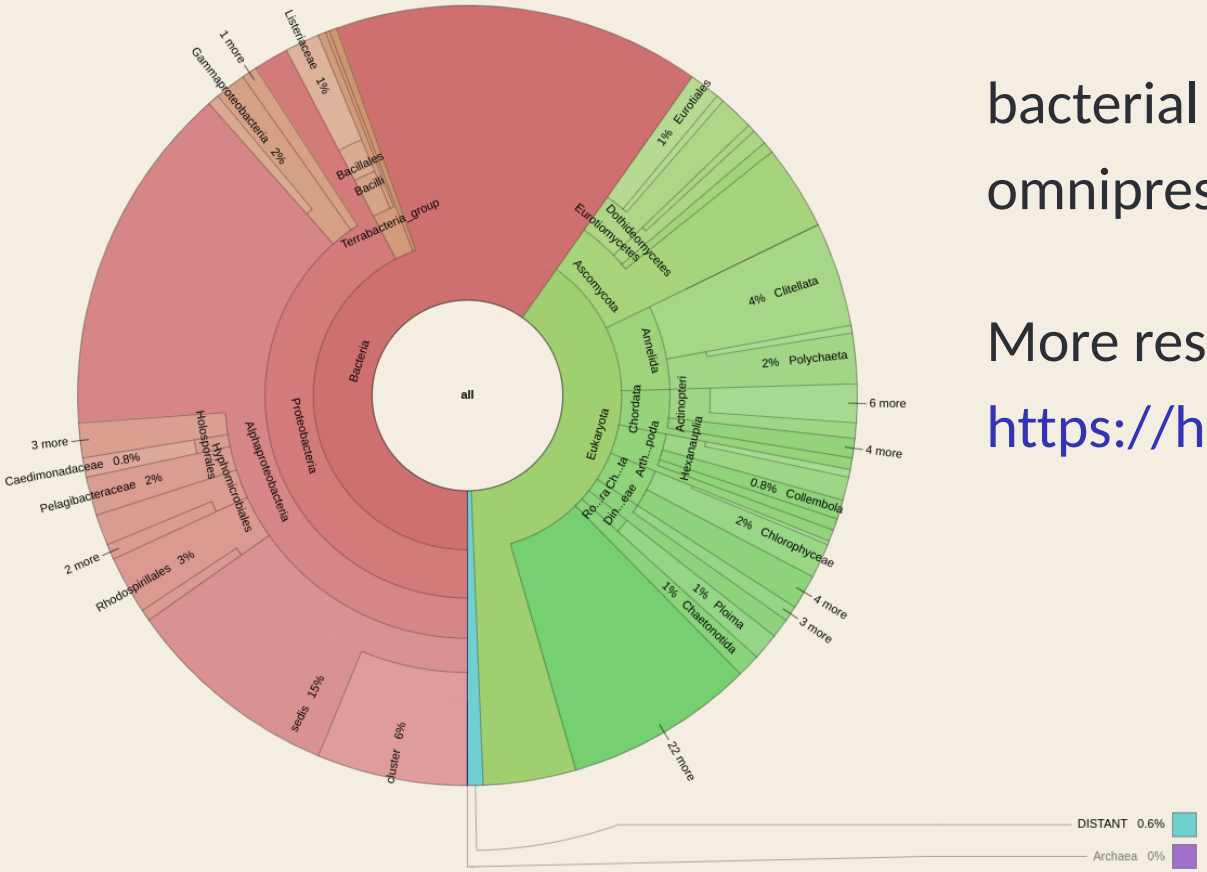
Methodology / Implementation

Dark mAtteR iNvesigator

Envo type	Sample type	Bioinfo pipeline(s)	# of ASVs	~% of sequence assignments per domain		
				eukaryotes	bacteria	archaea
marine	eDNA	QIIME2 - Dada2	13,376	**11**	88	**0.02**
		PEMA (d=10)	39,454	**25**	75	**0.1**
marine	bulk	PEMA (d=2)	193	**99**	**1**	**-**
marine	bulk	PEMA (d=2)	74	**97**	0	**-**
lotic	eDNA	PEMA (d=10)	1,940	**64**	34	**2**

Results: DARN using real-world data

with multiple sample types, primers, PCR protocols and bioinormatics pipelines



bacterial sequences are omnipresent in COI amplicon data

More results at:

https://hariszaf.github.io/darn/

Conclusions

on DARN and COI amplicon studies

- bacteria, algae, fungi etc. was verified to be present in COI amplicon data
- bacteria make up a significant proportion of sequences generated in COI based eDNA metabarcoding datasets
- dark matter seems to be particularly common in eDNA as compared to bulk samples
- DARN supports quality control and further investigation of the unassigned OTUs/ASVs and allows researchers to better understand the known unknowns
- We argue that the vast majority of these sequences represent microbial taxa, such as bacteria and archaea.