

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

developing computational approaches to better understand microbial assemblages

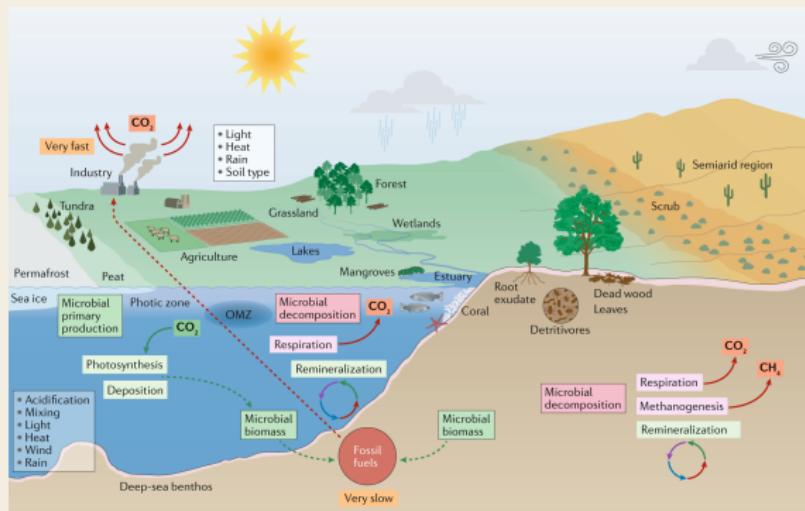
Haris Zafeiropoulos
PhD candidate



- 1. Chapter 1:** Introduction
- 2. Chapter 2:** Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
 - 2.1** PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & COI marker genes
 - 2.2** The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
- 3. Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
- 4. Chapter 4:** A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
- 5. Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
- 6. Chapter 6:** Os and 1s in marine molecular research: a regional HPC perspective
- 7. Chapter 7:** Conclusions

Microbial ecology & biogeochemical cycles

a corner-stone for life on earth



- composition
- functions
- interactions

→ power biogeochemical cycling

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

everyone is everywhere

Functional
potential
what

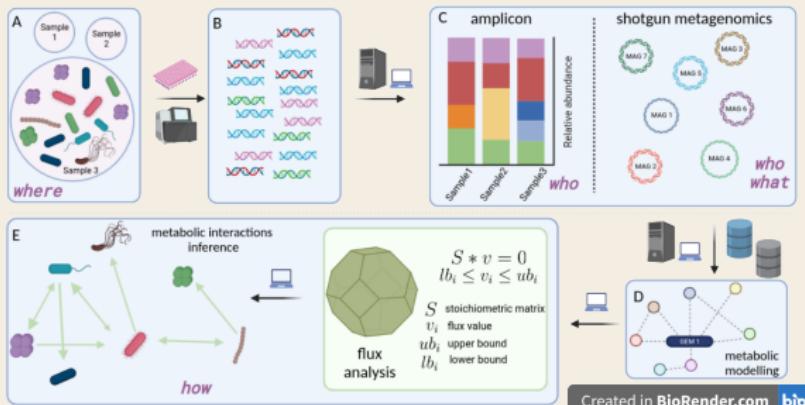
zero-sum game

Microbial
interactions
why / how

the entangled 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
- technology-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 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & COI marker genes
 - 2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4:** A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** Os and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

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. ...

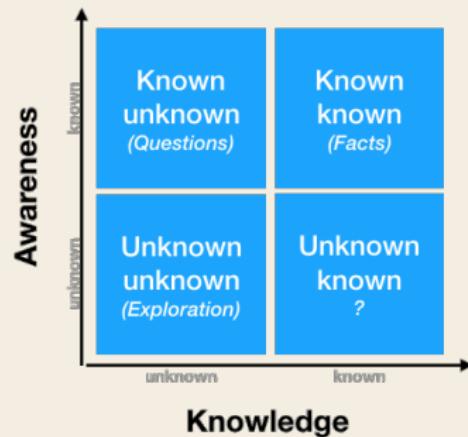
Methodology



- Sampling
 - Extraction
 - Bioinformatics
 - Biodiversity analysis

Bioinformatics challenges

for the analysis and the interpretation of amplicon data



PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & 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
- easy-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



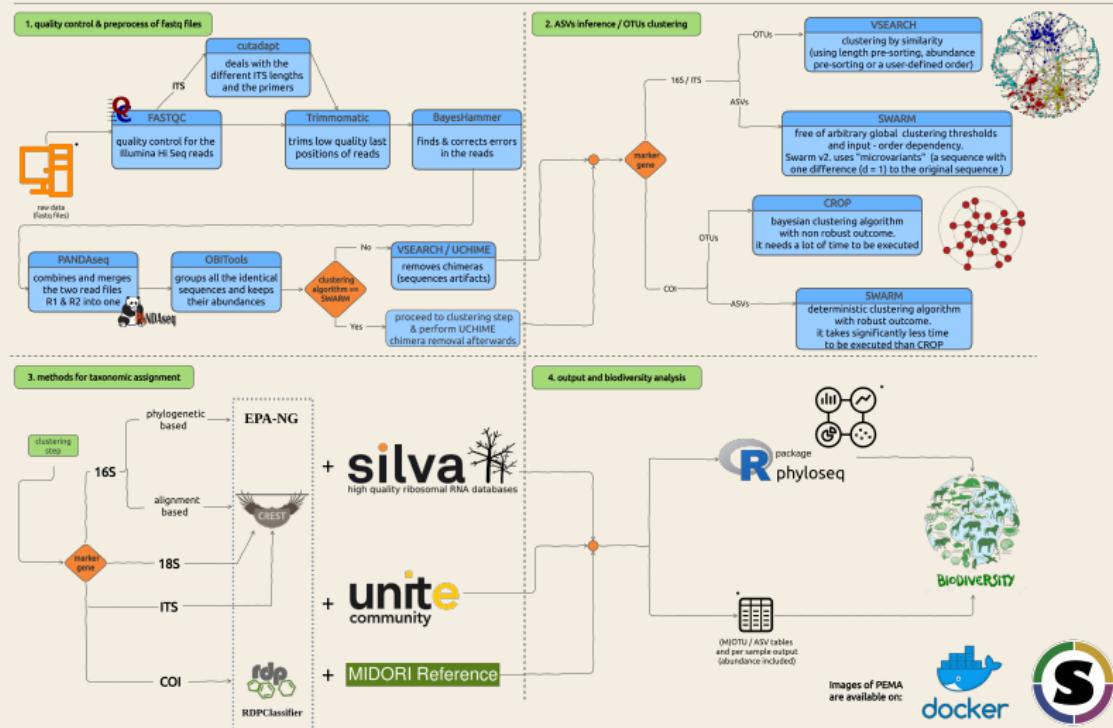
Containerization

High performance
computing

Results: PEMA features

an overview

PEMA in a nutshell



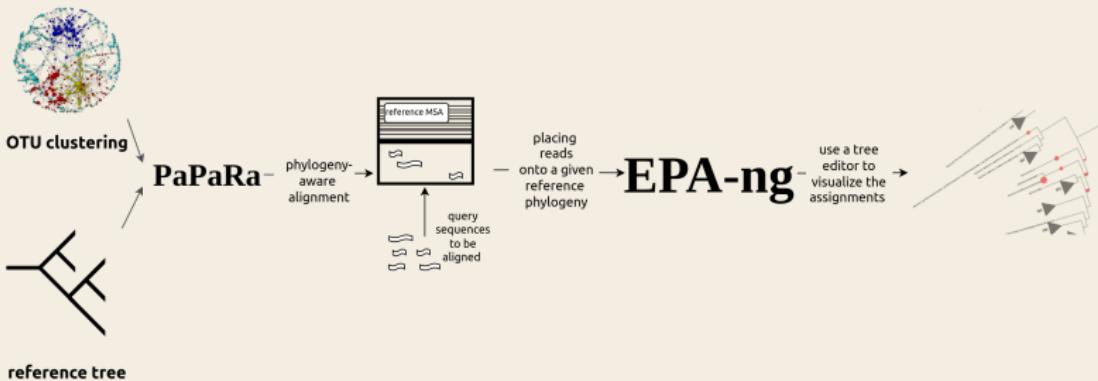
Results: PEMA features

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

A. create reference tree



B. phylogeny-based taxonomy assignment



Results: tuning effects

in mock communities

Mock communities using the 16S rRNA gene and multiple parameter sets
 (identification at the genus level):

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 no singletons)	Swarm (d = 25 strictness = 0.8)	Swarm (d = 30 strictness = 0.6)	Swarm (d = 30 strictness = 0.8)
TP	12	15	18	18	15	17	17	17	17	17
FP	2	2	21	11	6	5	5	4	5	5
FN	8	5	2	2	5	3	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	0.77
REC (TP / TP+FN)	0.6	0.75	0.9	0.9	0.75	0.85	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	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

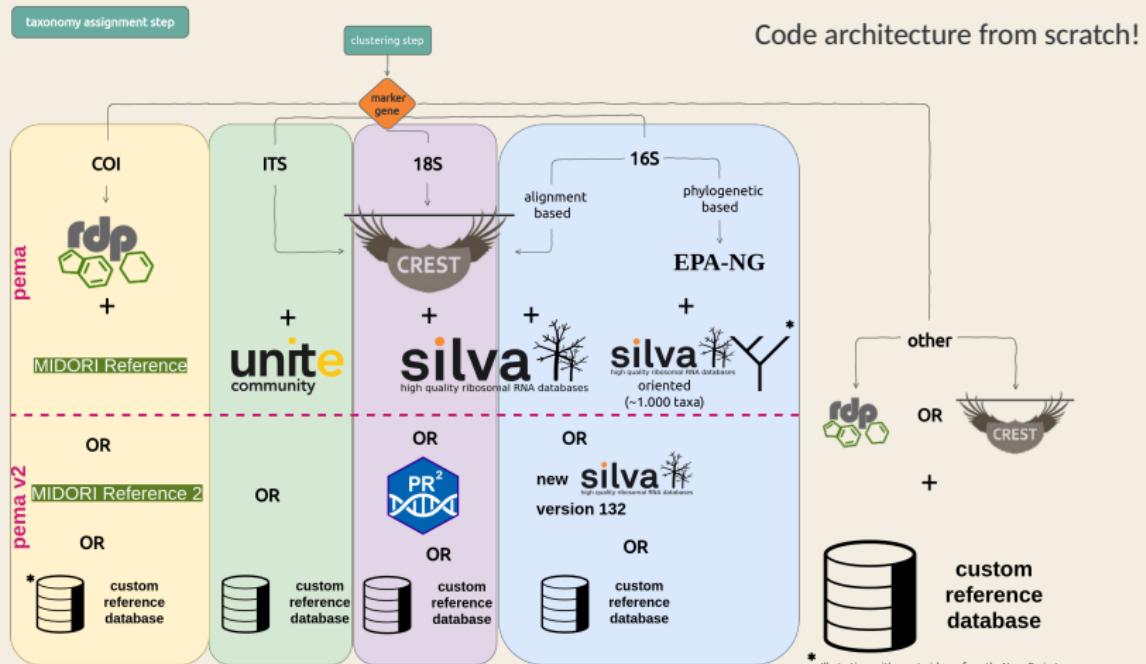
in real-world data

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

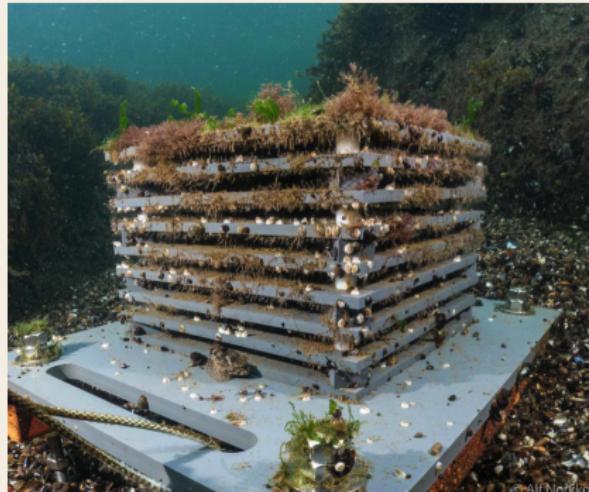
PEMA v.2

addressing some of the challenges



* Illustrations with an asterisk are from the Noun Project

Latest PEMa version *addressing the challenges of the community*



ASSEMBLE 
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

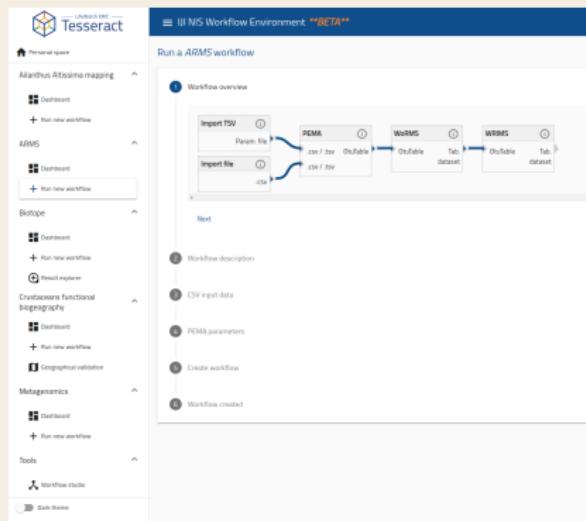
MBON
Marine Biodiversity
Observation Network

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
3. Combine with other tools

Conclusions

on PEMA and eDNA metabarcoding

- PEMA is accurate, execution-friendly and fast pipeline
- tuning is essential in metabarcoding analyses
- sequencing a mock community along with your samples can be of great help in parameter tuning
- computing resources required range; infrastructures may 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

To build a framework for extracting non-target, potentially unassigned (or assigned with low confidence) sequences from COI environmental sequence samples, hereafter referred to as “dark matter” as per Bernard et al. (2018).

We argue that the vast majority of these sequences represent microbial taxa, such as bacteria and archaea

Methodology / Implementation

Dark mAtteR iNvesigator

In COI amplicon studies,
a great number of OTUs/ASVs retrieved
either have no hits or
their hit has a low confidence

DARN estimates to what extent
the OTUs/ASVs retrieved in an
environmental sample represent
target taxa or not

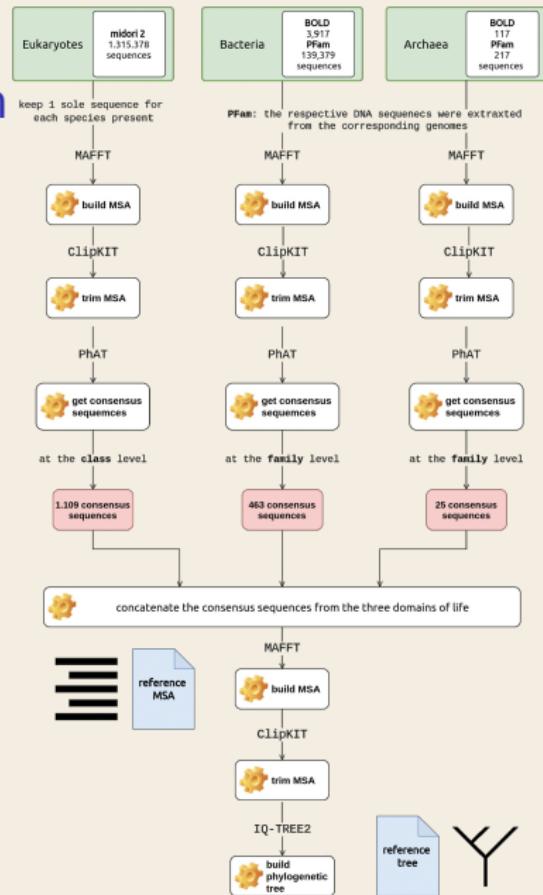


Figure from: [10.3897/mbmg.5.69657](https://doi.org/10.3897/mbmg.5.69657) 21/68

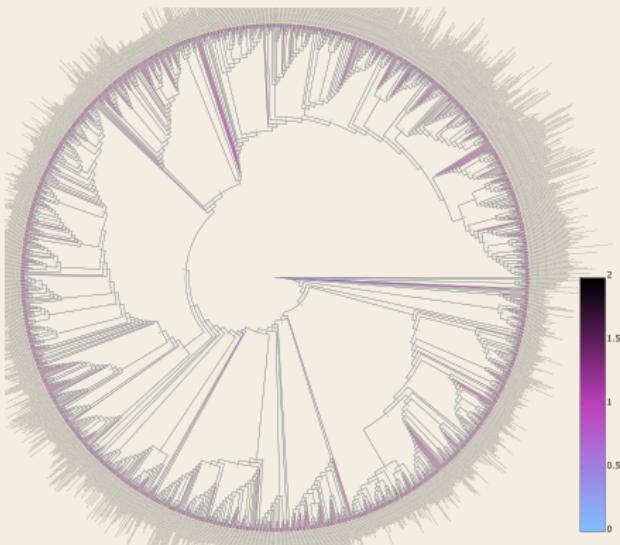
Methodology / Implementation

sequences retrieved

Resources	bacteria		archaea	
	# of sequences	# of strains	# of sequences	# of strains
BOLD	3,917	2,267	117	117
PFam-oriented	9,154	4,532	217	115
Total unique entries	11,421	6,798	334	201

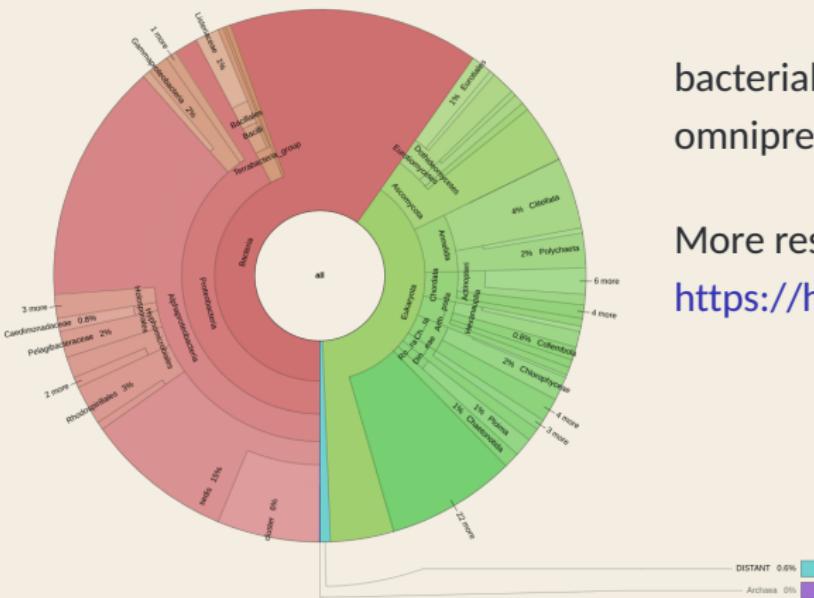
Methodology / Implementation

Phylogenetic tree of the COI consensus sequences retrieved



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

Results: DARN using real-world data with multiple sample types, primers, PCR protocols and bioinformatics 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

1. **Chapter 1:** Introduction
2. **Chapter 2:** Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
 - 2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & COI marker genes
 - 2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4:** A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** Os and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

Chapter 3: PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types

Aim of the study and contribution

To build a hypothesis generation resource based on associations between:

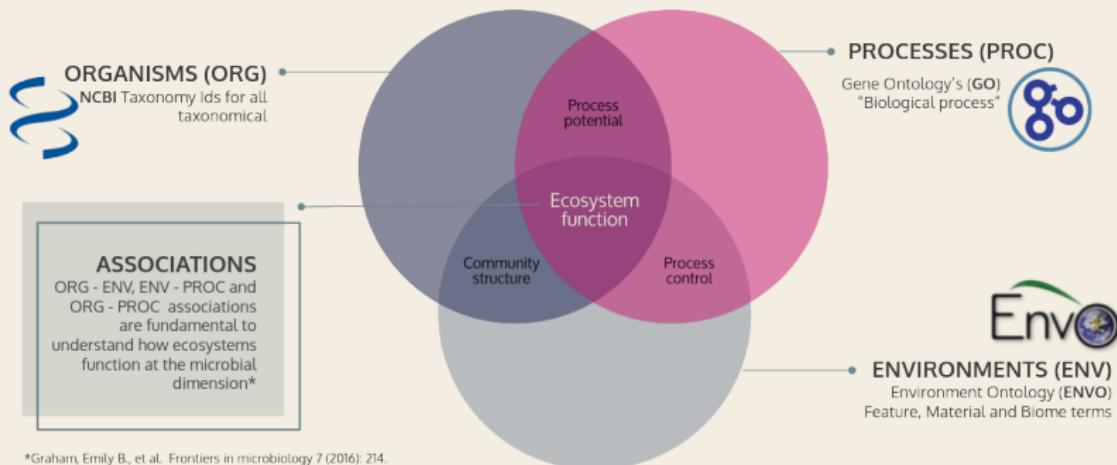
- *organisms* and the *environments* they inhabit
- *organisms* and the *biological processes* they are involved with
- *processes* and the *environments* where they occur

To this end, associations among such terms were exported from:

- the publicaly available literature
- genome and omics' studies, their results and their corresponding metadata

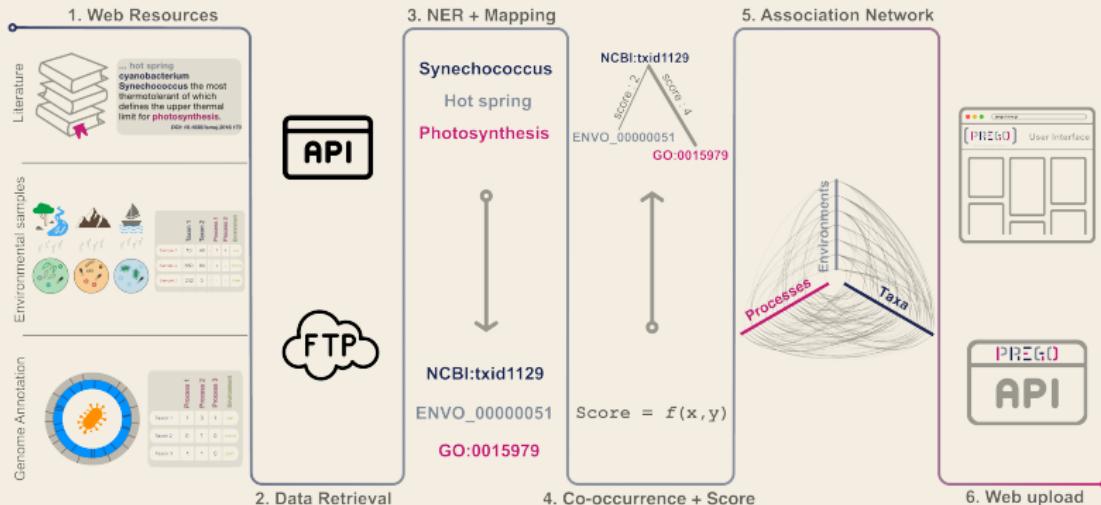
PREGO and its referring terms

the fundamental role of ontologies



Methods / Implementation

3 channels of information - 1 framework



Named Entity Recognition & the literature channel

exporting associations from publicly available publications

Identification of potentially important pathways missing from the model

EXTRACT	x
Protein	
Chemical compound	
Organism	
Environment	
Tissue	
Disease/phenotype	
Gene Ontology term	

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 hydrogentivorans* [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:	metres-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	mg485562	Cincinnati	USA	39.11, -84.5		
mgm4702469.3	06052015a_S3_L001_R2_001	Large lake biome	lake	water	mgs485566	mg485568	Cincinnati	USA	39.11, -84.5		
mgm4702471.3	06032015a_S1_L001_R2_001	Large lake biome	lake	water	mgs485554	mg485556	Cincinnati	USA	39.11, -84.5		

MG-RAST
metagenomics analysis server

Co-mentioning and scoring scheme

which are the most worthy and relevant associations

- genome annotation oriented associations: fixed scores
- associations in the *Environmental Samples* channel are scored based on the number of samples they co-occur.
- similarly, in the *Literature* channel, based on the number of publications

		Y = y	
		Yes	No
X = x	Yes	$c_{x,y}$	$c_{x,0}$
	No	$c_{0,y}$	$c_{0,0}$
	Total	$c_{.,y}$	$c_{.,0}$

Environmental samples score:

$$\text{score}_{x,y} = 2.0 * \sqrt{\frac{c_{x,y}}{c_{.,y}}}^a \quad (1)$$

Associations between entities of PREGO

after metadata retrieval and co-occurrence analysis

Channel	Source	Environments		Taxa		Taxa	
		- Processes	- Functions	Taxonomy	- Environments	- Processes	- Function
Literature	MEDLINE			Strains	69,968	590,630	384,079
	PubMed -	883,997	422,579	Species	778,877	3,501,635	1,961,920
	PMC OA			Total	1,669,608	7,969,310	4,613,827
	MG-RAST amplicon			Strains	13,645	-	-
		-	-	Species	39,007	-	-
				Total	53,439		
Environmental samples	MG-RAST metagenome			Strains	262,106		8,626,328
		-	620,846	Species	103,913	-	10,715,548
				Total	372,301		19,950,096
	MGnify amplicon			Strains	18	-	-
		-	-	Species	30,122	351	-
				Total	111,976	2,097	
Annotated Genomes and Isolates	JGI IMG isolates			Strains	8,229		3,461,693
		-	-	Species	42,141	-	13,216,559
				Total	50,888		16,821,850
	STRUO			Strains	-	-	1,803
		-	-	Species	-	-	4,070,195
				Total	-	-	4,079,312
	BioProject			Strains	3,263	7,473	
		-	-	Species	4,187	4,294	
				Total	7,641	12,169	
	Total			Strains	357,229	598,103	12,473,903
		All	883,997	Species	998,247	3,506,280	29,964,222
				Total	2,265,853	7,983,576	45,465,085

A real case hypothesis generation scenario

Posidonia and its microbiome



What about *Posidonia* ?

Literature suggests that

Planctomycetes

and especially *Blastopirellula*

and *Rhodopirellula*

are commonly

found in its microbiome.

Why so ?

Let us have a look [here!](#)

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 number of environmental types that have been associated with members of each phylum varies, as a phylum may be universally present, while others could be strongly niche-specific
- 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

1. **Chapter 1:** Introduction
2. **Chapter 2:** Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
 - 2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & COI marker genes
 - 2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4:** A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** Os and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

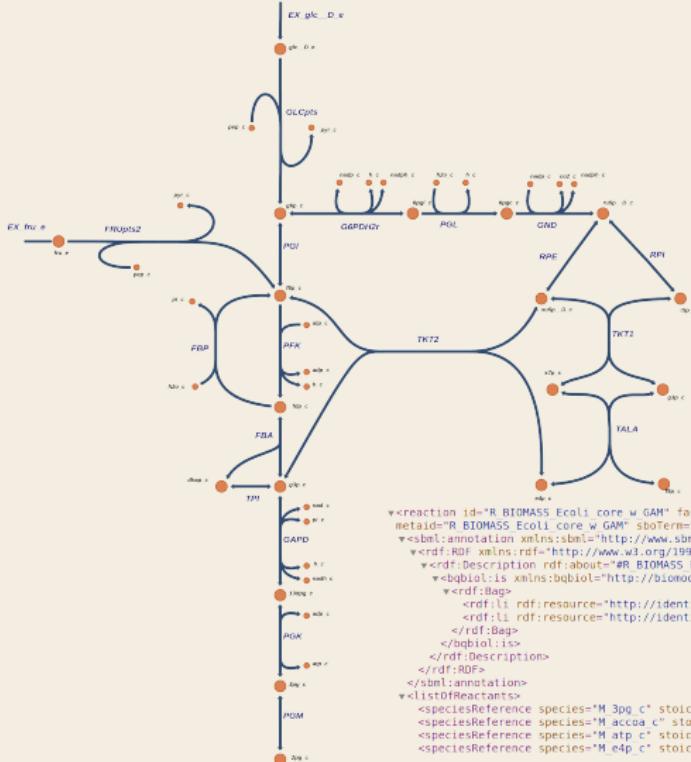
Chapter 4: A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks

Aim of the study and contribution

Flux sampling is a computationally intensive task, especially as the dimension of the polytopes derived from the metabolic model under study increases.

To allow flux sampling at high dimensional polytopes, such as those of multispecies communities or host-microbe cases, we introduce a Multi-phase Monte Carlo Sampling (MMCS) algorithm.

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

Genome-scale metabolic reconstruction approaches, pros and cons

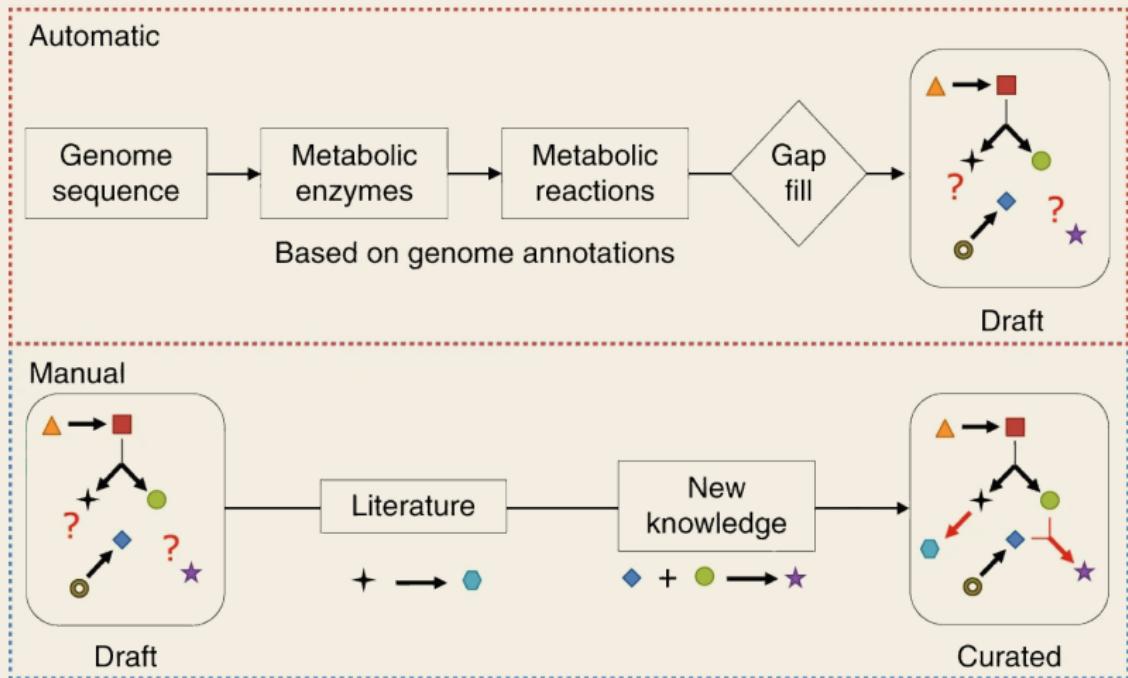


Figure from: Heirendt et al. Nature protocols 14.3 (2019): 639-702.

From a stoichiometric matrix

to a constraint-based model

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

Reactions

	R ₁	R ₂	R ₃	R ₄	R ₅
Metabolites	-1	0	0	0	0
▲	1	-1	0	0	0
■	0	1	-1	0	0
●	0	1	0	0	-1
◆	0	0	1	0	0
○	0	0	0	-1	0
◆	0	0	0	1	-1
★	0	0	0	0	1

S-matrix

$\times \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

Flux vector

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**.

From concentrations to fluxes to study changing environments

We can describe the mass balance of a chemical compound as the difference between the sum of the fluxes of all the reactions that form it and the sum of all that degrade it.

$$\frac{d\omega_i}{dt} = \sum_k s_{ik} v_k = \langle s_i, v \rangle$$

and thus:

$$\frac{d\omega}{dt} = S v$$

The region of steady states

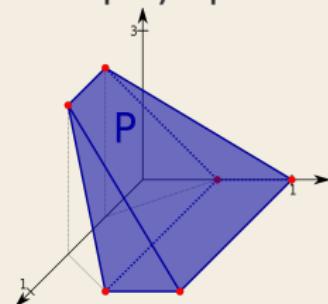
moving to full dimensional polytope

The *constraints* on the reactions fluxes.

$$Sv = 0, \quad (2)$$
$$v_{lb} \leq v \leq v_{ub}$$

$$S \in \mathbb{R}^{m \times n}, v \in \mathbb{R}^n$$

As a *full dimensional polytope*



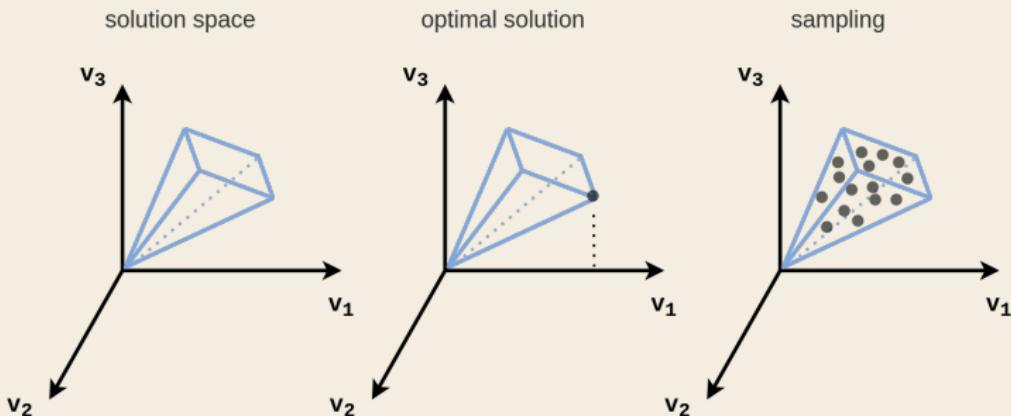
$$P := \{x \in \mathbb{R}^d | Ax \leq b\}$$

$N \in \mathbb{R}^{n \times d}$ denotes the matrix of the null space of S , i.e. $SN = 0_{m \times d}$.

By replacing v with Nx in Equation 2, we get the full dimensional polytope P , where
 $A = \begin{pmatrix} I_n N \\ -I_n N \end{pmatrix}$ and $b = \begin{pmatrix} v_{ub} \\ v_{lb} \end{pmatrix} N$, (in \mathbb{R}^d).

Flux sampling

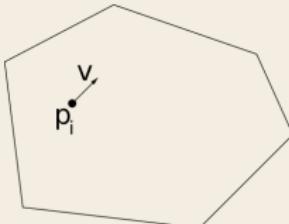
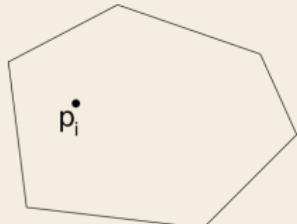
an alternative approach



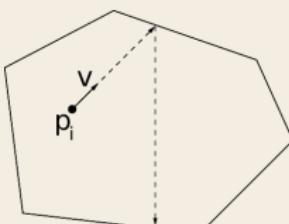
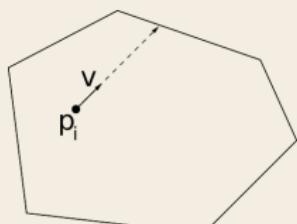
- enables the analysis of GEMs without the need of an objective function
- determines the feasible solution spaces for fluxes in a network based on a set of conditions as well as the probability of obtaining a solution

Billiard walk

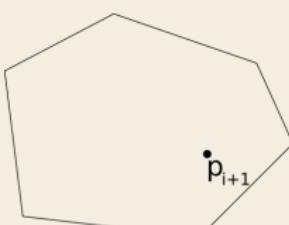
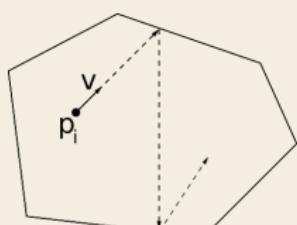
for random sampling



Generate the length of the trajectory $L \sim D$.



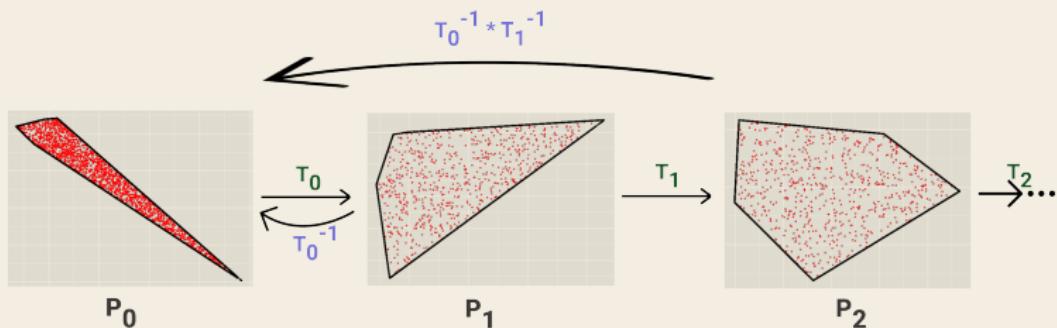
Pick a uniform direction v to define the trajectory.



The trajectory reflects on the boundary if necessary.

Return the end of the trajectory as $p_i + 1$.

Our Markov Chain Monte Carlo (MCMC) algorithm for flux sampling

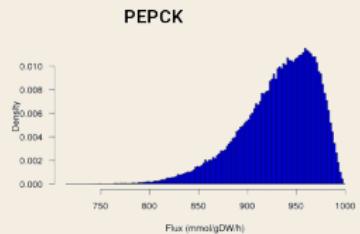
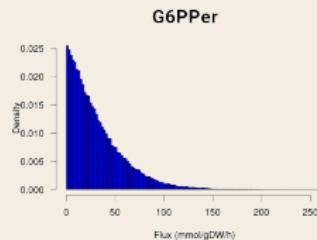
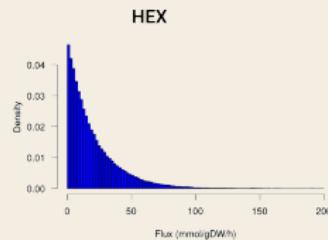


Steps of an MMCS phase

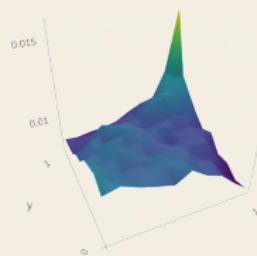
- **sampling step:** using a variant of the **Billiard walk**
- **rounding step:** calculate a linear transformation T_i that puts the sample into isotropic position and then apply it on P_i to obtain the polytope of the next phase
- check several statistic tests

Flux sampling output

marginal distributions and copulas



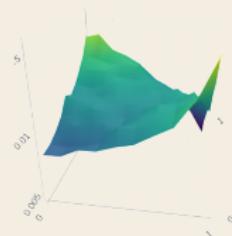
HEX - G6PPer



HEX:
 $\text{atp_c} + \text{glc_D_c} \rightarrow \text{adp_c} + \text{g6p_c} + \text{h_c}$

G6PPer:
 $\text{g6p_r} + \text{h2o_r} \rightarrow \text{pi_r} + \text{glc_D_r}$

HEX - PEPCK



PEPCK:
 $\text{gtp_c} + \text{oaa_c} \rightarrow \text{co2_c} + \text{gdp_c} + \text{pep_c}$

Result / experiments

sampling the largest single-species metabolic networks

model	m	n	d	MMCS		cobra	
				Time (sec)	N	Time (sec)	N
e_coli_core	72	95	24	6.50e-01	3.40e+03 (8)	7.20e+01	4.61e+06
iLJ478	570	652	59	9.00e+00	5.40e+03 (5)	4.54e+02	2.79e+07
iSB619	655	743	83	1.70e+01	8.20e+03 (5)	9.56e+02	5.51e+07
iHN637	698	785	88	2.00e+01	6.80e+03 (4)	1.03e+03	6.19e+07
iJN678	795	863	91	2.50e+01	8.10e+03 (4)	1.17e+03	6.62e+07
iNF517	650	754	92	1.70e+01	6.20e+03 (4)	1.33e+03	6.77e+07
iJN746	907	1054	116	5.70e+01	8.70e+03 (3)	2.22e+03	1.07e+08
iAB_RBC_283	342	469	130	5.20e+01	1.07e+04 (5)	7.85e+03	4.05e+08
iJR904	761	1075	227	2.98e+02	1.62e+04 (4)	8.81e+03	4.12e+08
iAT_PLT_636	738	1008	289	3.25e+02	1.04e+04 (2)	1.73e+04	6.68e+08
iSDY_1059	1888	2539	509	2.813e+03	2.31e+04 (3)	6.66e+04	2.07e+09
iAF1260	1668	2382	516	6.84e+03	5.33e+04 (6)	7.04e+04	2.13e+09
iEC1344_C	1934	2726	578	4.86e+03	3.95e+04 (4)	9.42e+04	2.67e+09
iJO1366	1805	2583	582	6.02e+03	5.14e+04 (5)	9.99e+04	2.71e+09
iBWG_1329	1949	2741	609	3.06e+03	4.22e+04 (4)	1.05e+05	2.97e+09
iML1515	1877	2712	633	4.65e+03	5.65e+04 (5)	1.15e+05	3.21e+09
Recon1	2766	3741	931	8.09e+03	1.94e+04 (2)	3.20e+05	6.93e+09
Recon2D	5063	7440	2430	2.48e+04	5.44e+04 (2)	~ 140 days	1.57e+11
Recon3D	8399	13543	5335	1.03e+05	1.44e+05 (2)	-	-

Conclusions

on sampling the flux space of metabolic models

- besides the computational, several challenges from the biological point of view
- essential insight (knock-out genes, host-microbe interactions etc)
- Recon3D includes 13, 543 reactions ($d = 5, 335$); sampling the flux space of metabolic models integrating several microbial GEMs is now possible

1. **Chapter 1:** Introduction
2. **Chapter 2:** Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
 - 2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & COI marker genes
 - 2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
3. **Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
4. **Chapter 4:** A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
5. **Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
6. **Chapter 6:** Os and 1s in marine molecular research: a regional HPC perspective
7. **Chapter 7: Conclusions**

Chapter 5: 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

Both amplicon and metagenome analysis was conducted to investigate the composition and the functional potential of microbial communities from the Tristomo marsh (Karpathos island, Greece)

Tristomo swamp in Karpathos

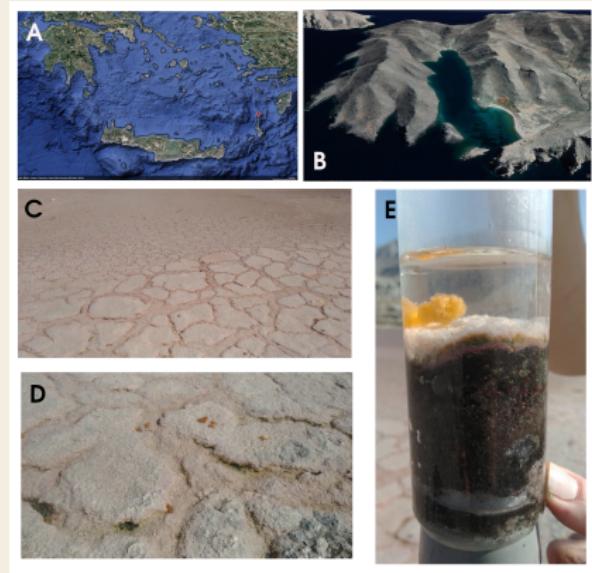
a seasonal brackish water marsh formed at the edge of a small plain

Type of samples:

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

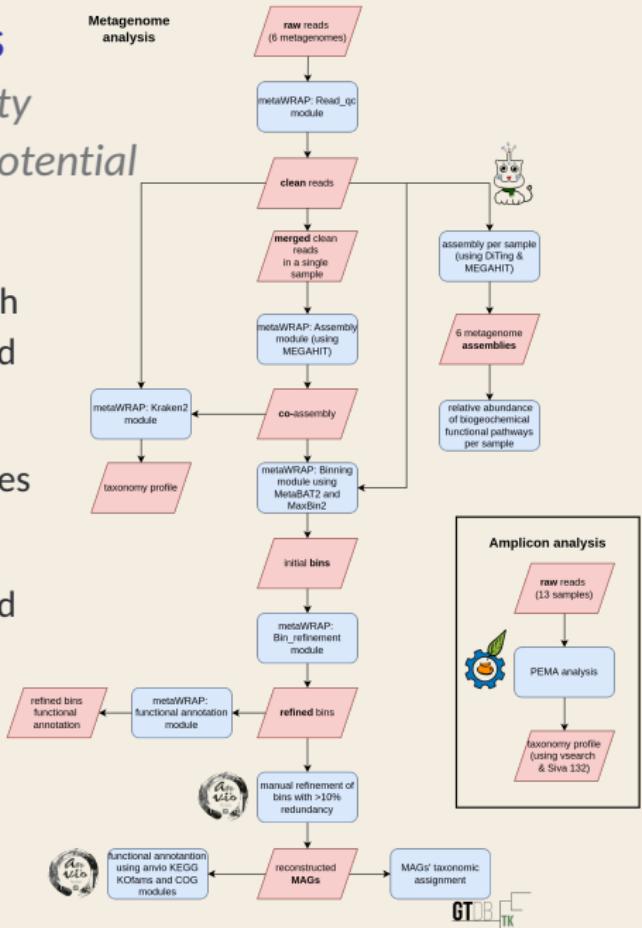
Sampling time points:

- July 2018
- November 2019



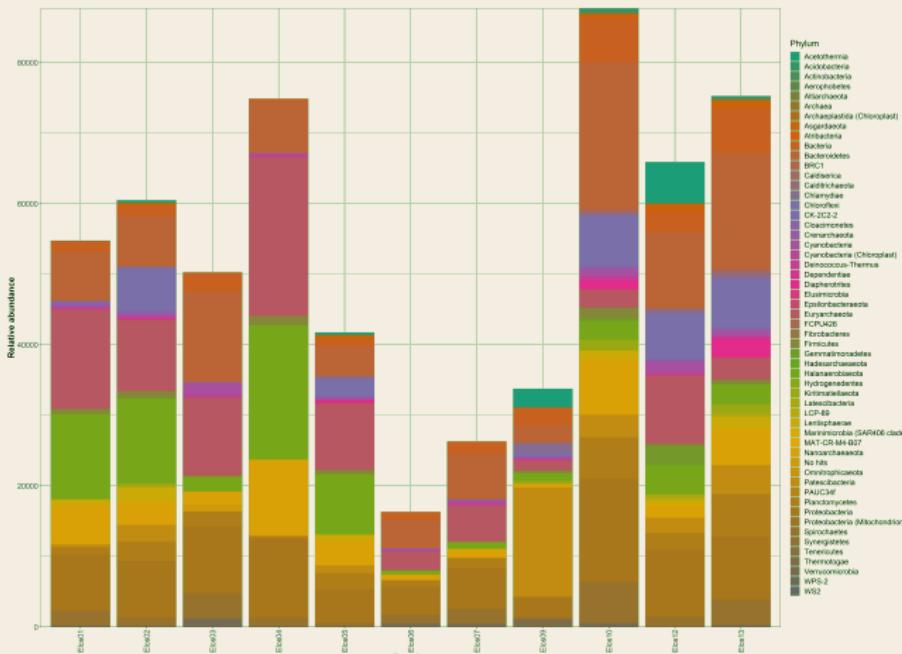
Bioinformatics analysis from raw data to community composition & functional potential

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



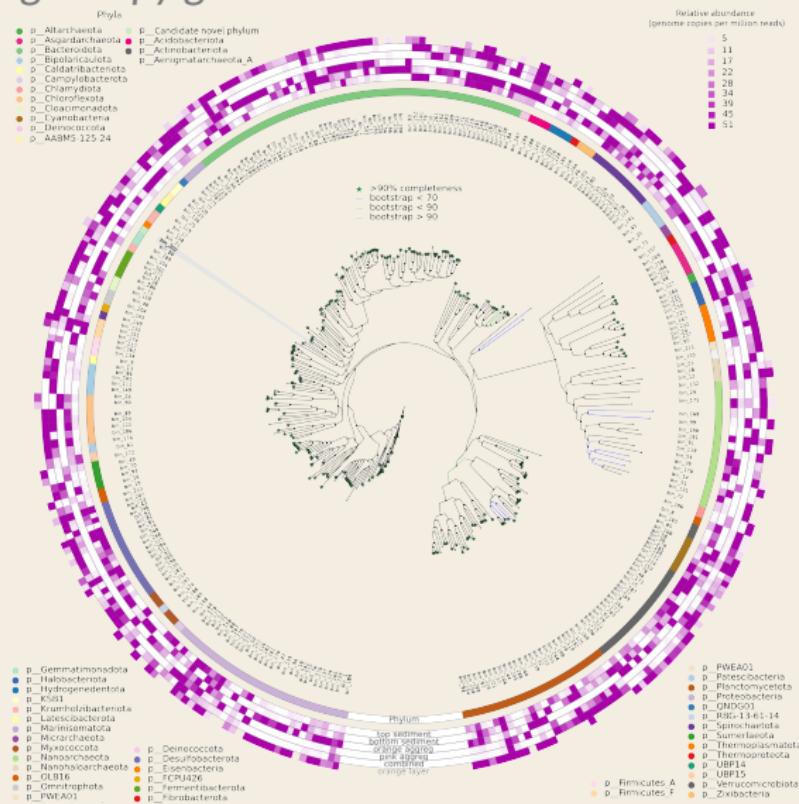
Abundances of the main microbial taxa, at the phylum level

based on 16S rRNA amplicon data

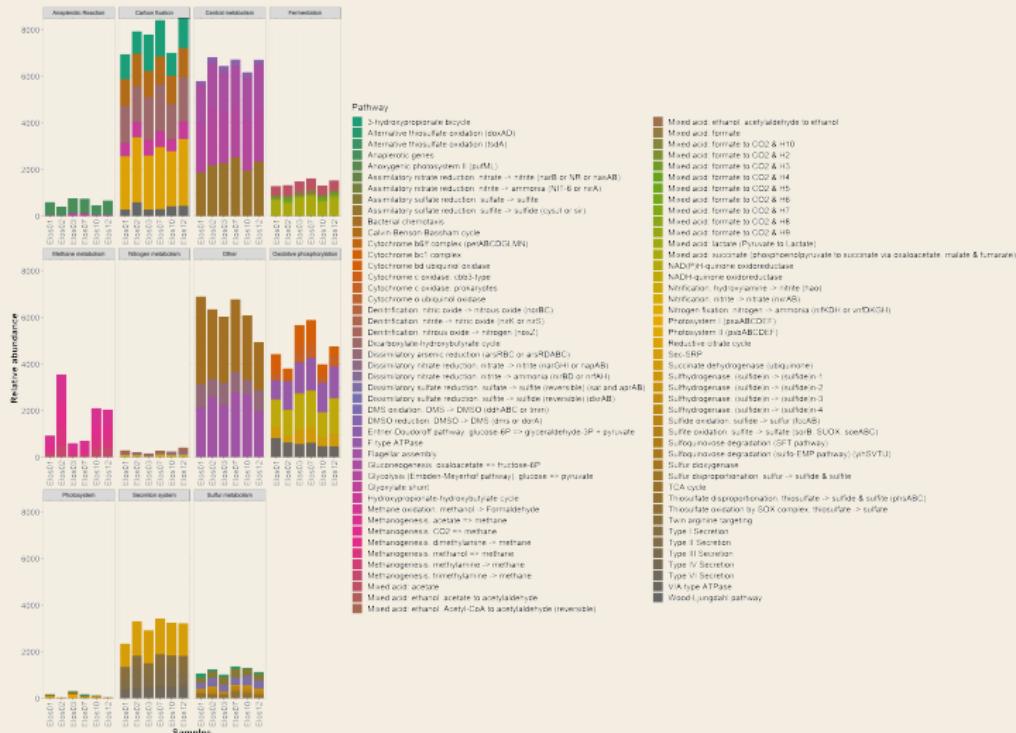


MAGs phylogeny

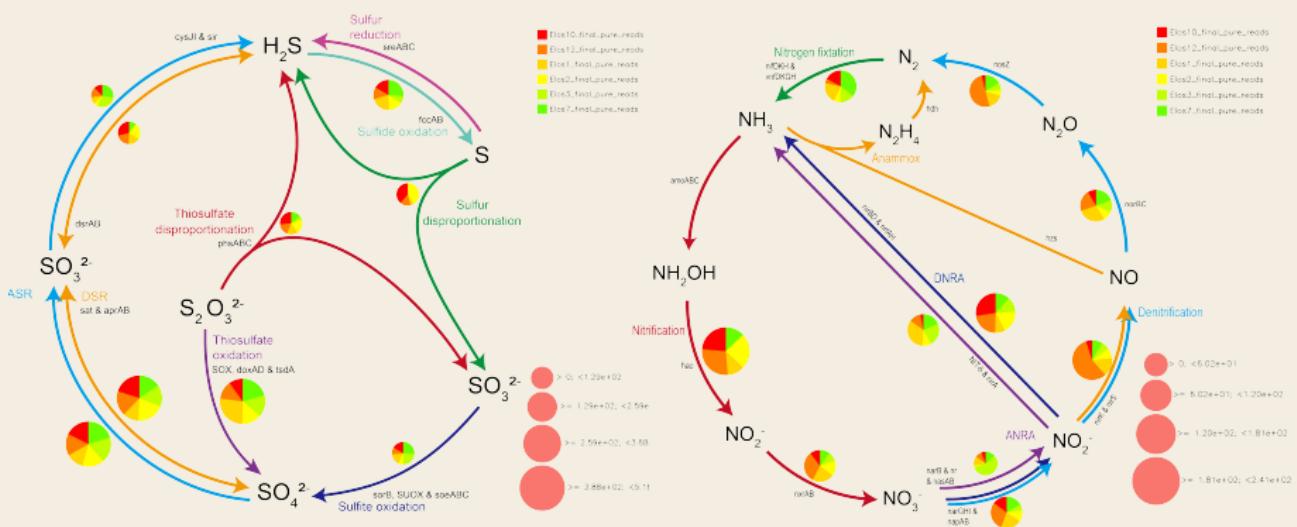
based on 25 single-copy genes



Metabolic pathways per biogeochemical cycle and their relative abundance at each sample



The S and the N cycle using KEGG annotation terms



Conclusions

and future work on hypersaline microbial mats

- temporal change and seasonal development of the microbial hypersaline mats under study - during winter months, both the salt crust and the layering of the microbial mat disappears - is necessary for the survival of the microorganisms by 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

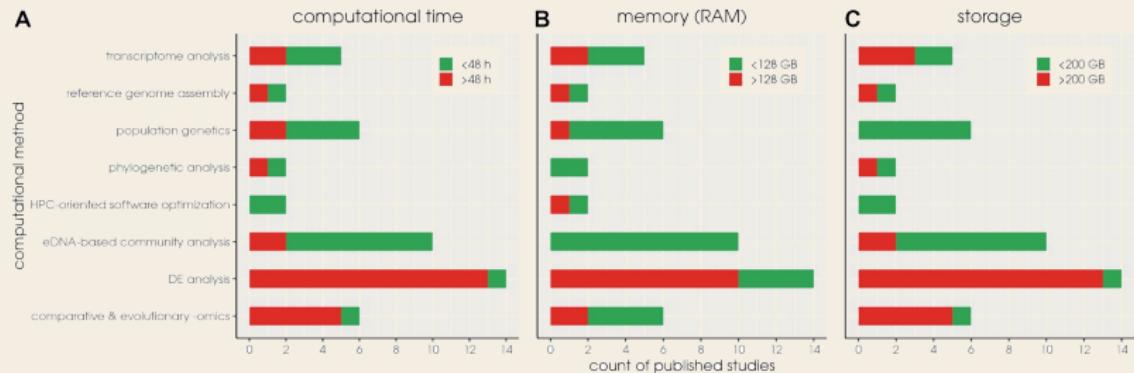
- 1. Chapter 1:** Introduction
- 2. Chapter 2:** Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
 - 2.1** PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & COI marker genes
 - 2.2** The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
- 3. Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
- 4. Chapter 4:** A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
- 5. Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
- 6. Chapter 6:** Os and 1s in marine molecular research: a regional HPC perspective
- 7. Chapter 7: Conclusions**

Chapter 6: Os 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

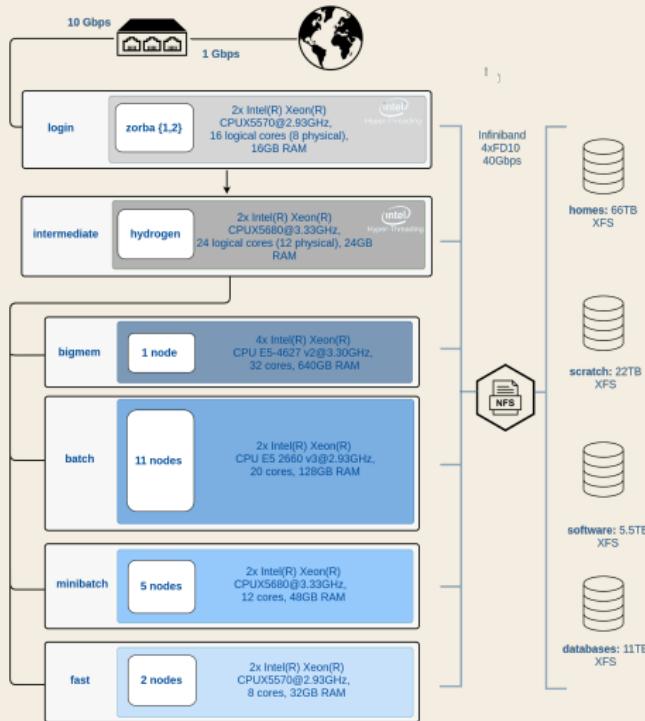
Computational requirements for trivial bioinformatic tasks



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

Zorbas: the HPC facility of IMBBC

a Tier 2 (regional) HPC facility



Block diagram of the
Zorba architecture

- 1. Chapter 1:** Introduction
- 2. Chapter 2:** Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment
 - 2.1** PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS & COI marker genes
 - 2.2** The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data
- 3. Chapter 3:** PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types
- 4. Chapter 4:** A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks
- 5. Chapter 5:** Deciphering the functional potential of a hypersaline marsh microbial mat community
- 6. Chapter 6:** Os and 1s in marine molecular research: a regional HPC perspective
- 7. Chapter 7: Conclusions**

Chapter 7: Conclusions

- Bioinformatics approaches enhance microbial diversity assessment 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" This could provide us with great insight on

the evolvability of metabolic decisions and on how such decisions affect microbial coexistence in the communities.

Wrap-up

software tools



a pipeline for eDNA metabarcoding analysis

github.com/hariszaf/pema



github.com/hariszaf/darn



github.com/lab42open-team/ github.com/GeomScale/dingo

the prego* repositories



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). Os 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



Acknowledgments

people and more

My promtors:

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. Pavlouri C.

My corner:

Would not be here
if it was not with you.

