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

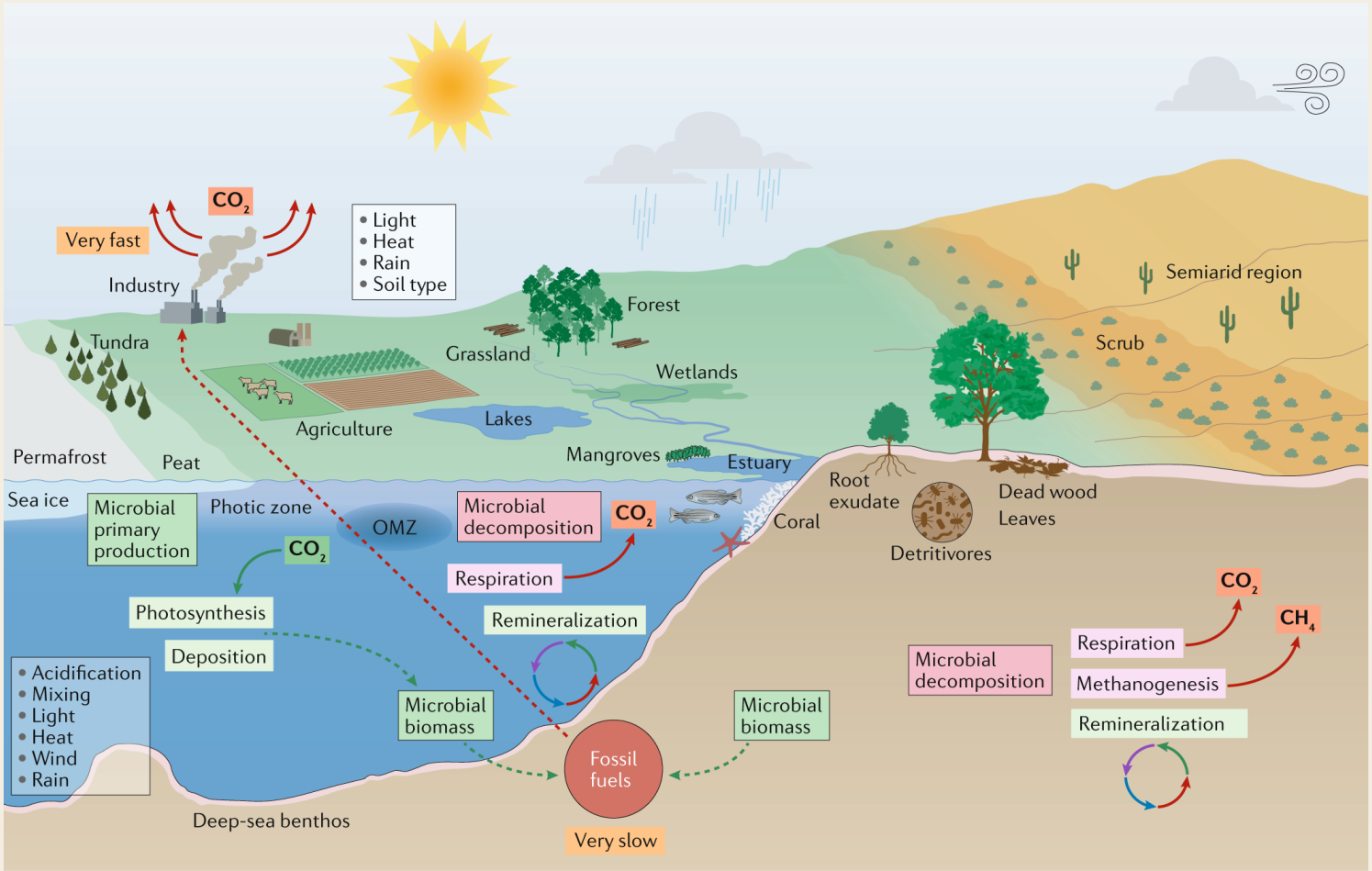


Figure from: Nature Reviews Microbiology 17.9 (2019): 569-586.

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

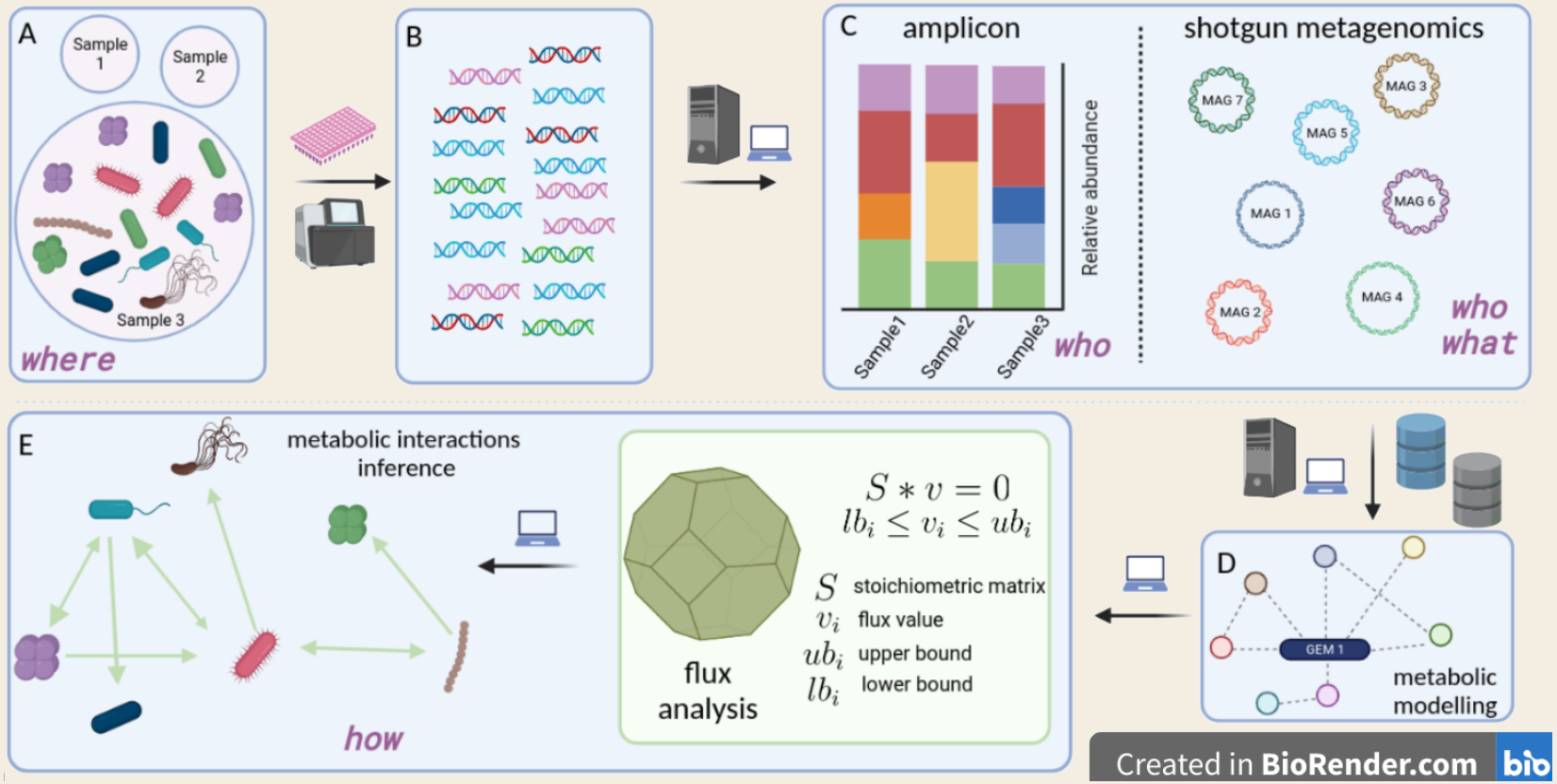
processes

underlying mechanisms
how

intreactions, fluxes

Reverse ecology

transforming ecology into a high-throughput field



From raw reads to community analysis

not a straight-forward way

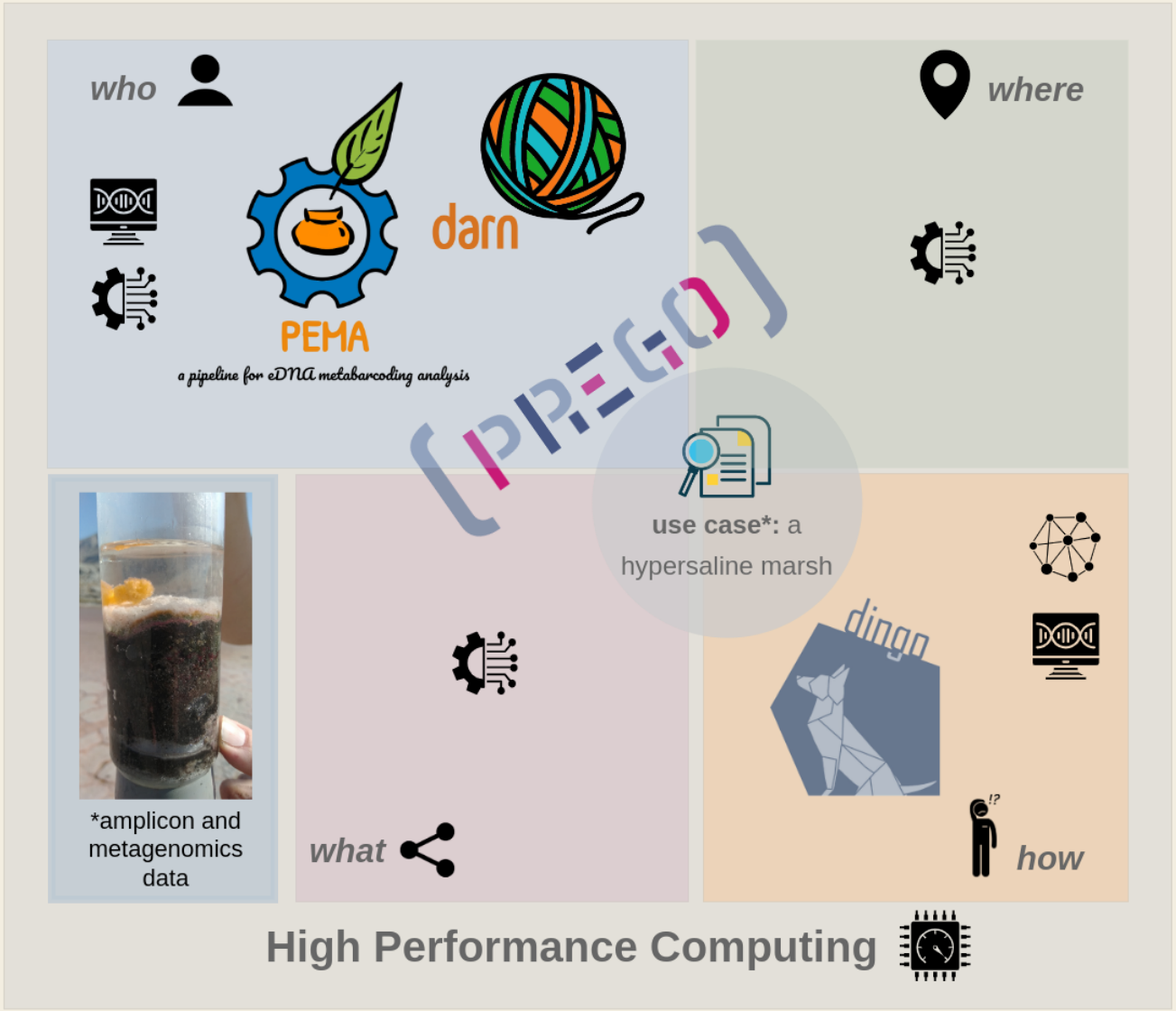
- sequencing technology oriented
- **eDNA-**, sample process-, primers-, PCR-oriented
- **complex bioinformatics analysis**
- **analysis reproducibility**
- **computing requirements**
- **algorithm-oriented**
- versioning of reference databases & tools
- standards, ontologies
- **data integration applications**

Aims and objectives

- build algorithms and software to address on-going challenges in microbiome data analysis

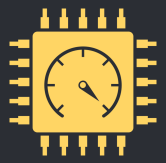
- identify taxa & functions with a key role in microbial community assemblages in hypersaline sediments

Graphical abstract of this PhD thesis



0s and 1s in marine molecular research

a regional HPC

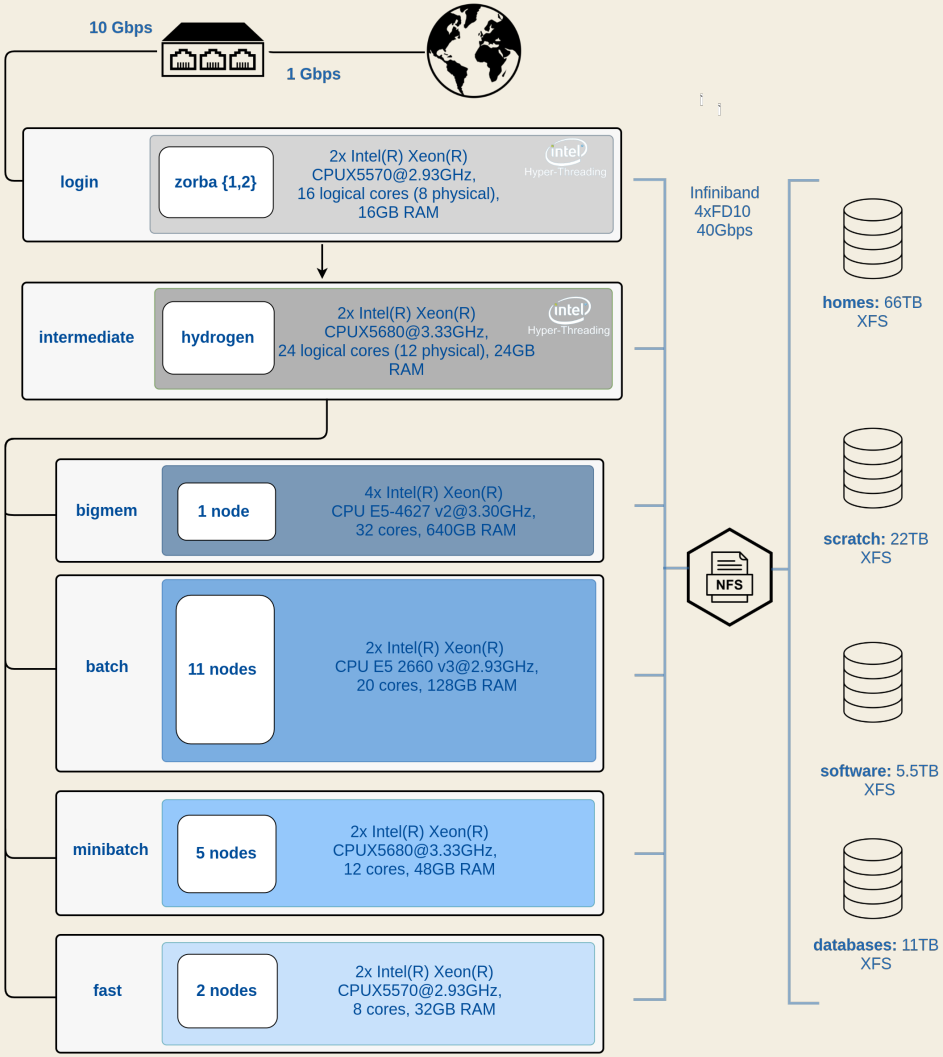


challenge category: compuring requirements

aim & contribution: to present insights from a thorough analysis of the research supported by the IMBBC HPC facility

Zorbas: the HPC facility of IMBBC

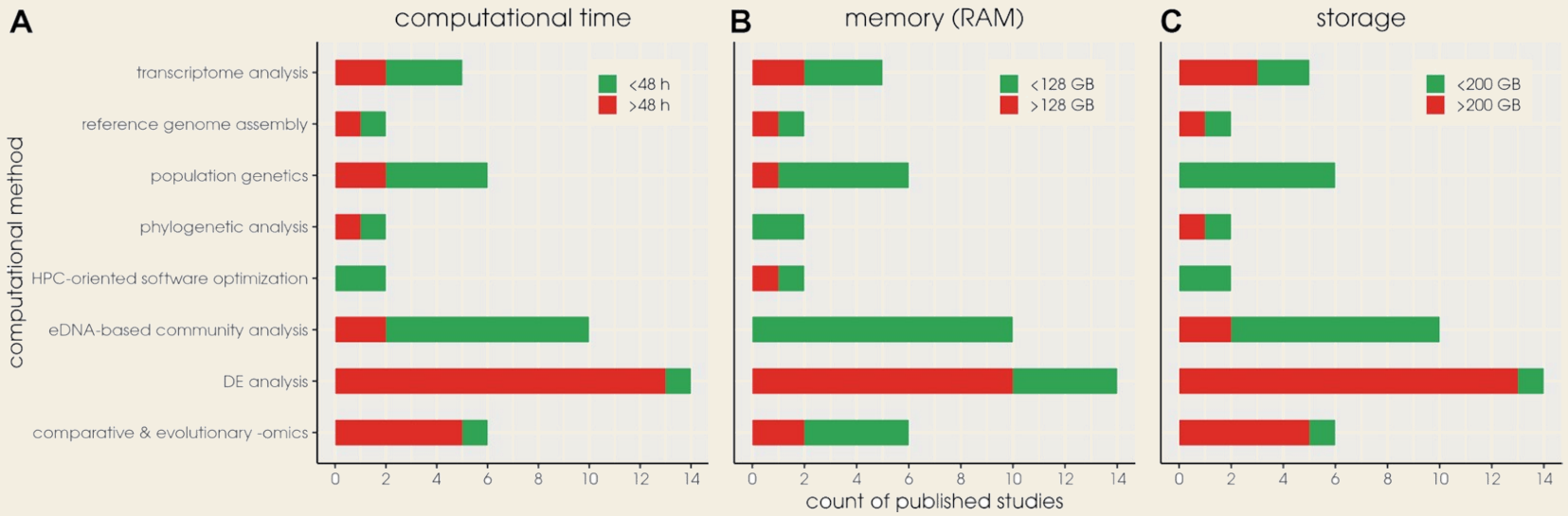
a Tier 2 (regional) HPC facility



Block diagram of the Zorba architecture

Computational requirements

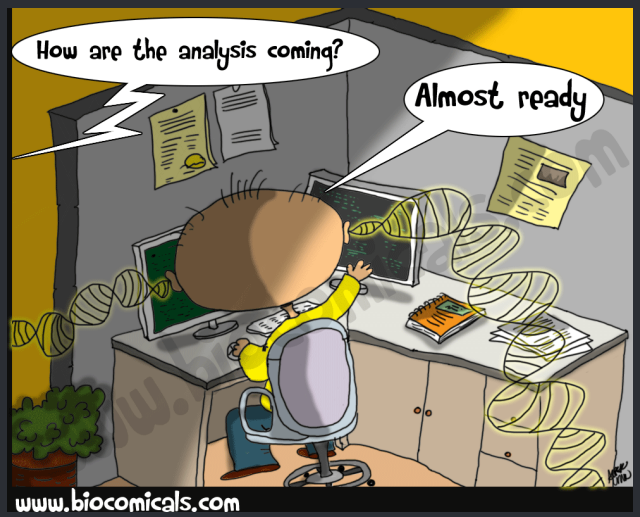
for trivial bioinformatic tasks



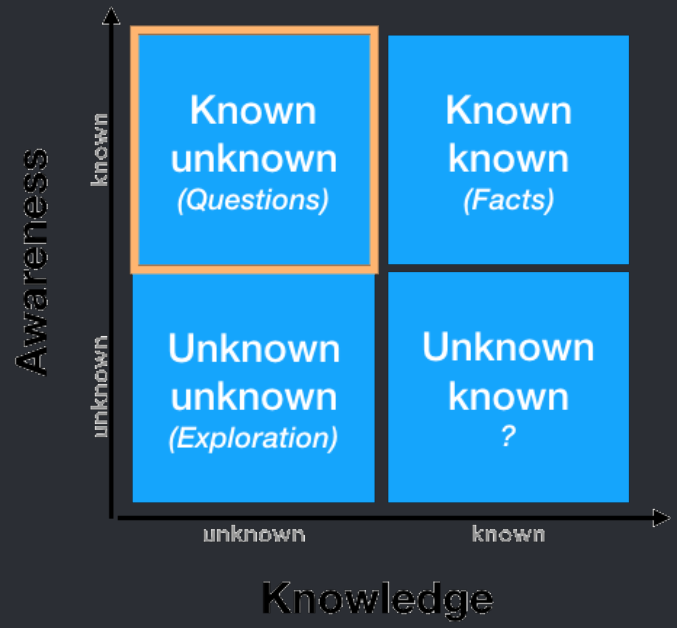
Red bars denote published research with high resource requirements
of the various computational methods employed at the IMBBC HPC facility

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

challenge category: complex bioinformatics analysis, analysis reproducibility

aim & 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

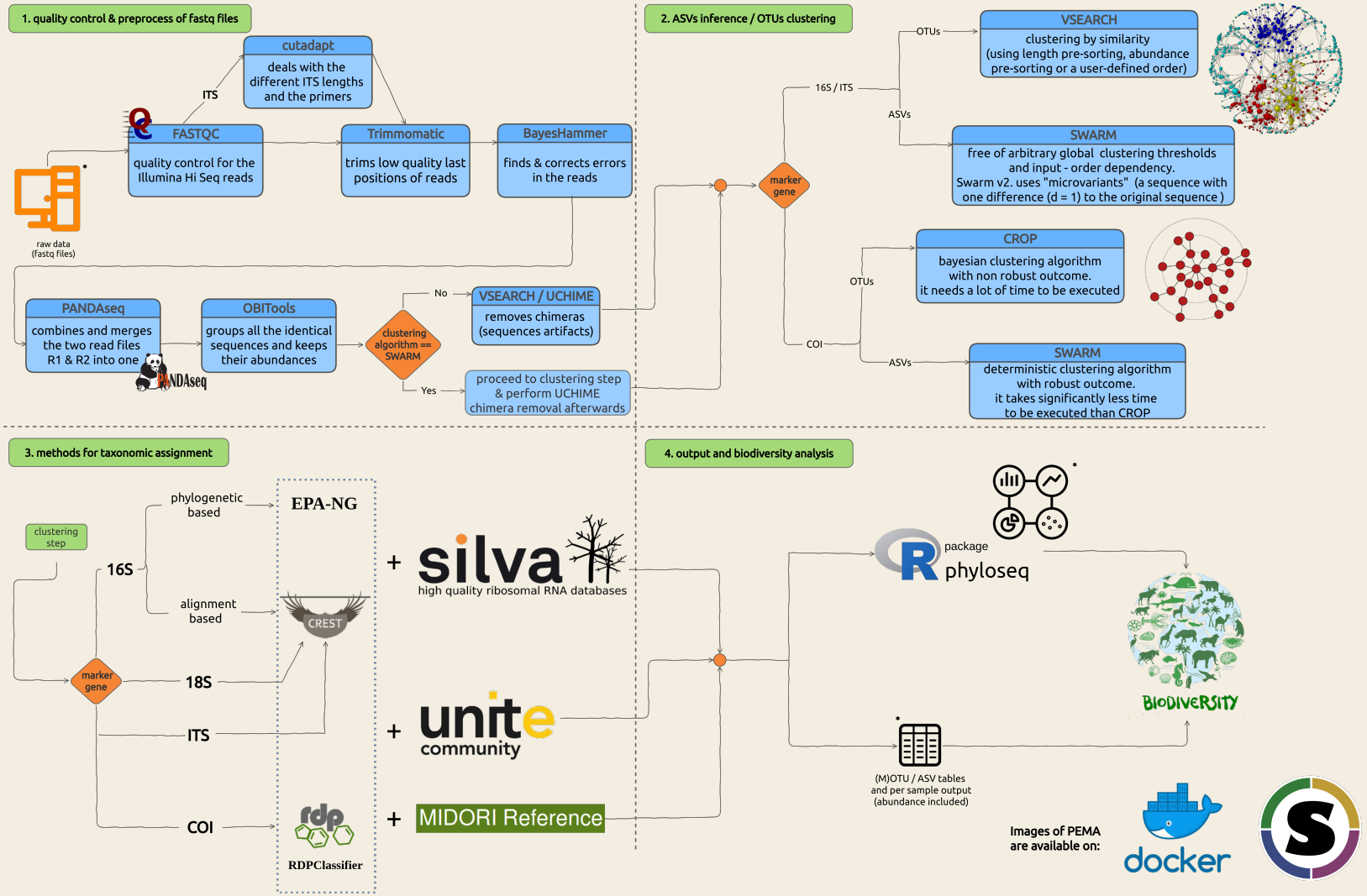
High performance computing

PEMA features

an overview

PEMA

PEMA in a nutshell



* Illustrations with asterisk are from the Noun Project

Results: tuning effetcts

in mock communities

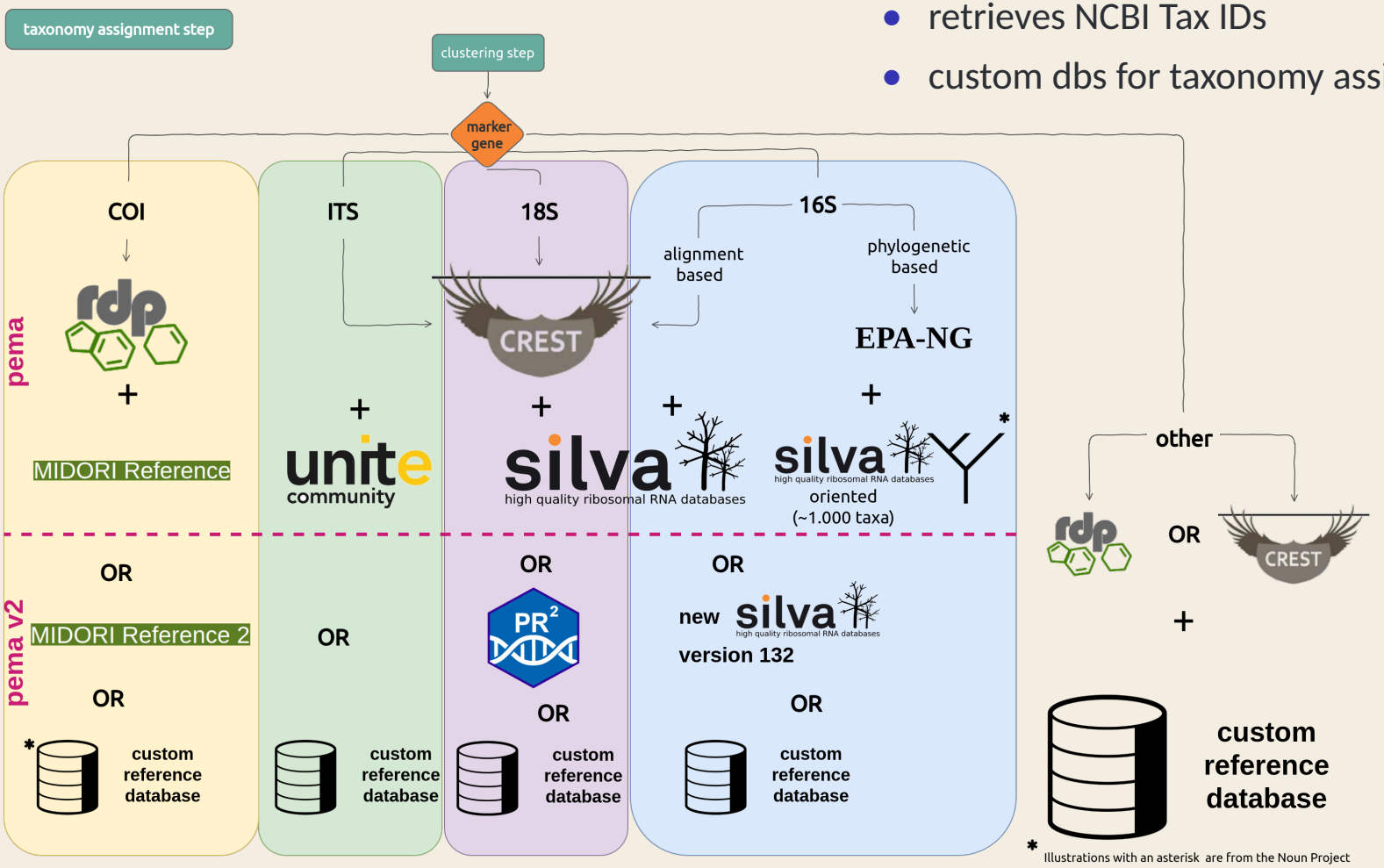
Mock communities using the 16S rRNA gene and multiple parameter sets (identification at the genus level; part of the initial table):

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

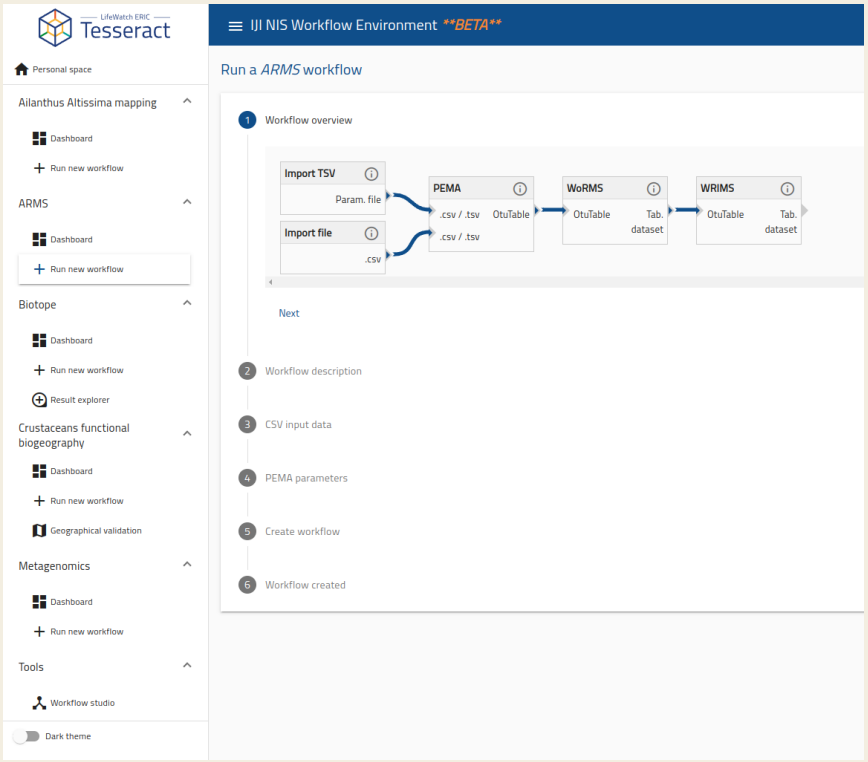
addressing some of the challenges

- 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

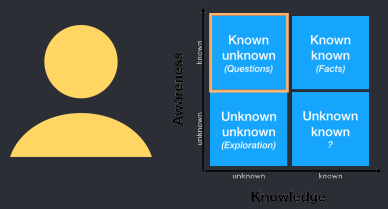
Conclusions

on PEMA and eDNA metabarcoding

- **parameters tuning** is essential in metabarcoding analyses; **mock communities** among samples under study benefit to that end
- data derived from **well-studied** and **unexplored** environments differ in their analysis
- **e-infrustuctures** benefit studies with great number of samples and CLI non-familiar users

DARN

dark matter invesigator tool



challenge category: eDNA - oriented issues

aim & contribution:

extract "dark matter" from COI amplicon data

i.e. non-target, unassigned or assigned with low confidence sequences.

Methodology / Implementation

sequences retrieved

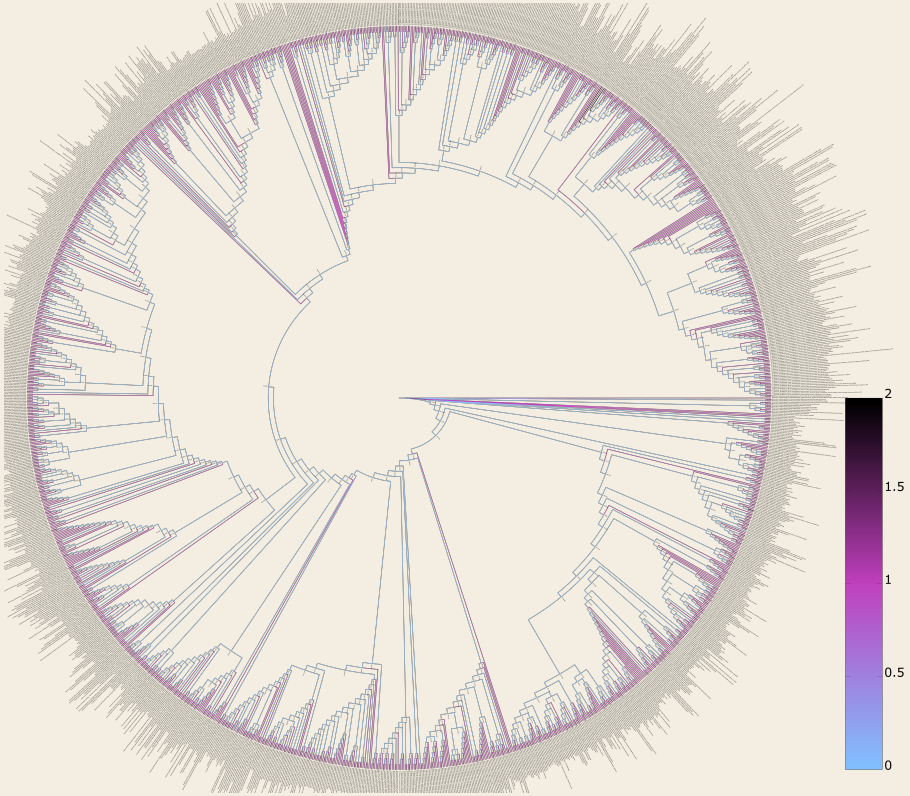
Resources	bacteria		archaea		eukaryotes	
	# of sequences	# of strains	# of sequences	# of strains	# of sequences	# of species
BOLD	3,917	2,267	117	117		
PFam-oriented	9,154	4,532	217	115		
Midori 2					1,315,378	183,330
Total unique entries	**11,421**	**6,798**	**334**	**201**	**1,315,378**	**183,330**

↓

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

Methodology / Implementation

Phylogenetic tree of the COI consensus sequences retrieved

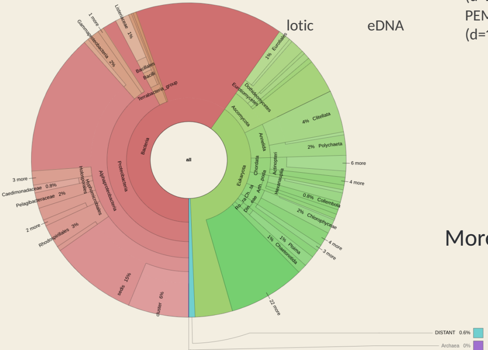


the consensus sequences have been placed in their corresponding taxonomic branches, proving the tree valid

DARN using real-world data

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

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**



More results at: https://hariszaf.github.io/darn/

Conclusions

on DARN and COI amplicon studies

- dark matter is widely common in eDNA samples compared to bulk ones
- bacteria make up a significant proportion of sequences generated in COI based eDNA metabarcoding datasets
- including non-eukaryotic COI sequences in reference databases could benefit the method

dingo

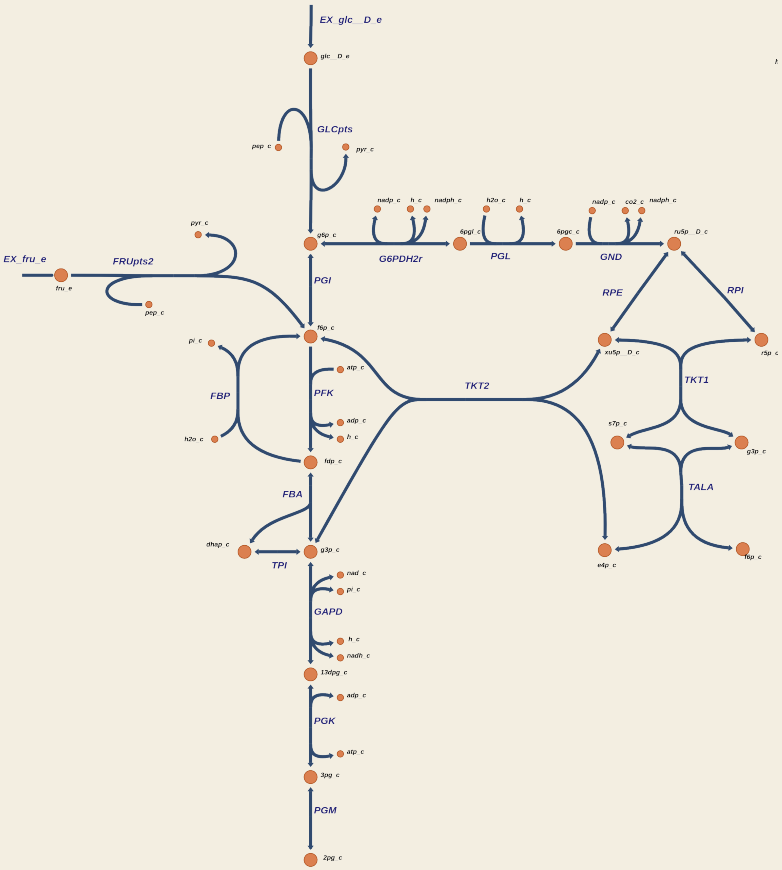
a new MCMC algorithm for sampling the flux Space of metabolic networks

challenge category: algorithm-oriented

aim & contribution:
support flux sampling at high dimensional polytopes such as those of complex organisms and multispecies and/or host-microbe communities

Metabolic modelling

and the biomass function



Metabolic models allow us

to move from a metabolic map

to mathematical structures

the study of which may provide

fundamental biological insight

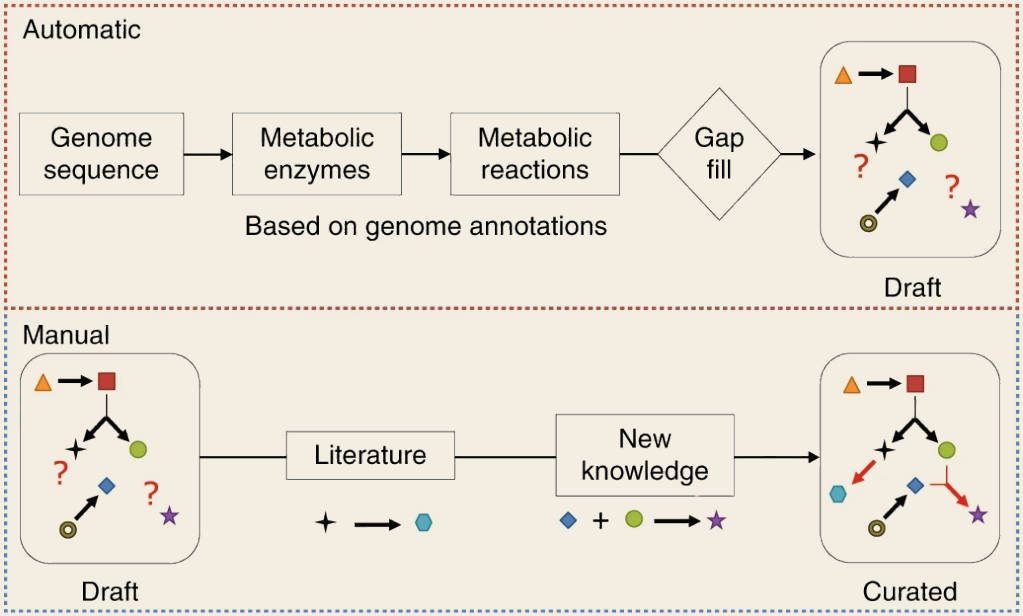
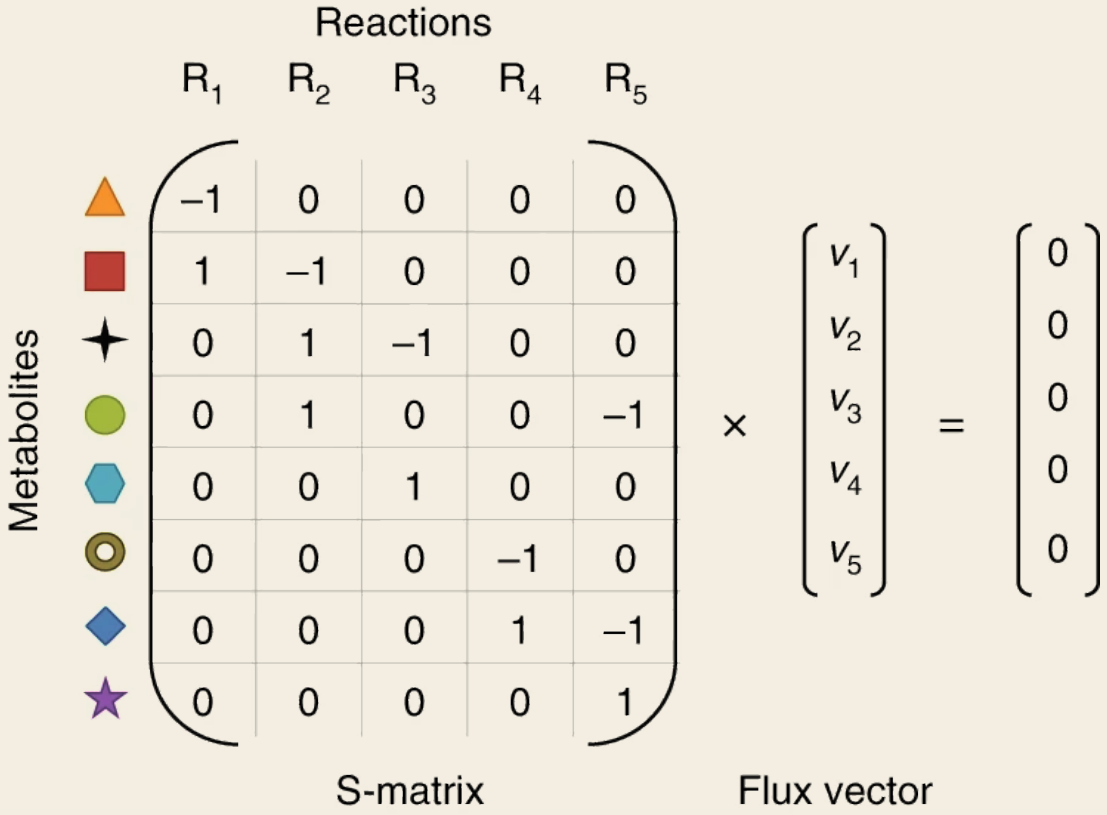


Figure from: Heirendt et al. Nature protocols 14.3 (2019): 639-702.

From a stoichiometric matrix

to a constraint-based model



In a **steady state**
the production rate
of each metabolite
equals its consumption rate

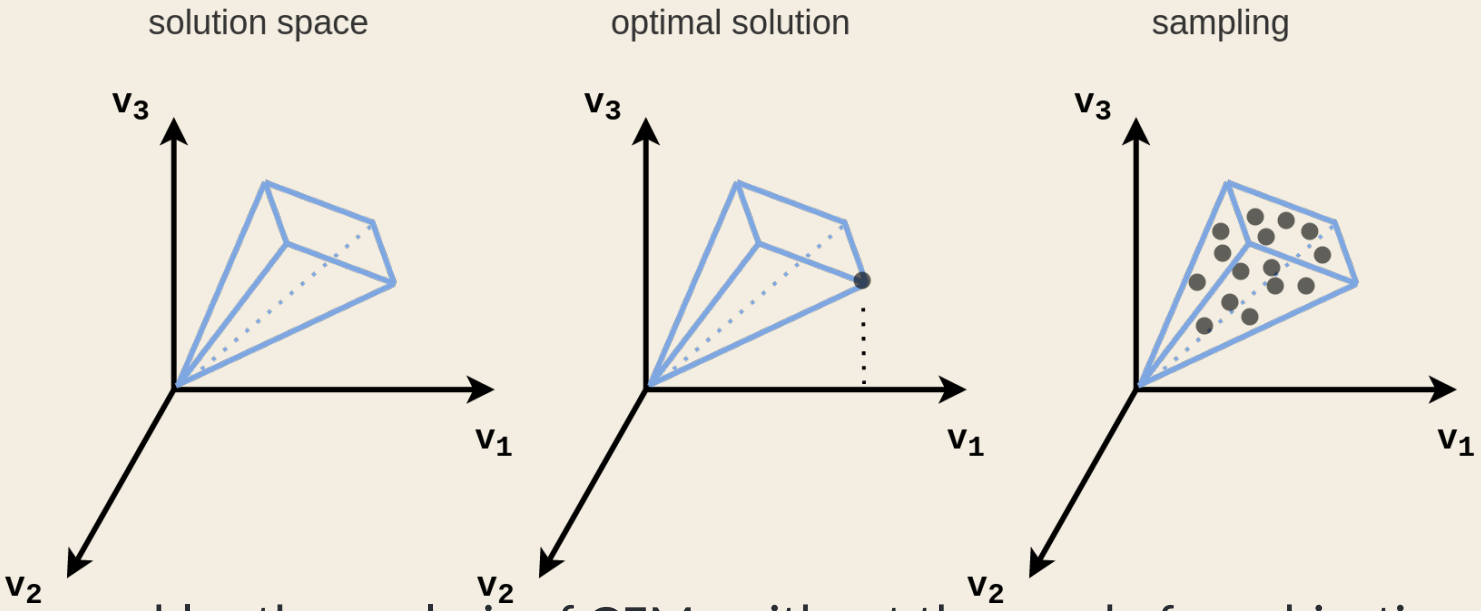
The **flux vector** is a vector with
the value of
each reaction flux
in a certain steady
state.

The steady state assumption
is ensured by
the **zero-vector**.

Figure from: Heirendt et al. Nature protocols 14.3 (2019): 639-702.

Flux sampling

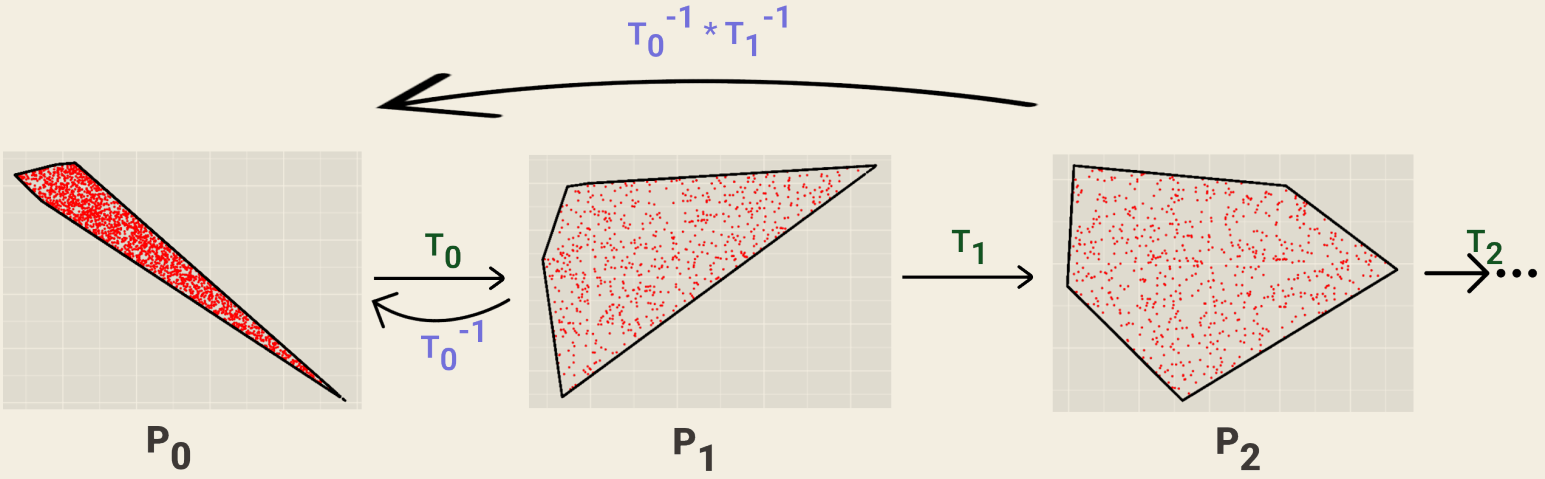
an alternative approach



- enables the analysis of GEMs without the need of an objective function
- determines the feasible solution spaces for fluxes in a network based on a set of conditions as well as the probability of obtaining a solution

A new Markov Chain Monte Carlo algorithm

for flux sampling

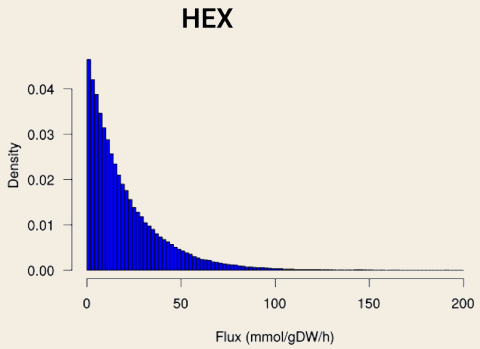


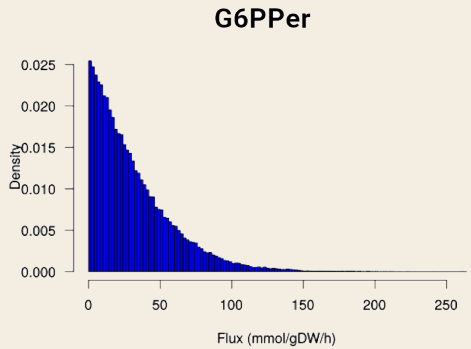
Steps of an MMCS phase

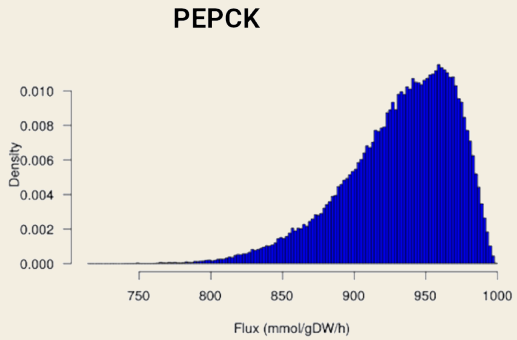
- **sampling step:** using a variant of the **Billiard walk**
- **rounding step:** calculate a linear transformation T_i that puts the sample into isotropic position and then apply it on P_i to obtain the polytope of the next phase
- check several statistic tests

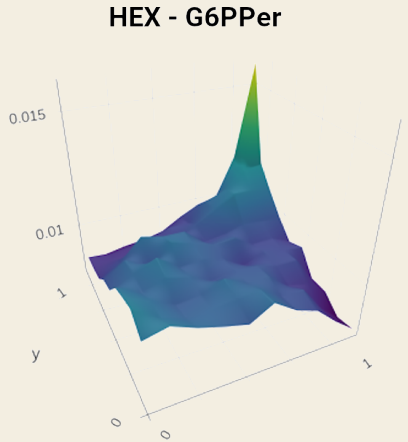
Flux sampling output

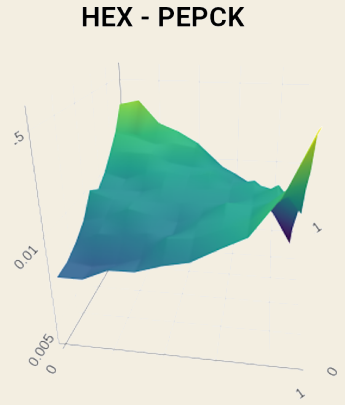
marginal distributions and copulas











HEX:
atp_c + **glc__D_c** → adp_c + g6p_c + h_c :

G6PPer:
g6p_r + h2o_r → pi_r + **glc__D_r**

PEPCK:
gtp_c + oaa_c → co2_c + gdp_c + pep_c

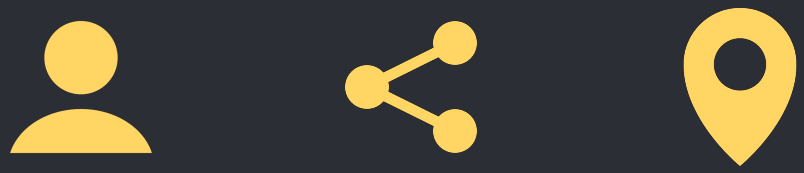
Conclusions

on sampling the flux space of metabolic models

- flux sampling provides essential insight (knock-out genes, host-microbe interactions etc)
- our multiphase MCMC algorithm enables sampling on the so-far largest metabolic model (Recon3D including 13, 543 reactions; $d = 5, 335$) is now possible
- sampling the flux space of community metabolic models is now possible

PREGO

a literature- and data-mining resource to associate microorganisms, biological processes & environment types



challenge category: data integration applications

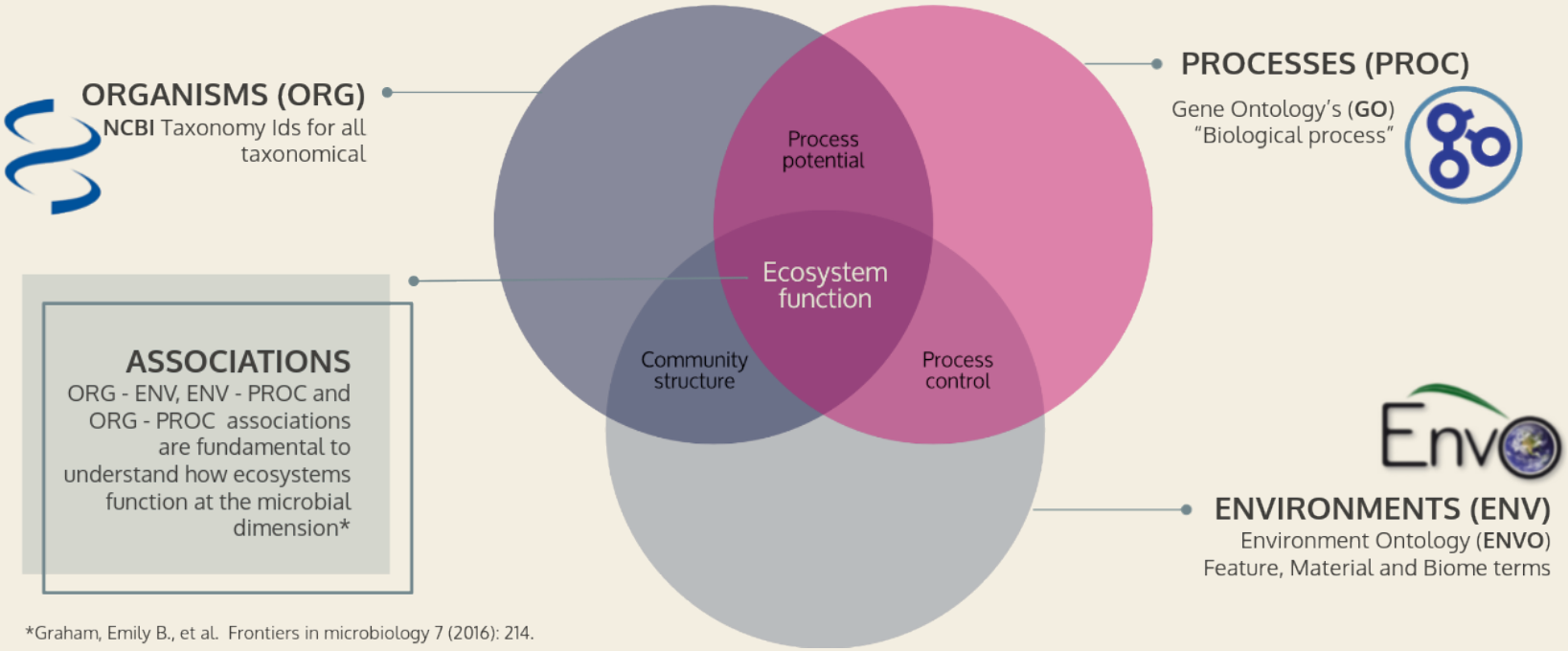
aim & contribution:

To build a hypothesis generation resource based on associations between:

- *organisms* and the *environments* they inhabit
- *organisms* and the biological *processes* they are involved with
- *processes* and the *environments* where they occur

PREGO and its referring terms

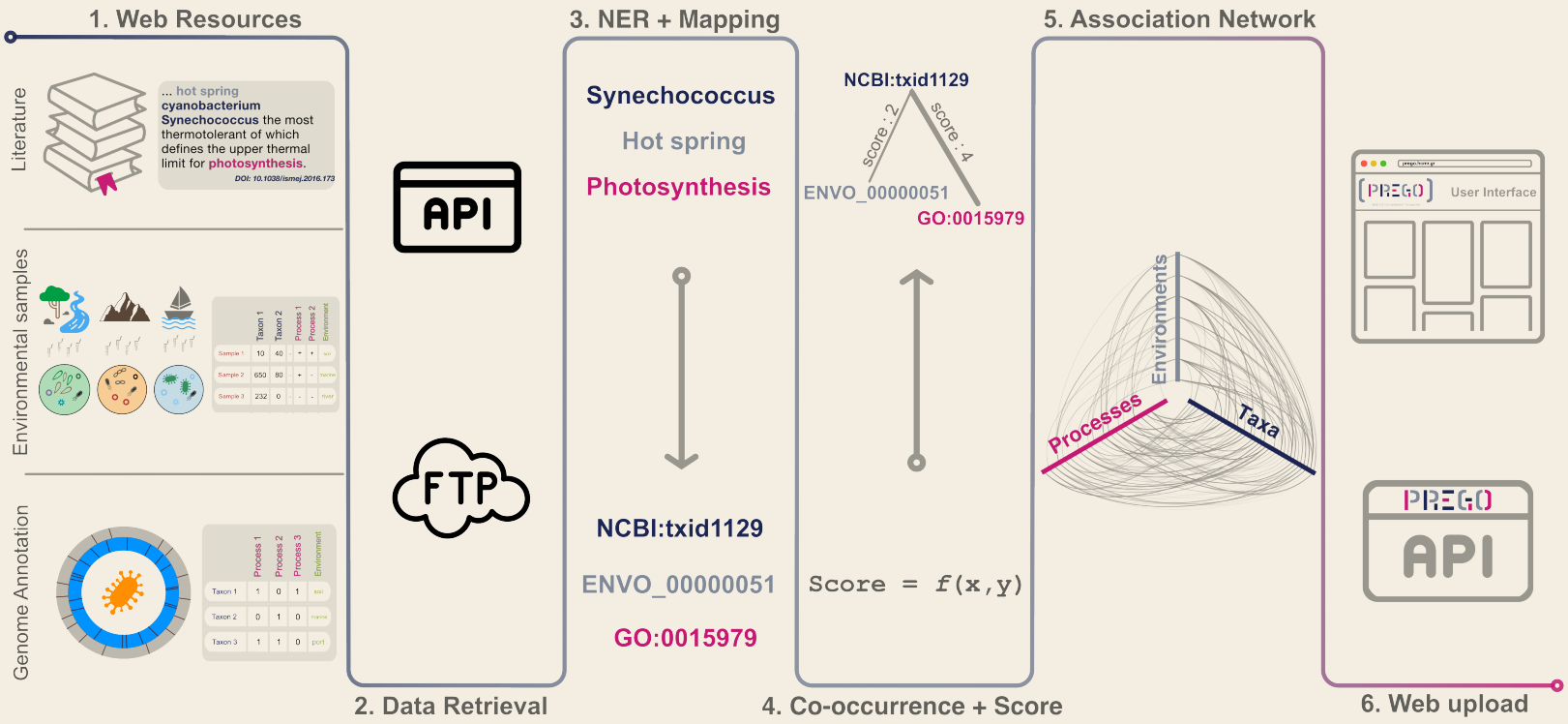
the fundamental role of ontologies



*Graham, Emily B., et al. Frontiers in microbiology 7 (2016): 214.

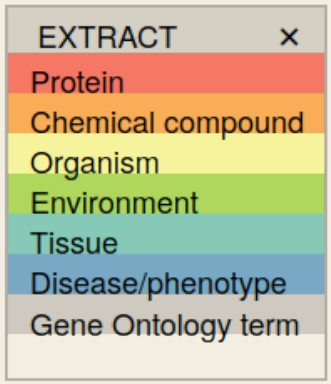
Methods / Implementation

3 channels of information - 1 framework



Named Entity Recognition

counting co-occurences to export associations



Identification of potentially important pathways missing from the model

From the metagenomic bins, we were able to identify two metabolic processes that were not previously included in the model. A number of MAGs (bin.59, bin.15, bin.73) clustered to the KEGG genomes of freshwater sulfur-oxidizing autotrophs capable of denitrification, *Sulfuritalea hydrogenivorans* [41], and *Sulfuricella denitrificans* [42]. These MAGs contained the diagnostic genes for carbon fixation (*rbcLS*), sulfur cycling (*dsrAB*), and denitrification (*nosZ*). One MAG (bin.59) also clustered with iron oxidizing autotroph *Sideroxydans lithotrophicus* ES-1. Bin.59 is the most relatively abundant bin from 17 to 21 m depth. Thus, if this MAG is associated with iron oxidation, it also contains sulfur-cycling genes that add to metabolic flexibility, which was previously observed [40]. The model did not include sulfide oxidation with nitrate, so it is unclear from the current model predictions where this process is expected to occur within the water column to compare to the MAG distributions.

Example text from Arora-Williams et al. Microbiome 6.1 (2018): 1-16.

The *Environmental Samples* and the *Annotated Genomes and Isolates* channels

the role of metadata

Sample metadata [-]

Collection date:	11/1/11
Elevation:	200
Environment (biome):	soil
Environment (feature):	nosZ
Environment (material):	soil DNA
Environmental package:	MIGS/MIMS/MIMARKS.soil
Geographic location (depth):	15-20cm
Instrument model:	454 GS FLX Titanium
Investigation type:	miens-survey
NCBI sample classification:	410658
Project name:	EcoFINDERS

Project Information	
Cultured	No
Ecosystem	Environmental
Ecosystem Category	Aquatic
Ecosystem Subtype	Oceanic
Ecosystem Type	Marine

MG-RAST ID	name	biome	feature	material	sample	library	location	country	coordinates	download
mgm4702467.3	06032015b_S2_L001_R2_001	Large lake biome	lake	water	mgs485560	mgl485562	Cincinnati	USA	39.11, -84.5	metadata submitted
mgm4702469.3	06052015a_S3_L001_R2_001	Large lake biome	lake	water	mgs485566	mgl485568	Cincinnati	USA	39.11, -84.5	metadata submitted
mgm4702471.3	06032015a_S1_L001_R2_001	Large lake biome	lake	water	mgs485554	mgl485556	Cincinnati	USA	39.11, -84.5	metadata submitted

Source databases that are integrated in PREGO

and the number of items retrieved

Source	# Items	Data Type	Metadata	License
MEDLINE and PubMed	33 million	abstracts (text)	no	NLM Copyright
PubMed Central OA Subset	2.7 million	full article (text)	no	CC for Commercial, non-commercial
JGI IMG	9,644	Isolates Annotated genomes	yes	JGI Data Policy
Struo	21,276	Annotated genomes	no	MIT, CC BY-SA 4.0
BioProject	18,752	Annotated genomes with abstracts (text)	yes	INSDC policy
MG-RAST	16,096	marker gene samples	yes	CC0
	7,965	metagenomic samples	yes	CC0
MGnify	10,500	marker gene samples	yes	CC-BY, CC0

PREGO in action

looking for environments a taxon is present

Desulfatiglans anilini DSM 4660 [1121399]

Synonyms: Desulfatiglans anilini DSM 4660, D. anilini DSM 4660, D anilini DSM 4660, Desulfatiglans DSM 4660, Desulfatiglans str. DSM 4660 ...

Environments | Biological Processes | Molecular Function | Documents | Downloads

Literature

Search:

Name	Z-score	Confidence
Oil seep	3.0	★★☆☆☆
Marine mud	3.0	★★☆☆☆
Marine sediment	2.8	★★☆☆☆
Brackish water	2.4	★★☆☆☆
Oil reservoir	2.2	★★☆☆☆
Anaerobic sediment	2.0	★☆☆☆☆
Cold seep	1.8	★☆☆☆☆
Contaminated sediment	1.7	★☆☆☆☆
Neritic sub-littoral zone	1.4	★☆☆☆☆
Oil spill	1.4	★☆☆☆☆
Petroleum	1.1	★☆☆☆☆
Sea floor	1.1	★☆☆☆☆

Showing 1 to 12 of 12 entries

Environments | Biological Processes | Molecular Function | Documents | Downloads

Literature

Search:

Name	Z-score	Confidence
benzoyl-CoA catabolic process	4.3	★★★☆☆
Benzene catabolic process	3.6	★★☆☆☆
Acetone metabolic process	3.5	★★☆☆☆
Phenanthrene catabolic process	3.5	★★☆☆☆
Sulfate reduction	3.3	★★☆☆☆
Ketone body catabolic process	3.2	★★☆☆☆
Naphthalene catabolic process	3.2	★★☆☆☆
Alkane catabolic process	3.1	★★☆☆☆
Benzoate catabolic process	3.0	★★☆☆☆
Denitrification pathway	2.8	★★☆☆☆
Sulfide ion homeostasis	2.6	★★☆☆☆
Ketone catabolic process	2.6	★★☆☆☆
Methanogenesis	1.8	★☆☆☆☆
Electron transport chain	1.0	★☆☆☆☆

Showing 1 to 14 of 14 entries

The PREGO knowledge-base is available at http://prego.hcmr.gr/.

Conclusions

on PREGO and its associations

- Similar number of molecular functions in all cases indicates the robustness of the main metabolic processes required for life
- The *Literature* provide us with a great number of high-quality associations, while the *Environmental Samples* one will gain more and more ground as omics' dataset keep increasing exponentially, retrieving associations that might not be described in the corresponding literature

Deciphering the functional potential of a hypersaline marsh microbial mat community

Aim of the study and contribution

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

Tristomo swamp in Karpathos

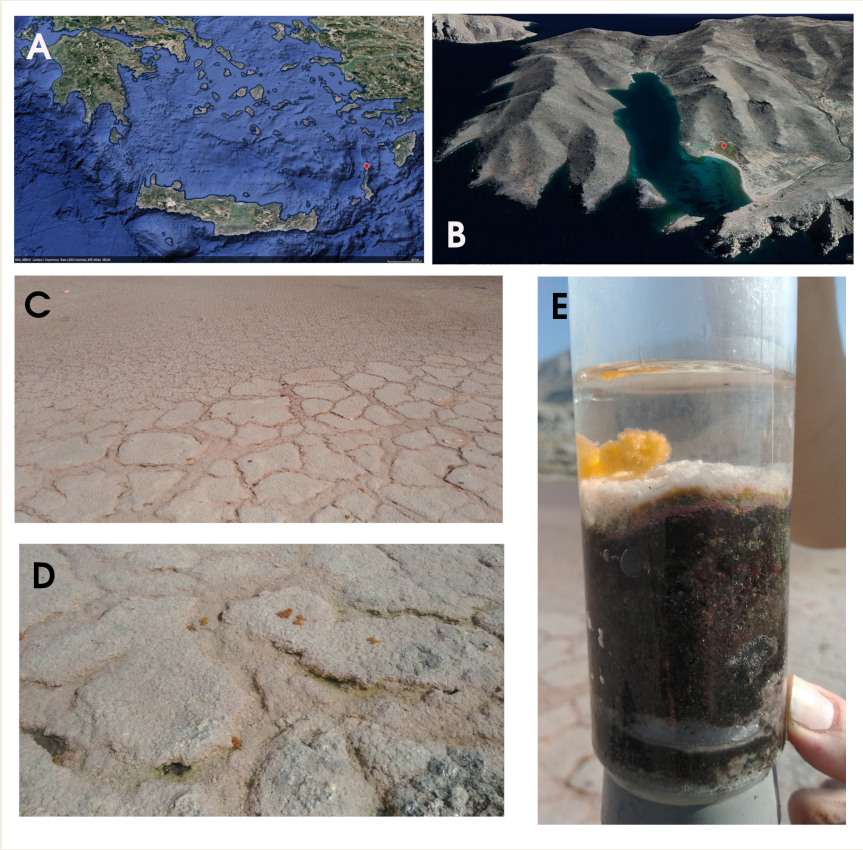
a seasonal brackish water marsh

Type of samples:

- from clearly observed mats, top - bottom layers
- if no clearly observed mats samples with no slicing
- aggregate samples

Sampling time points:

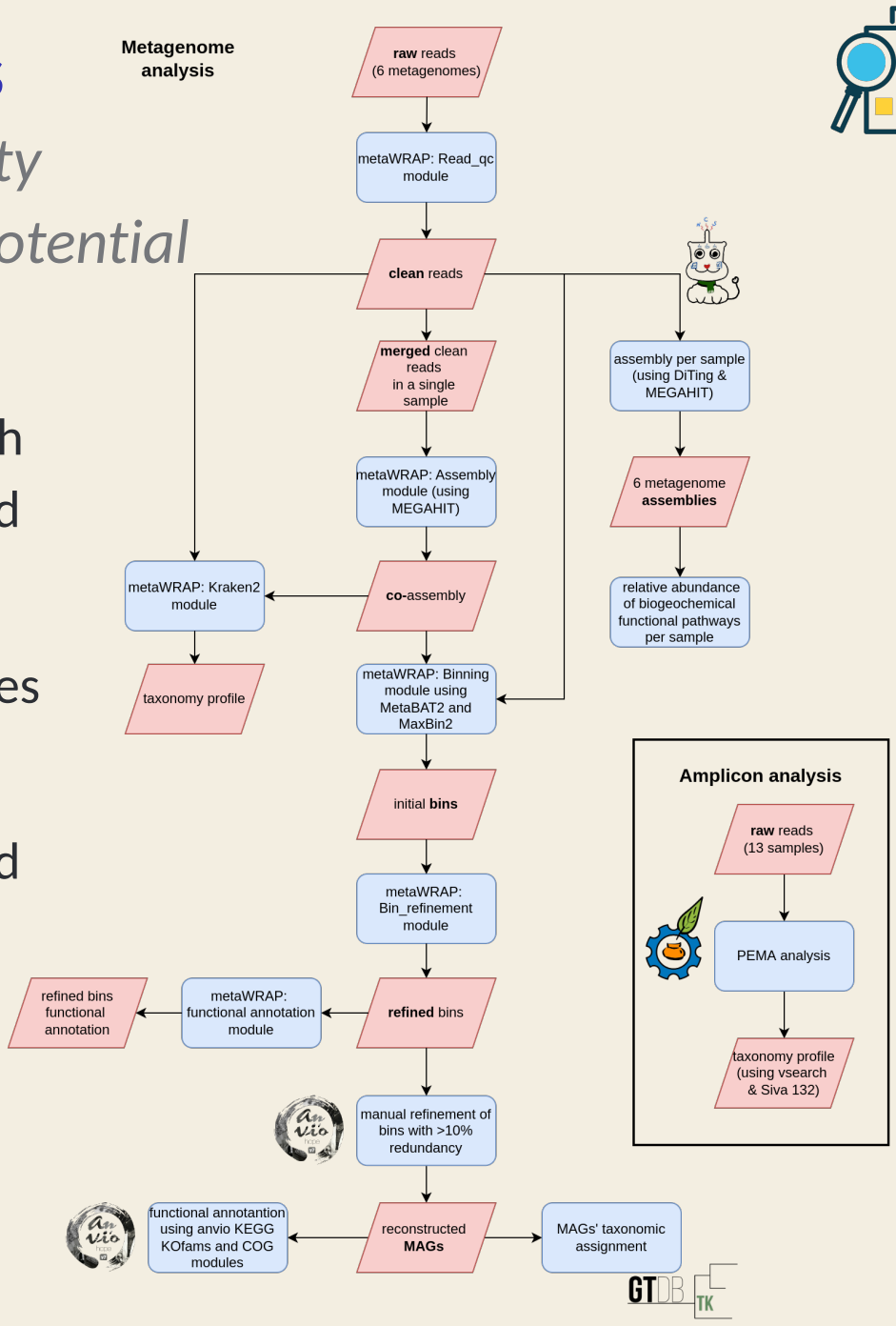
- July 2018
- November 2019



Bioinformatics analysis

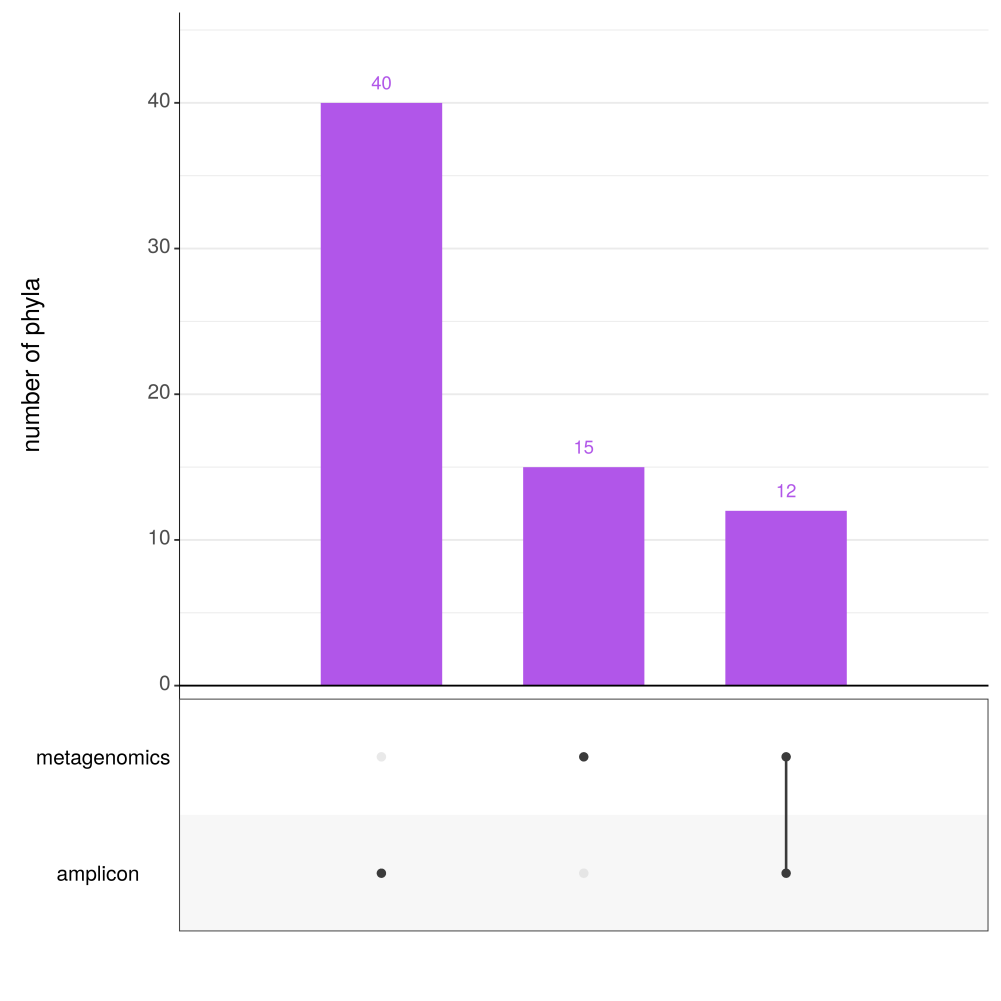
from raw data to community composition & functional potential

- metagenomic reads were both co - assembled and assembled at the sample level
- taxonomic & functional profiles per sample were retrieved
- MAGs were reconstructed and annotated



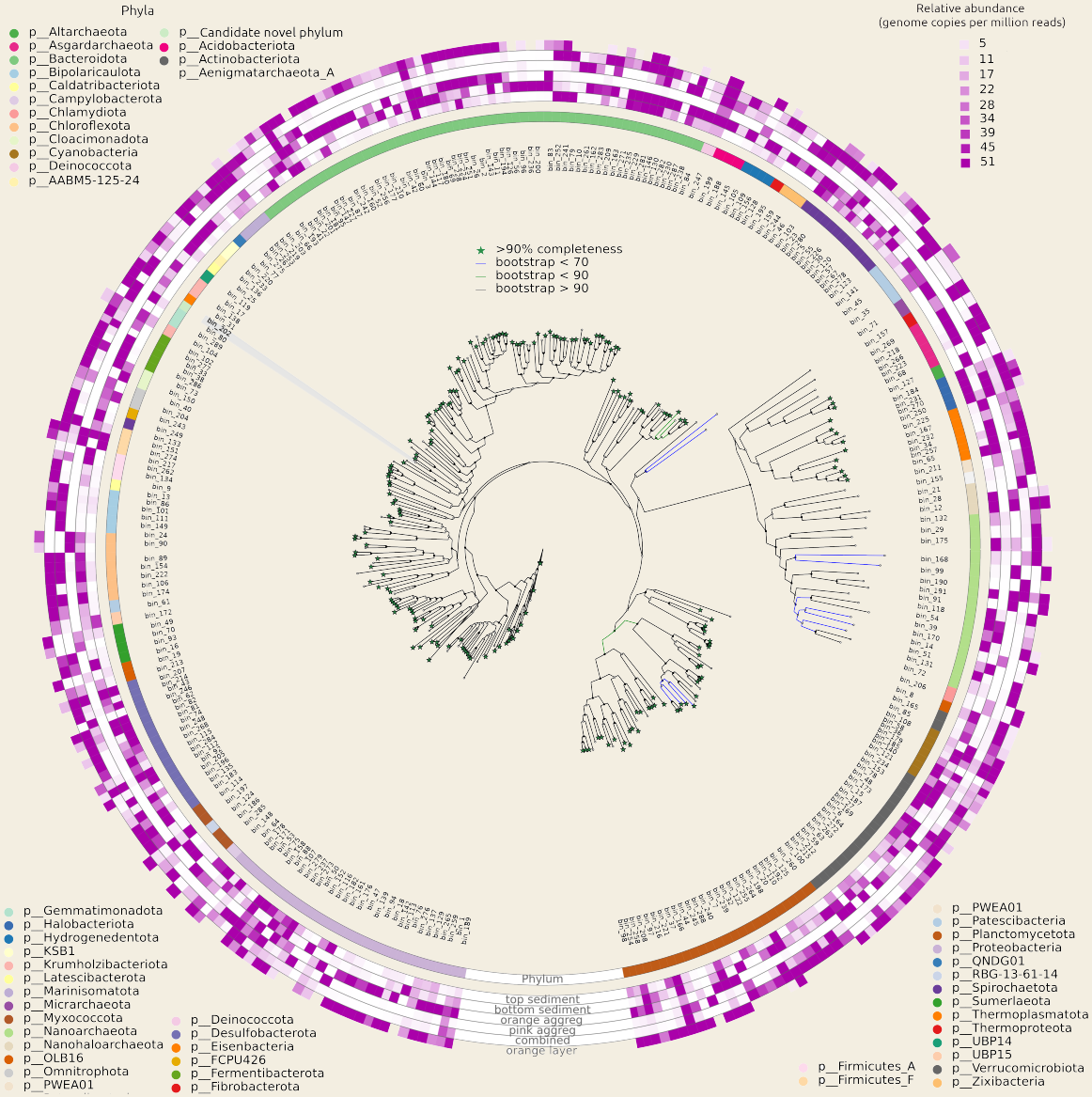
Amplicon vs metagenomics

number of phyla retrieved per method



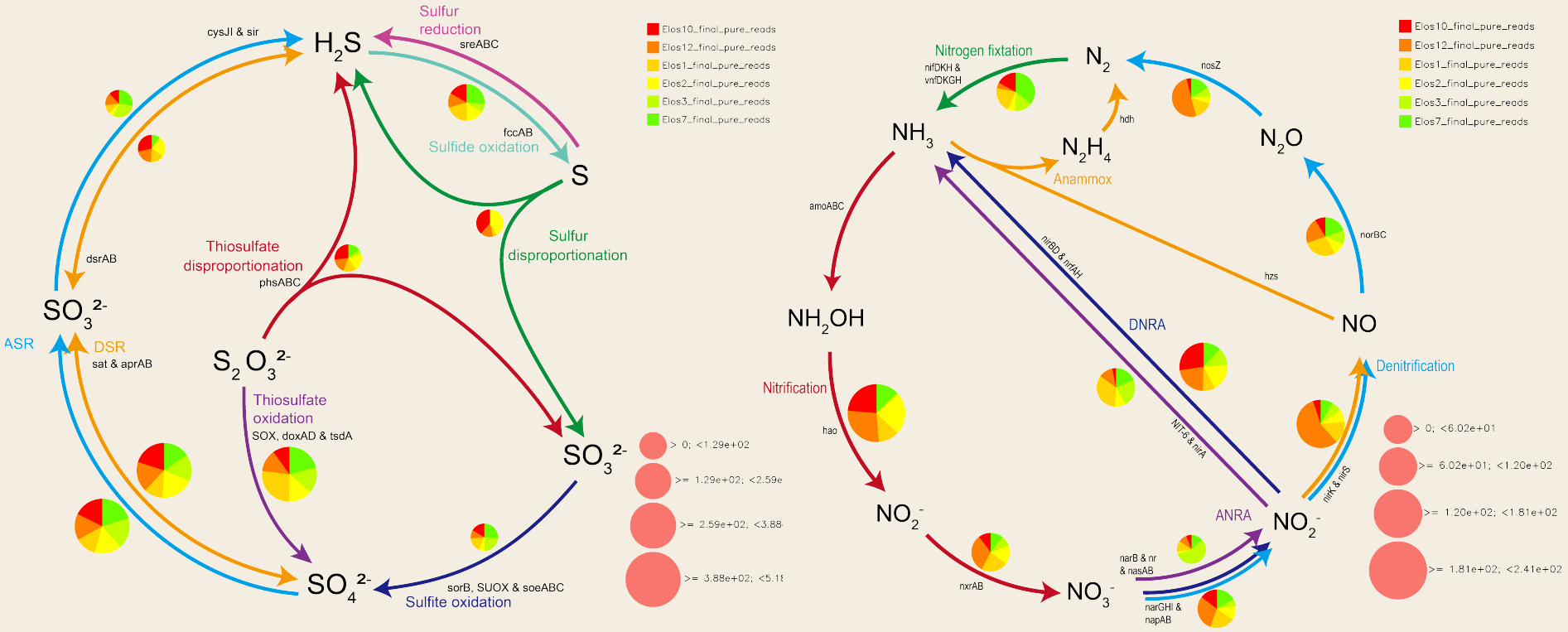
MAGs phylogeny

based on 25 single-copy genes



The S and the N cycle

using KEGG annotation terms



Conclusions

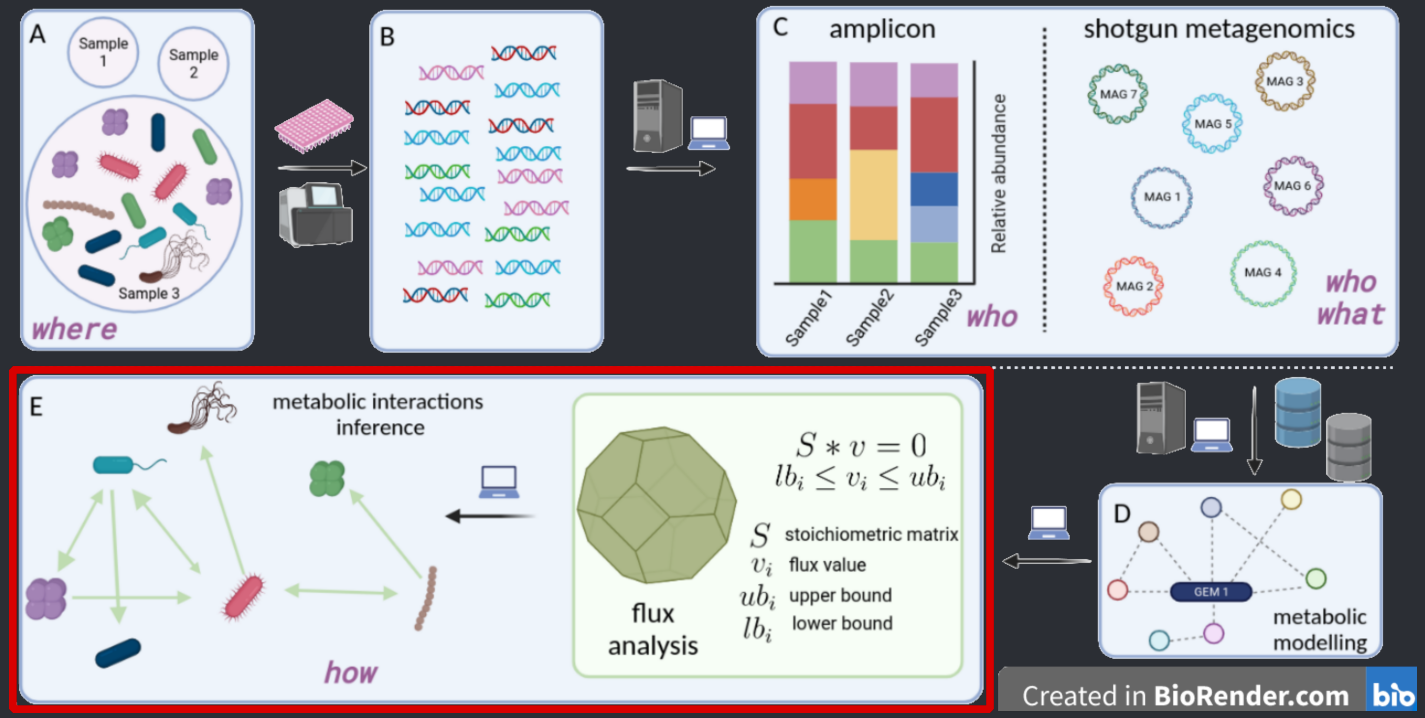
based on metagenomics analysis in the Tristomo marsh

- seasonality plays a key-role for the microbial mats under study to survive, ensuring oxygenic photosynthesis for a while
- anaplerotic reactions, that are abundant in our samples, may play an important role in replenishing the intermediates of the TCA cycle
- metabolic modelling can shed further light on the effects of the environmental challenges on the mat construction

General conclusions

- Bioinformatics approaches enhance microbial diversity assesment based on HTS data
- Containerization technologies and e-infrastructures provide the means for computational capacity and reproducibility
- High quality metadata enable efficient exploitation of sequencing data in a meta-analysis level
- Markov Chain Monte Carlo approaches enable flux sampling in high-dimensional polytopes
- Hypersaline mats host a great range of novel taxa & their functioning might be subject to anaplerotic reactions

Future perspectives

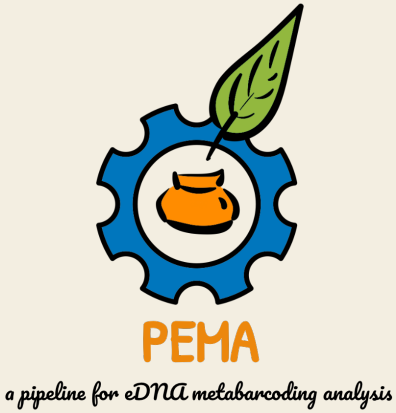


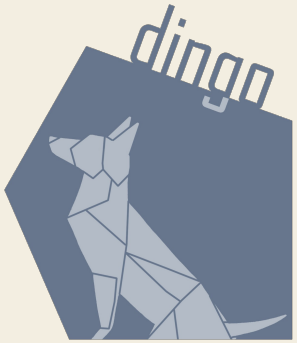
A (bit) more holistic framework

"a combination of quantitative high - throughput experiments and predictive metabolic models can elucidate the genotype - phenotype map of microbial metabolic strategies" - Bajic and Sanchez (2020)

Wrap-up

software tools





github.com/hariszaf/pema

github.com/hariszaf/darn

github.com/lab42open-team/ the `prego*` repositories

github.com/GeomScale/dingo

Publications

[1] **Zafeiropoulos, H.**, Paragkamian, S., Ninidakis, S., Pavlopoulos, G.A., Jensen, L.J. & Pafilis, E. (2022). PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types. Microorganisms 10(2), 293.

[2] **Zafeiropoulos, H.**, Gargan, L., Hintikka, S., Pavloudi, C. & Carlsson, J. (2021). The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data. Metabarcoding and Metagenomics, 5, e69657.

[3] Chalkis, A., Fisikopoulos, V., Tsigaridas, E. & **Zafeiropoulos, H.** (2021). Geometric algorithms for sampling the flux space of metabolic networks, 37th International Symposium on Computational Geometry (SoCG 2021).

[4] **Zafeiropoulos, H.**, Gioti, A., Ninidakis, S., Potirakis, A., Paragkamian, S., ... & Pafilis, E. (2021). 0s and 1s in marine molecular research: a regional HPC perspective. GigaScience, 10(8), giab053.

[5] Polymenakou, P.N., Nomikou, P., **Zafeiropoulos, H.**, ..., Kyrpides, N.C., Kotoulas, G. & Magoulas, A. (2021). The santorini volcanic complex as a valuable source of enzymes for bioenergy. Energies, 14(5), p.1414.

[6] **Zafeiropoulos, H.**, Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C. & Pafilis, E. (2020). PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. GigaScience, 9(3), giaa022.

[7] Pavloudi, C. & **Zafeiropoulos, H.** (2022) Deciphering the community structure and the functional potential of a hypersaline marsh microbial mat community ***(under review at FEMS Microbiology Ecology)***

[8] Garza, D.R., Gonze, D., **Zafeiropoulos, H.**, Liu, B. & Faust, K., (2022) Metabolic models of human gut microbiota: advances and challenges ***(under review at Cell systems)***

[9] Paragkamian, S., Sarafidou, G., ..., **Zafeiropoulos, H.**, Arvanitidis, C., Pafilis, E. & Gerovasileiou, V. Automating the curation process of historical literature on marine biodiversity using text mining: the DECO workflow ***(accepted in Frontiers in Marine Science)***

Acknowledgments

funding & grants

GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION

H.F.R.I.
Hellenic Foundation for Research & Innovation

Interreg
EUROPEAN UNION
Balkan-Mediterranean
RECONNECT

EOSC-Life

elixir

EMBO

DNAqua-Net
EU COST ACTION

Google
Summer of Code

Acknowledgments

people and more

My promotors:

Dr. Pafilis E.

Prof. Nikolaou Chr.

Prof. Ladoukakis

the rest of my 7-member committee:

Prof. Dina Lika

Prof. Panagiotis Sarris

Prof. Jens Carlsson

Prof. Karoline Faust

Special thanks to:

PhD Paragkamian S.

Dr. Gargan L.

Dr. Hintikka S.

Mr. Ninidakis St.

Mr. Potirakis Ant.

Dr. Chalkis A.

Dr. Fisikopoulos V.

Prof. Tsigaridas E.

My mojo:

Dr. Pavloudi C.

My corner:

Would not be here

if it was not with you.

GeomScale