

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

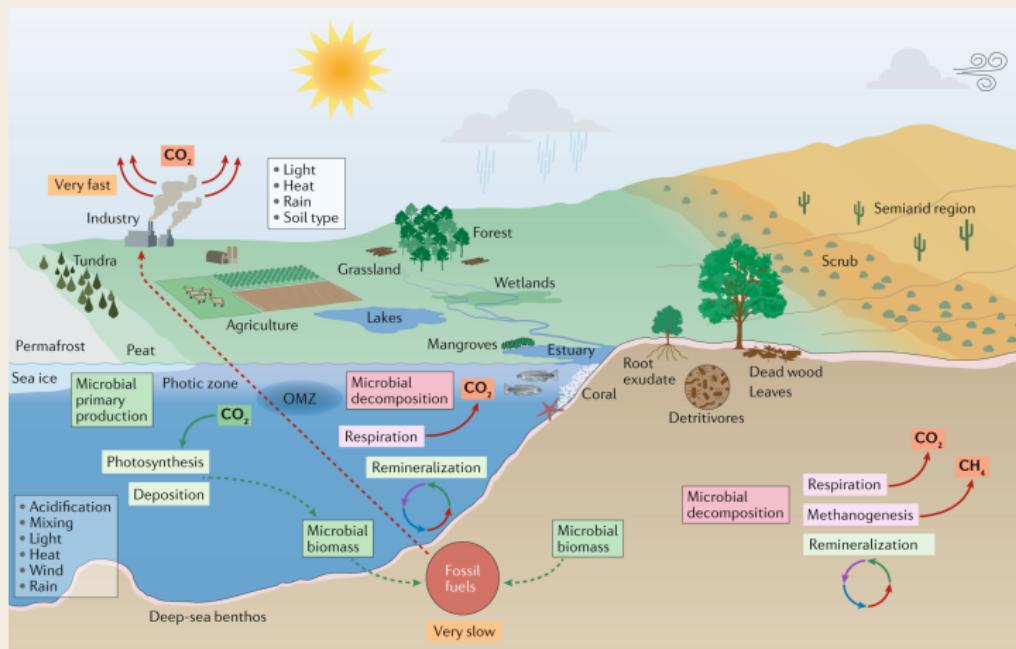


Figure from: Nature Reviews Microbiology 17.9 (2019): 569-586.

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

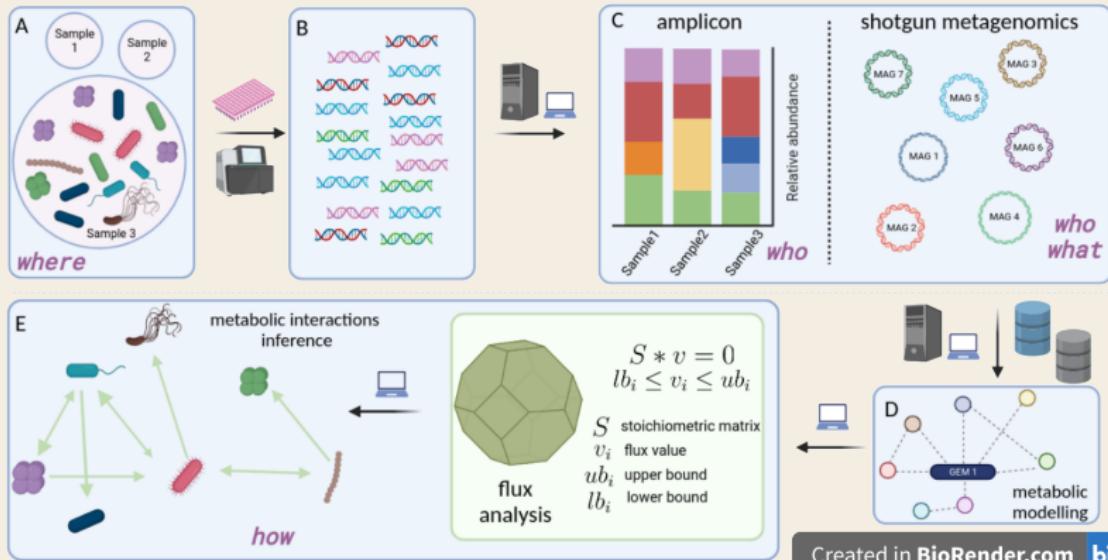


underlying
mechanisms
how

interactions, fluxes

Reverse ecology

transforming ecology into a high-throughput field



Created in BioRender.com

From raw reads to community analysis

not a straight-forward way

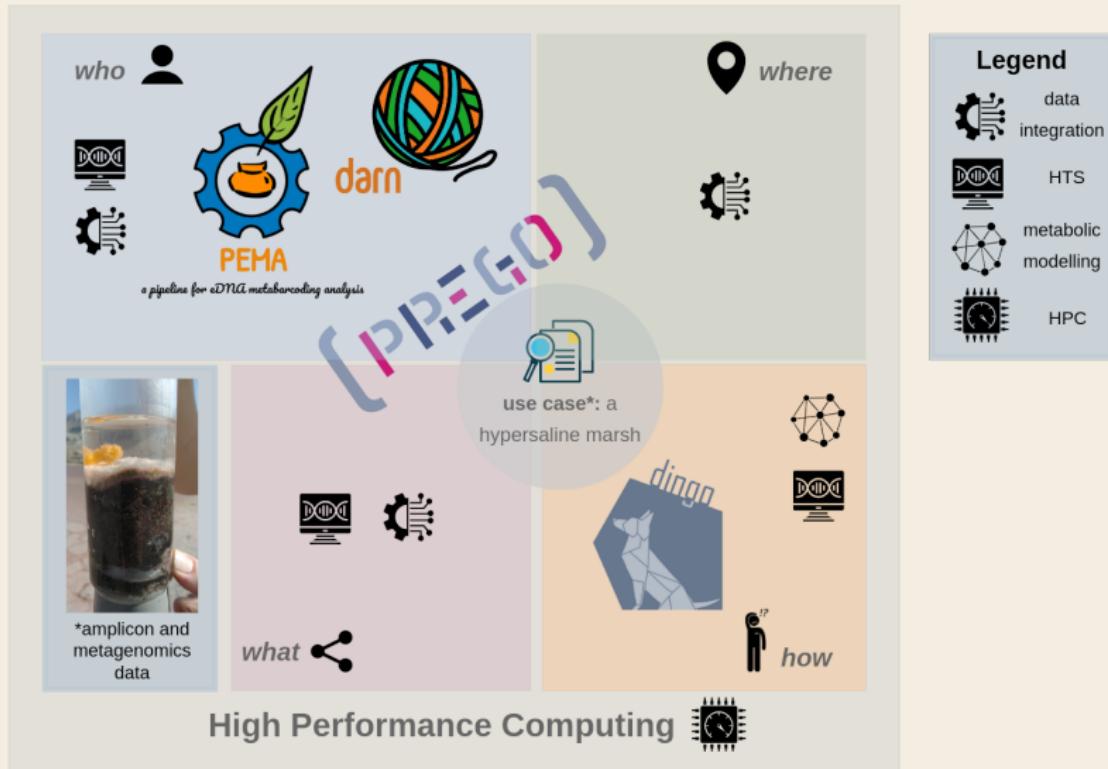
- computing requirements
- multi-step analyses
- reproducibility
- eDNA and PCR limitations
- algorithmic - oriented
- data integration applications



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



Os and 1s in marine molecular research *a regional HPC*

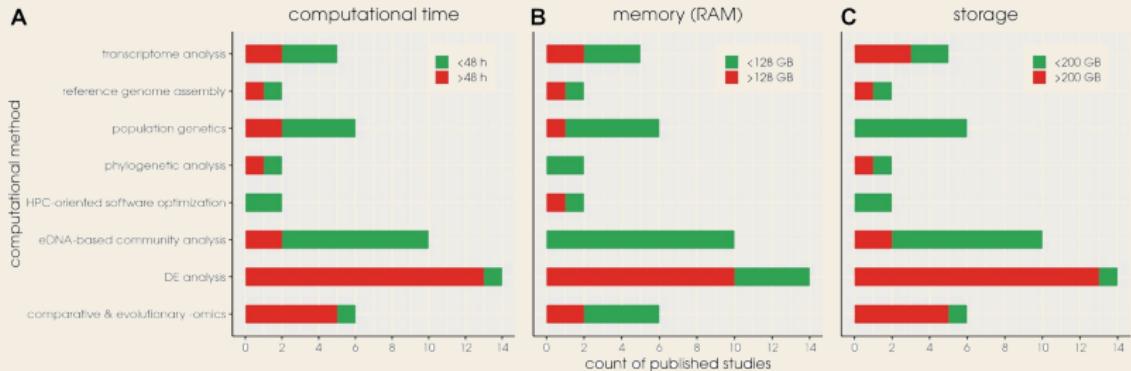


challenge category: computing requirements

aim & contribution: a retrospective analysis of resource allocation and computational methods supported by a regional HPC facility



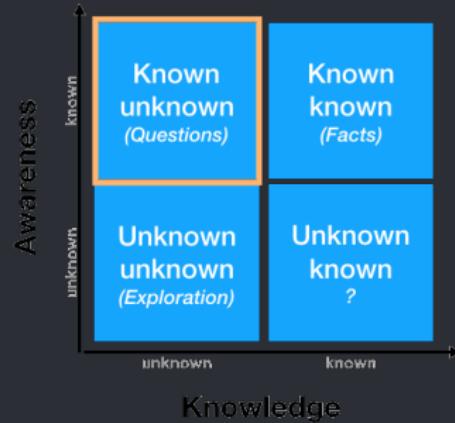
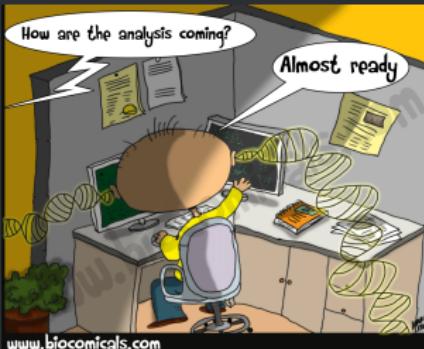
Computational requirements for trivial bioinformatic tasks



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 & reproducibility

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

Legend

	data integration
	HTS
	metabolic modelling
	HPC

who



a pipeline for eDNA metabarcoding analysis



where



*amplicon and
metagenomics
data

use case*: a
hypersaline marsh



what



how



High Performance Computing

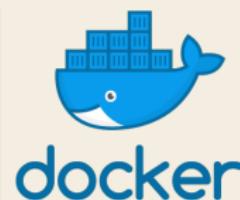


Methods / Implementation

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

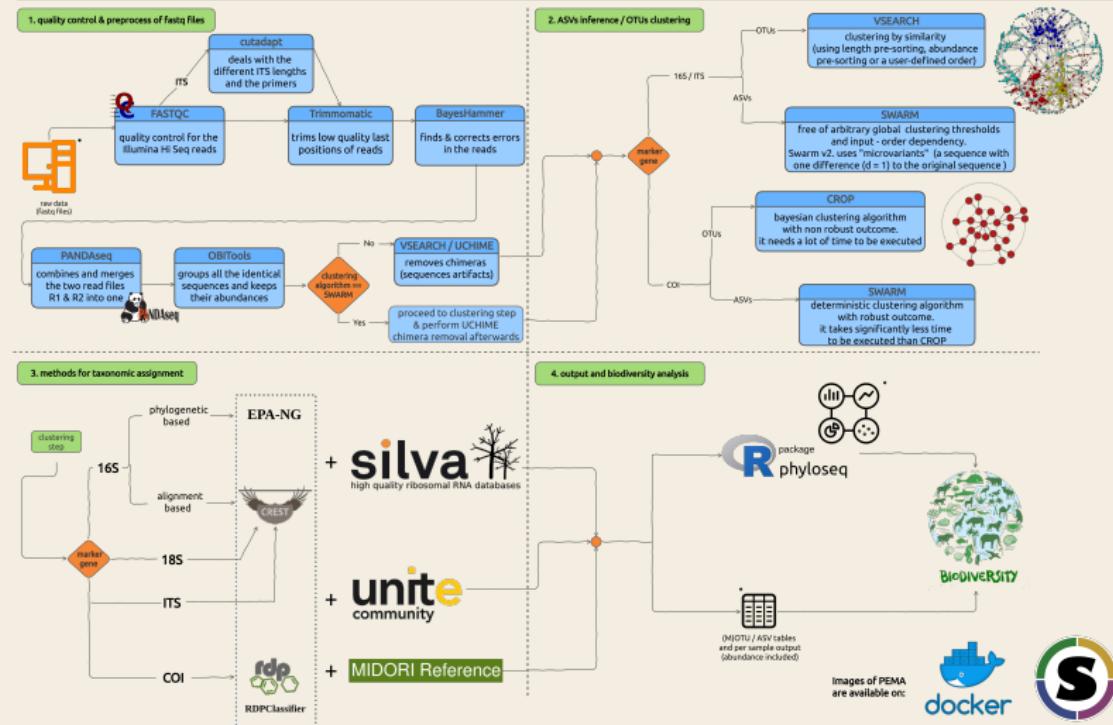
High performance
computing



PEMA features

an overview

PEMA in a nutshell





Results: tuning effects in mock communities

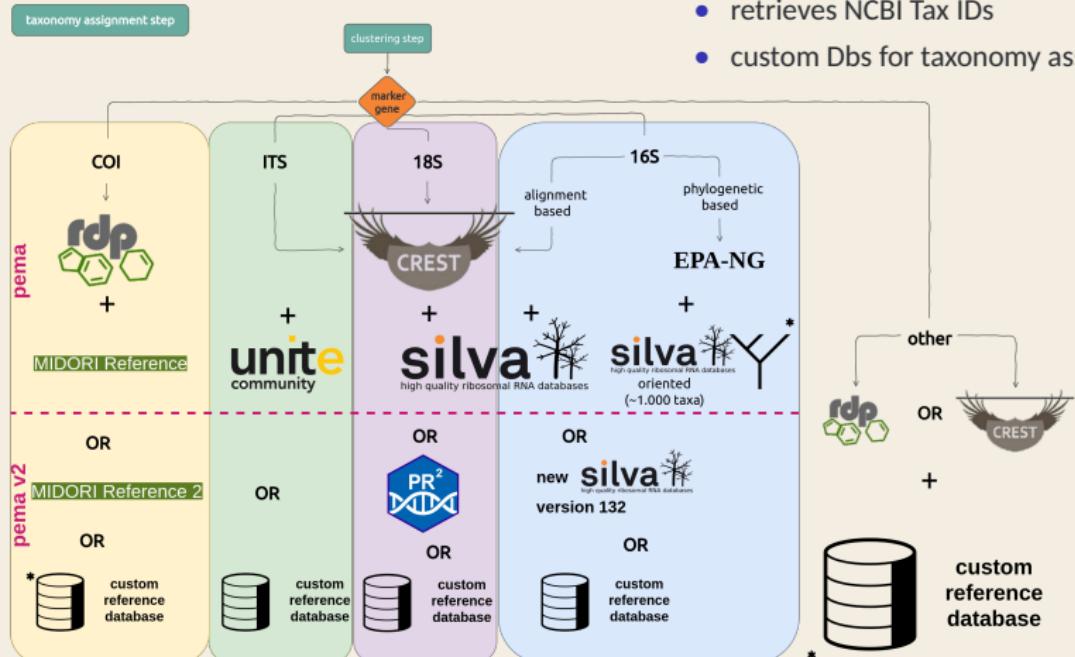
Mock communities using the 16S rRNA gene and multiple parameter sets
(identification at the genus level; part of the initial table):

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



* Illustrations with an asterisk are from the Noun Project



Moving at the large scale

PEMA @ infrastructures

The screenshot shows the III NIS Workflow Environment interface. On the left is a sidebar with categories like Tesseract, ARMS, Biotope, and Tools. The main area is titled "Run a ARMS workflow" and shows a "Workflow overview" with a diagram of four steps: Import TSV (Param file), PEMA (CSV / Tab), WaRMS (dataset), and WRMS (dataset). Below the diagram are steps: "Workflow description", "CSV input data", "PEMA parameters", "Create workflow", and "Workflow created".

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

1. web - interface make analysis even easier
2. researchers without access to HPC/clouds are now able to run big scale analyses
3. combine with other tools

Elixir Greece:
<https://elixir-greece.org/>
Hypatia cloud infrastructure:
<https://hypatia.athenarc.gr/>



Conclusions

on PEMA and eDNA metabarcoding

- **parameters tuning** is essential in metabarcoding analyses; **mock communities** among samples under study benefit to that end
- workflow managers & containers enable complex and reproducible workflows
- **e-infrastuctures** benefit studies with great number of samples and CLI non-familiar users

DARN

dark matter investigator tool

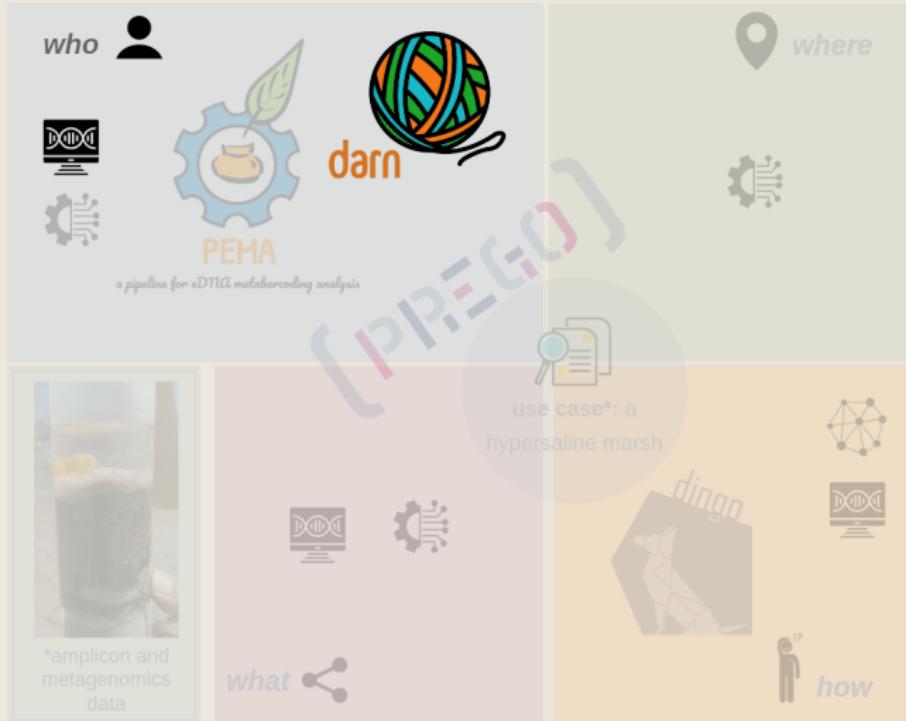


challenge category: eDNA - oriented issues

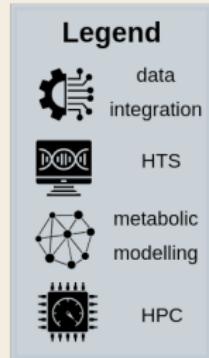
aim & contribution:

extract "dark matter" from COI amplicon data

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



High Performance Computing





Building a COI-oriented tree of life

sequences retrieved

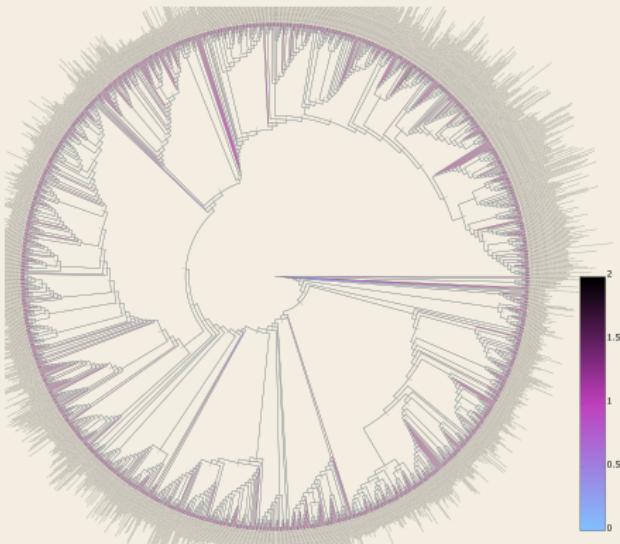
Resources	bacteria		archaea		eukaryotes	
	# of sequences	# of strains	# of sequences	# of strains	# of sequences	# of species
BOLD	3,917	2,267	117	117		
PFam-oriented	9,154	4,532	217	115		
Midori 2					1,315,378	183,330
Total unique entries	11,421	6,798	334	201	1,315,378	183,330



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



Reference phylogenetic tree of the COI consensus sequences retrieved

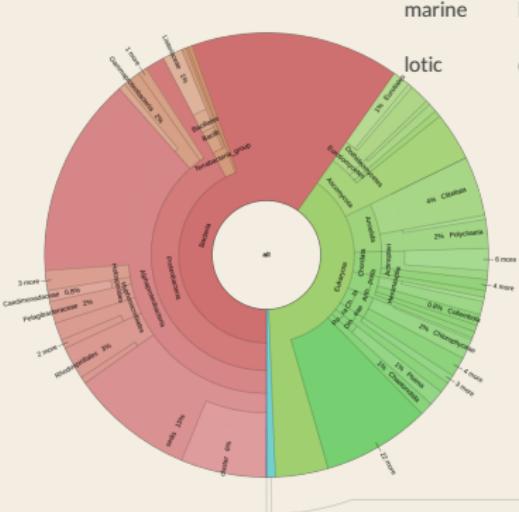


The consensus sequences have been placed in their corresponding taxonomic branches, proving the tree valid.



DARN using real-world data

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



Env 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	-
		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



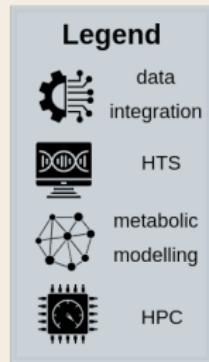
aim & contribution: support flux sampling at high dimensional polytopes such as those of complex organisms and multispecies and/or host-microbe communities

Definitions

- *flux*: the rate of turnover of molecules through a reaction
- *polytope*: a bounded polyhedron
- *flux sampling*: calculation of a sufficiently large number of uniformly distributed points in the polytope derived from a metabolic model

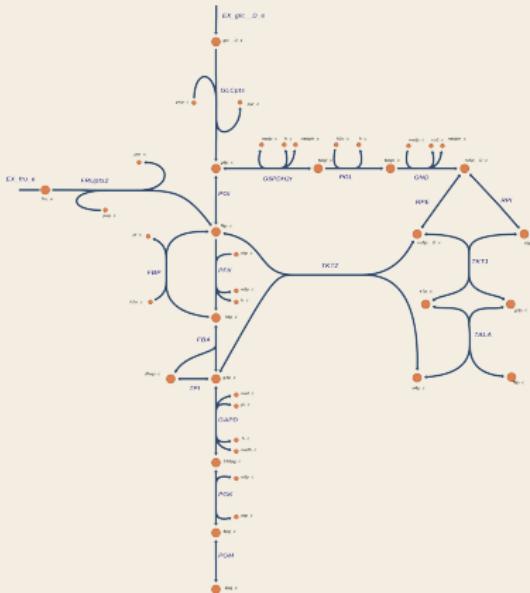


High Performance Computing





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.

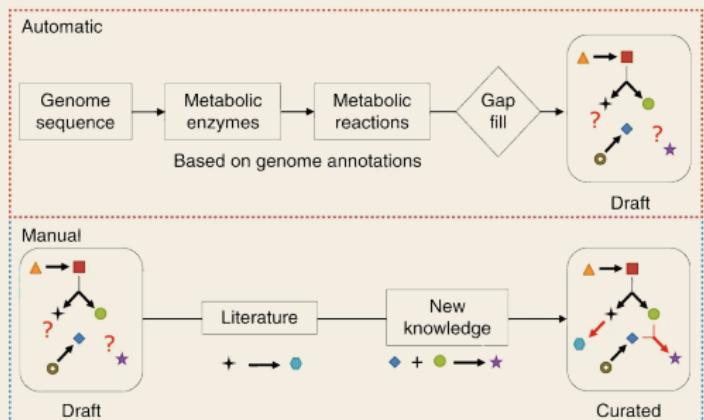


Figure from: Heirendt et al. Nature protocols 14.3 (2019): 639-702.



From a stoichiometric matrix to a constraint-based model

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

Flux vector

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

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

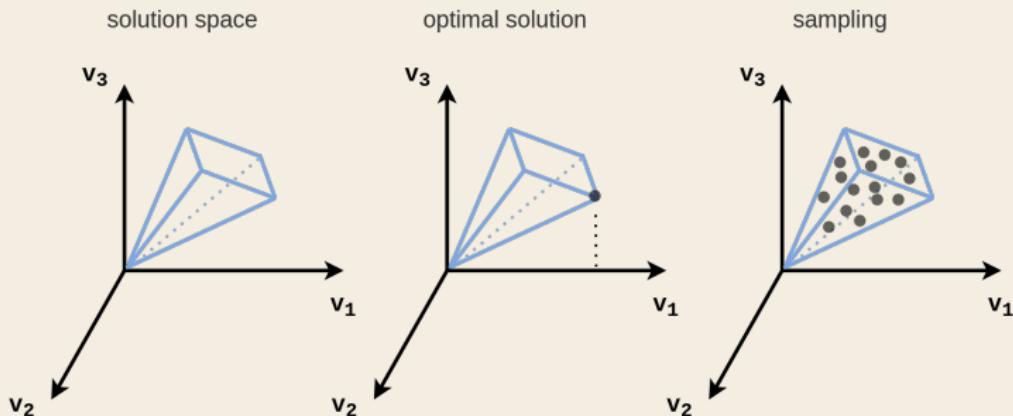
The **flux vector** is a vector with
the value of
each reaction flux
in a certain steady
state.

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



Flux sampling

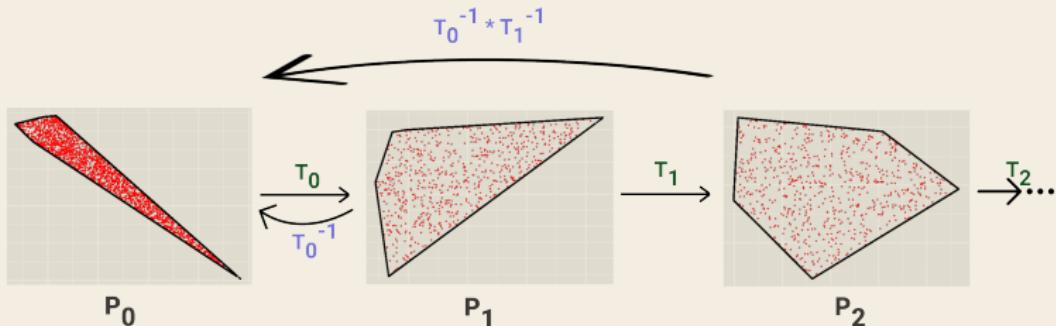
an alternative approach



- 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



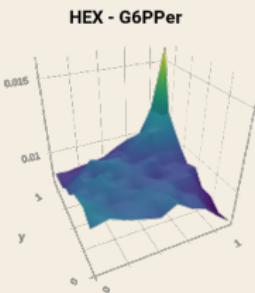
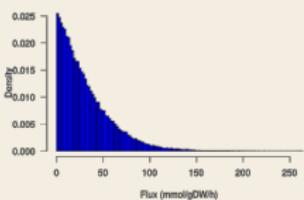
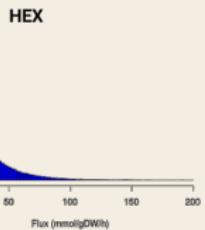
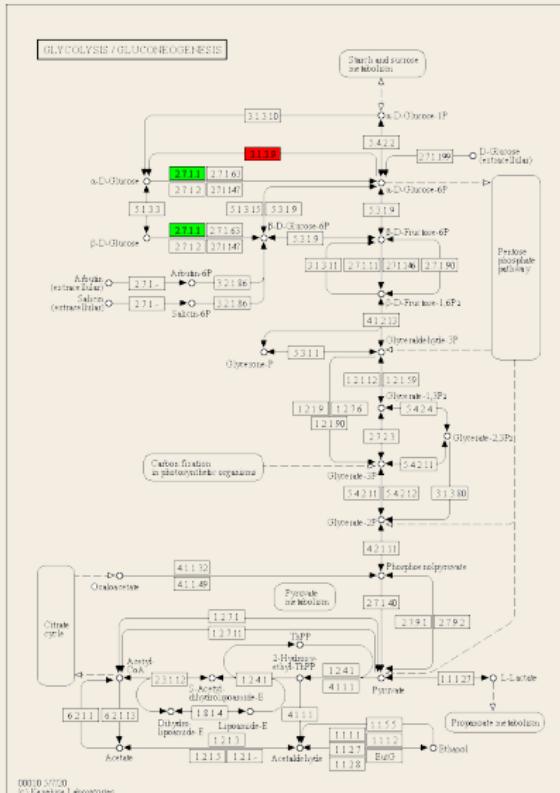
A new Markov Chain Monte Carlo algorithm for flux sampling



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:

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

G6PPer:

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



Conclusions

on sampling the flux space of metabolic models

- flux sampling provides essential insight (knock-out genes, host-microbe interactions etc)
- manual curation of the models is essential
- our multiphase MCMC algorithm enables sampling on the so-far largest metabolic model (13543 reactions; $d = 5335$)
- 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*

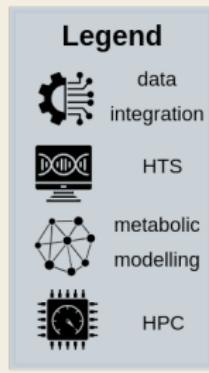
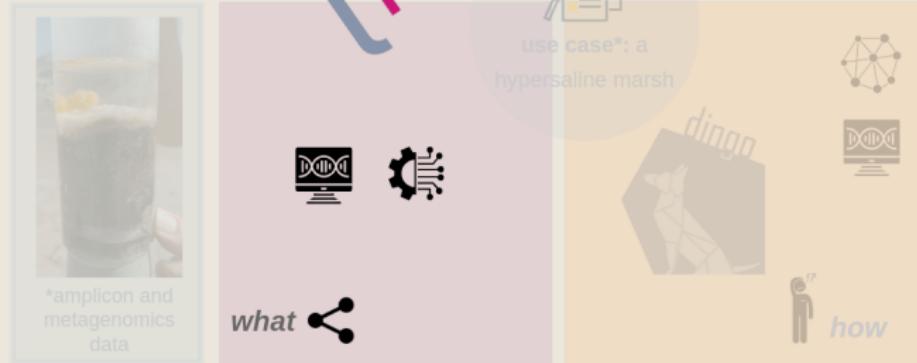


challenge category: data integration applications

aim & contribution:

To build a hypothesis generation resource based on associations between:

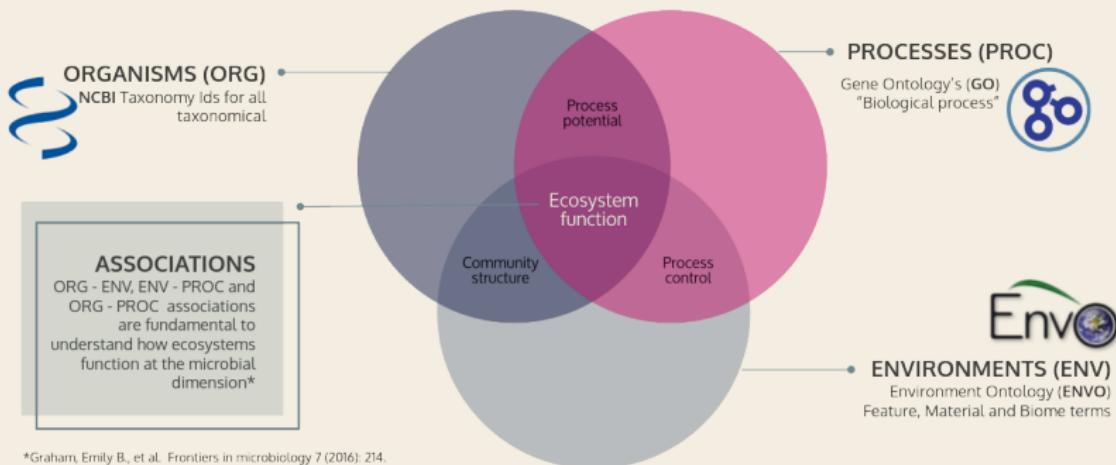
- *organisms and the environments they inhabit*
- *organisms and the biological processes they are involved with*
- *processes and the environments where they occur*



High Performance Computing

PREGO and its referring terms

the fundamental role of ontologies



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

Named Entity Recognition

counting co-occurrences to export associations

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 hydrogenivorans* [41], and *Sulfuricella denitrificans* [42]. These MAGs contained the diagnostic genes for **carbon fixation** (*rbcL.S*), **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.

Publicly available omics datasets and 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

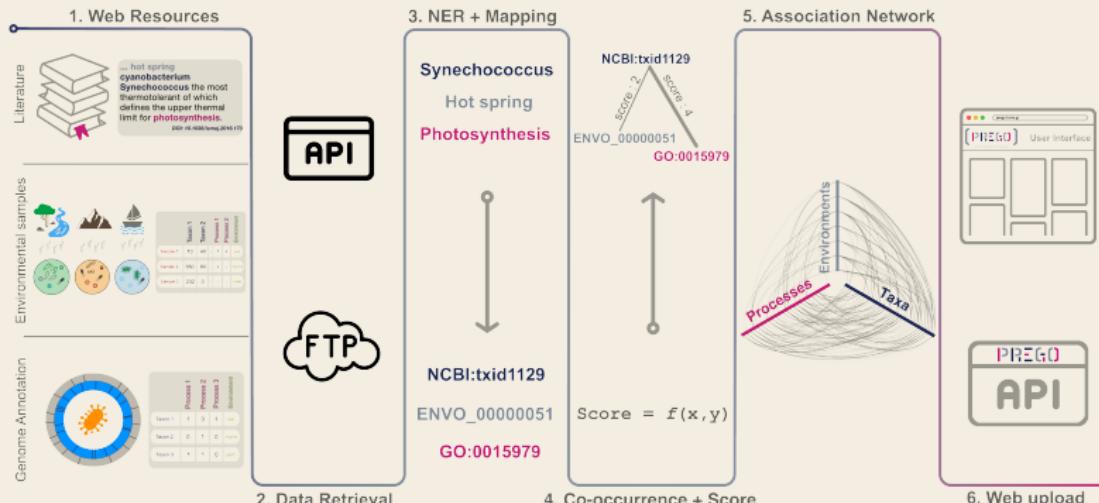
Project Information	
Cultured	No
Ecosystem	Environmental
Ecosystem Category	Aquatic
Ecosystem Subtype	Oceanic
Ecosystem Type	Marine

MG-RAST
metagenomics analysis server

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	 	

PREGO methodology

3 channels of information - 1 framework



Source databases 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	CCO
	7,965	metagenomic samples	yes	CCO
MGnify	10,500	marker gene samples	yes	CC-BY, CCO

PREGO in action

with a taxon as an entry point

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

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	★★★★★
Nertic sub-litoral 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

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

- metadata can give substantial added value in experimental data
- data integration methods may lead to associations not mentioned in the literature
- more and more valid associations as omics' dataset keep increasing exponentially and metadata standards gain space

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.

Legend

	data integration
	HTS
	metabolic modelling
	HPC

who



PEMA

a pipeline for eDNA metabarcoding analysis

where



use case*: a
hypersaline marsh



how



*amplicon and
metagenomics
data

what



High Performance Computing



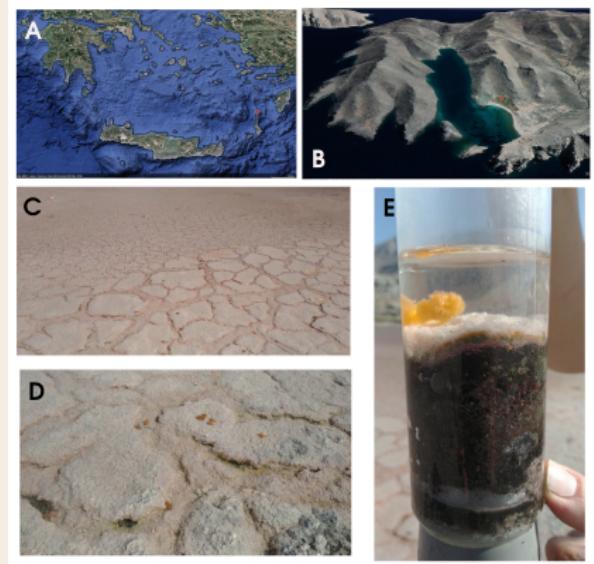


Tristomo swamp in Karpathos

a seasonal brackish water marsh

Sample	Type	Season
Elos01	top	summer
Elos02	bottom	summer
Elos03	orange aggregate	summer
Elos07	pink aggregate	summer
Elos10	combined*	winter
Elos12	combined	winter

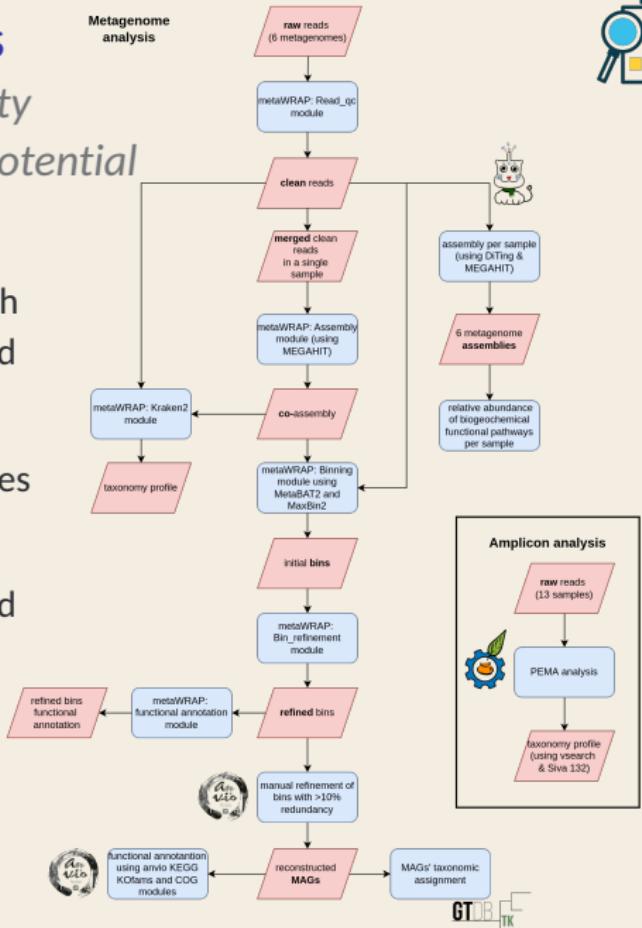
*combined: samples with no slicing



Bioinformatics analysis from raw data to community composition & functional potential

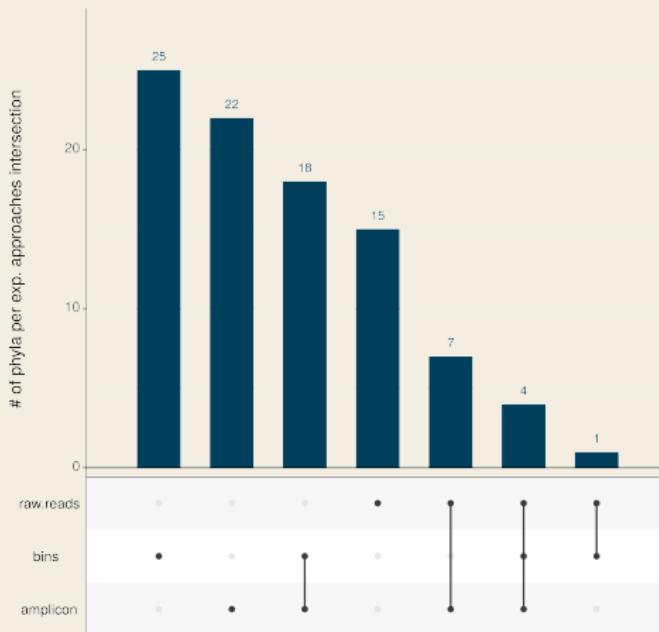


- metagenomic reads were both co - assembled and assembled at the sample level
- taxonomic & functional profiles per sample were retrieved
- MAGs were reconstructed and annotated



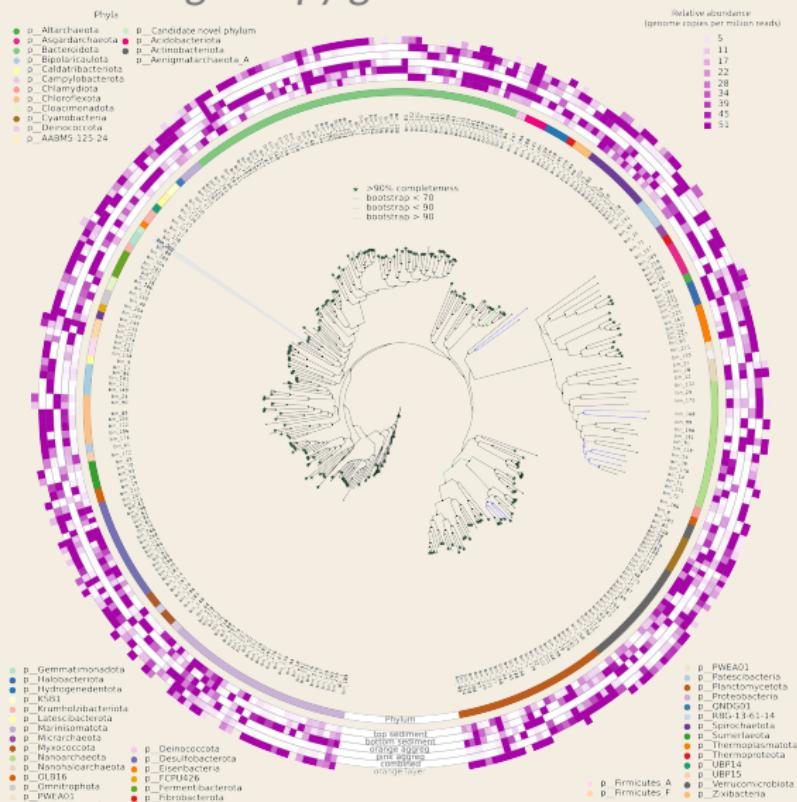
Amplicon vs metagenomics

number of phyla retrieved per method



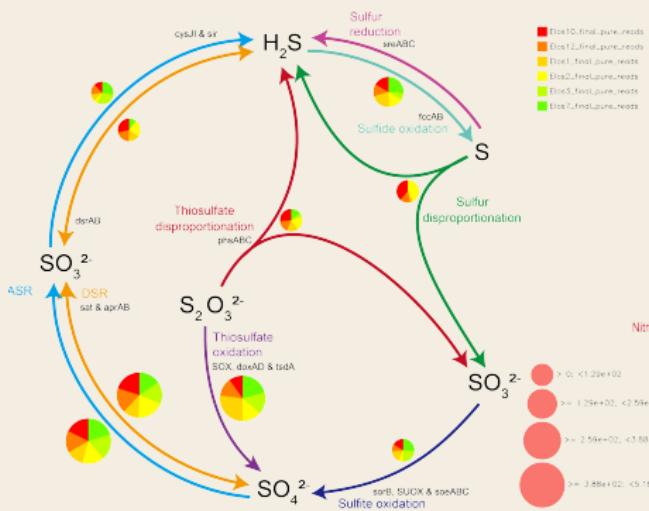
MAGs phylogeny

based on 25 single-copy genes

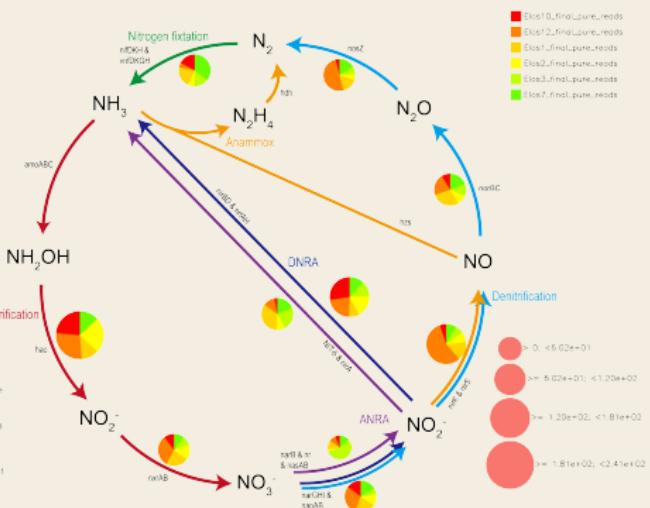




The S and the N cycle using KEGG annotation terms



Sample	Type	Season
Elos01	top	summer
Elos02	bottom	summer
Elos03	orange aggregate	summer
Elos07	pink aggregate	summer
Elos10	combined	winter
Elos12	combined	winter





Conclusions

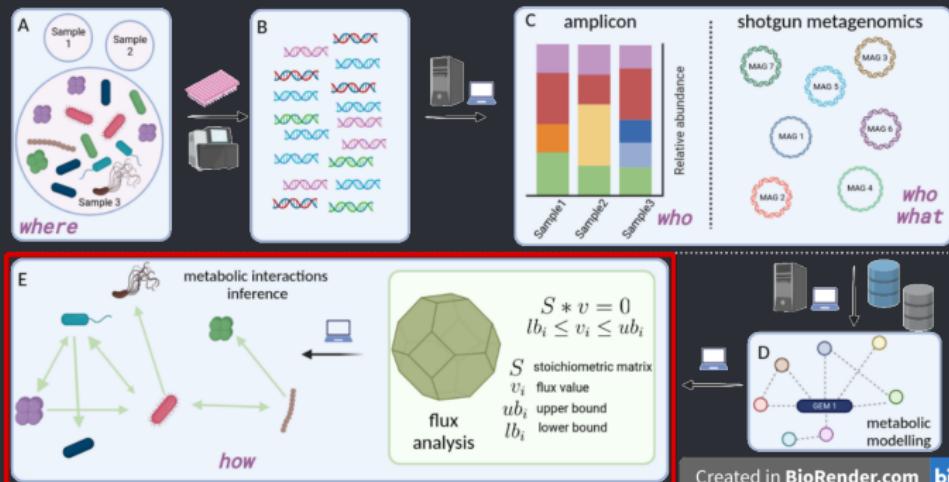
based on metagenomics analysis in the Tristomo marsh

- shotgun metagenomics are a powerful tool in the discovery of novel taxa
- seasonality plays a key-role for the mats under study, ensuring oxygenic photosynthesis
- anaplerotic reactions may play a key-role in replenishing the intermediates of the TCA cycle
- metabolic modelling can shed further light on the mechanisms governing the mat communities

General 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



Created in BioRender.com

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



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



Wrap-up 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] 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.
- [4] 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)*.
- [5] 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.
- [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] 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*)
- [8] 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*)
- [9] 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*)
- [10] Chalkis, A., Fisikopoulos, V., Tsigaridas, E. & Zafeiropoulos, H. dingo: a Python library for metabolic networks analysis (*under preparation*)

Acknowledgments

funding & grants



GENERAL SECRETARIAT FOR
RESEARCH AND INNOVATION



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



Acknowledgments

people and more

My promtors:

Prof. Emmanuel Ladoukakis drs. Savvas Paragkamian Dr. Christina Pavloudi

Prof. Christoforos Nikolaou Dr. Laura Gargan

Dr. Evangelos Pafilis

Dr. Sanni Hintikka

the rest committee

Mr. Stelios Ninidakis

members:

Mr. Antonis Potirakis

Prof. Konstadia (Dina) Lika

Dr. Apostolos Chalkis

Prof. Panagiotis Sarris

Dr. Vissarion Fisikopoulos

Prof. Jens Carlsson

Prof. Elias Tsigaridas

Prof. Karoline Faust

My mojo:

My corner:

Would not be here

if it was not with you.



Computing infrastructures an alternative for the most!



We will develop a workflow for the analysis of Genomic Observatories data that will allow researchers to deal better with the increasing amount of data.



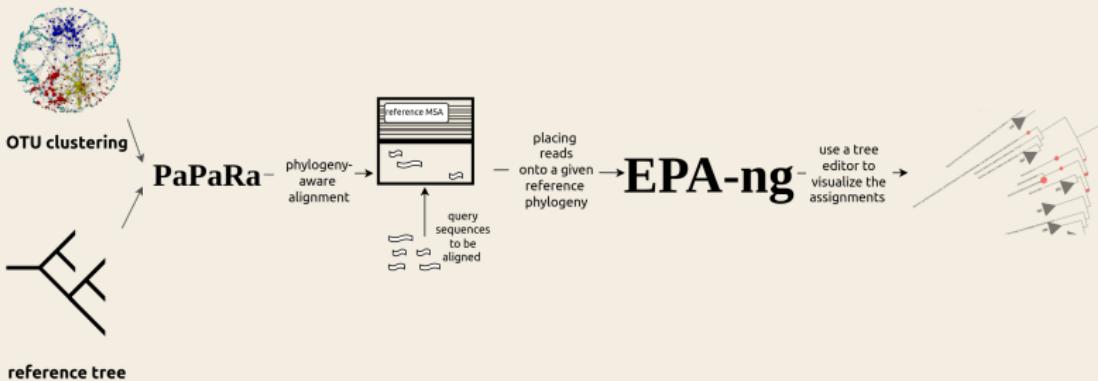
PEMA features

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

A. create reference tree



B. phylogeny-based taxonomy assignment



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!](#)

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 in which they co-occur
- similarly, in the *Literature* channel, based on the number of publications

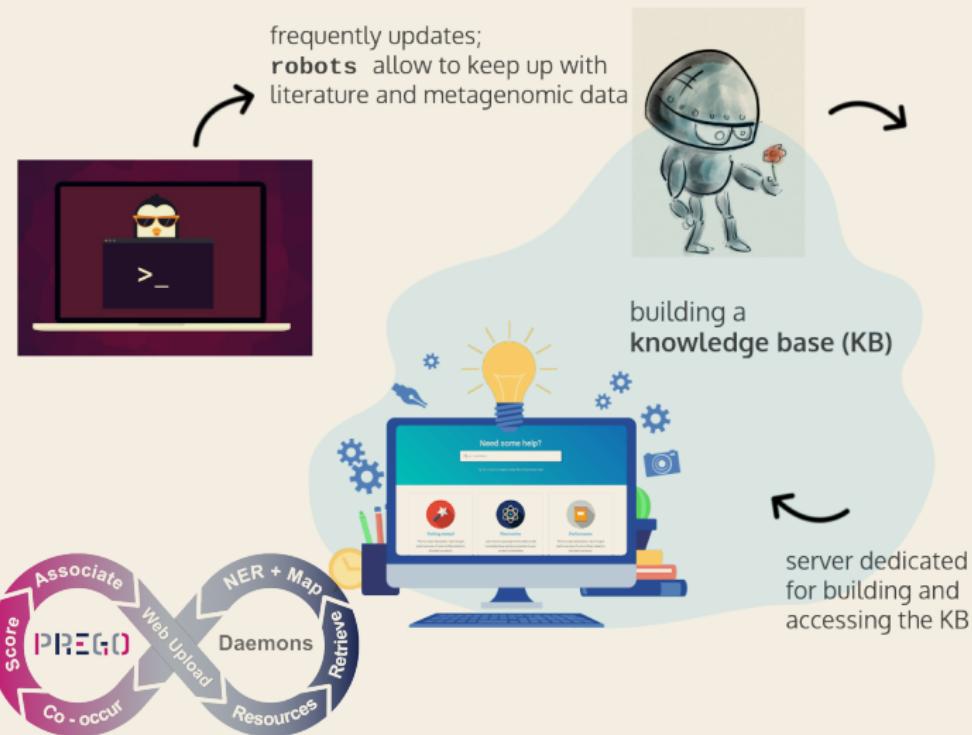
		$Y = y$		Total
		Yes	No	
$X = x$	Yes	$c_{x,y}$	$c_{x,0}$	$c_{x..}$
	No	$c_{0,y}$	$c_{0,0}$	$c_{0..}$
	Total	$c_{..y}$	$c_{.,0}$	$c_{...}$

Environmental samples score:

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

Building a knowledge-base

development and information technology operations



© Thanos Dailianis

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:

$$\frac{d\omega}{dt} = Sv$$

The region of steady states

moving to full dimensional polytope

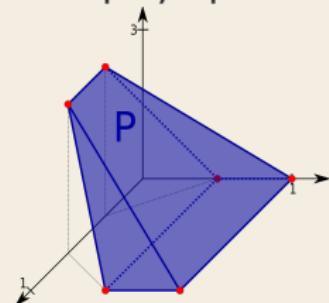
The *constraints* on the reactions fluxes.

$$Sv = 0, \quad (2) \quad v = Nx$$

$$v_{lb} \leq v \leq v_{ub}$$

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

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).

Computational geometry basics

A powerful approach to obtain well **roundness** is to put P in *near isotropic position*. In general, $K \subset \mathbb{R}^d$ is in isotropic position if the uniform distribution over K is in isotropic position, that is $\mathbb{E}_{X \sim K}[X] = 0$ and $\mathbb{E}_{X \sim K}[X^T X] = I_d$, where I_d is the $d \times d$ identity matrix.

Thus, to put a polytope P into isotropic position one has to generate a set of uniform points in its interior and apply to P the transformation that maps the point-set to isotropic position; then iterate this procedure until P is in c -isotropic position for a constant c .

Random walk performance metrics

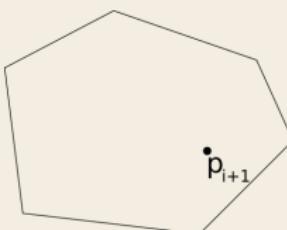
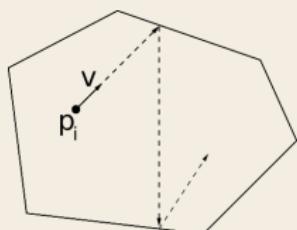
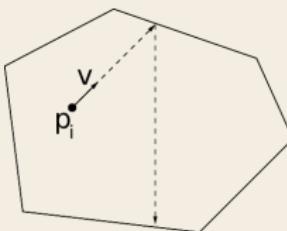
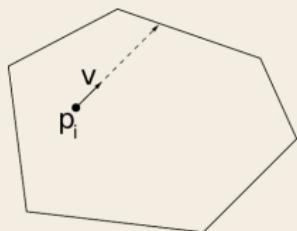
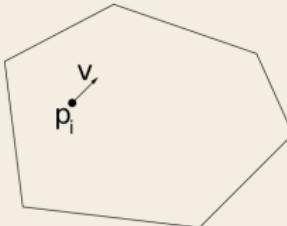
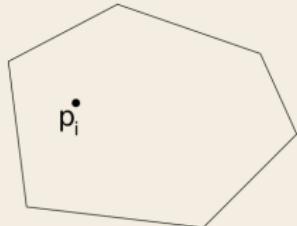
- **cost per iteration:** the number of operations the algorithm needs to sample a single point
- **mixing time:** the number of samples the algorithm needs to burn in order to lose dependency from previous iterations (i.i.d)
- **cost per sample:** the total number of operations the algorithm needs to sample an i.i.d point

MCMC convergence diagnostics:

- **Effective Sample Size (ESS):**
the number of effectively independent draws from the target distribution that the Markov chain is equivalent to
- **Potential Scale Reduction Factor (PSRF)**

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 $p_i + 1$.

Find possible targets against SARS-CoV-2

a flux sampling application

Bioinformatics, 36(26), 2020, i813–i821

doi: 10.1093/bioinformatics/btaa813

ECCB2020

OXFORD

Systems

FBA reveals guanylate kinase as a potential target for antiviral therapies against SARS-CoV-2

Alina Renz^{1,2,*}, Lina Widerspick¹ and Andreas Dräger^{1,2,3,*}

¹Computational Systems Biology of Infections and Antimicrobial-Resistant Pathogens, Institute for Bioinformatics and Medical Informatics (IBMI) and ²Department of Computer Science, University of Tübingen, Tübingen 72076, Germany and ³German Center for Infection Research (DZIF), partner site Tübingen, Germany

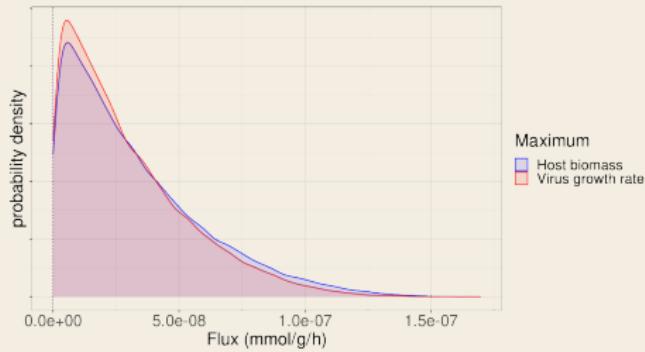
- Renz *et al.* 2020 built the biomass function of SARS-CoV-2 to build a host - virus network
- Using FBA they computed an optimal steady state using
 - (i) human biomass maintenance
 - (ii) virus growth rate
- They found reaction GK1 as a possible anti-viral target

Find possible targets against SARS-CoV-2

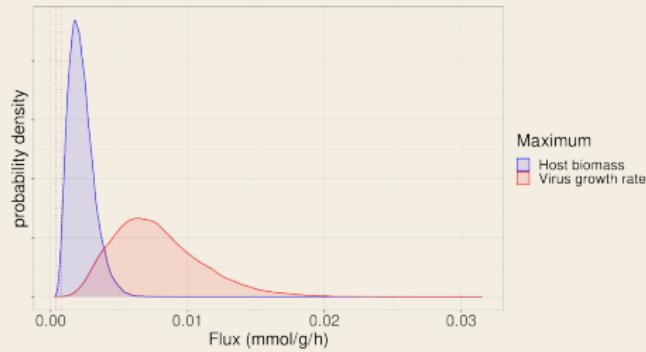
a flux sampling application



Reaction TYMSULT



Reaction GK1



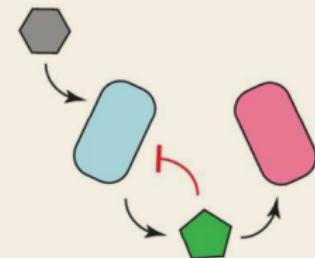
- Check if the flux distribution of a reaction changes.
- Find possible anti-viral targets and study further.

For more about this example case, you may check this [blog-post](#).

Further applications of metabolic flux sampling



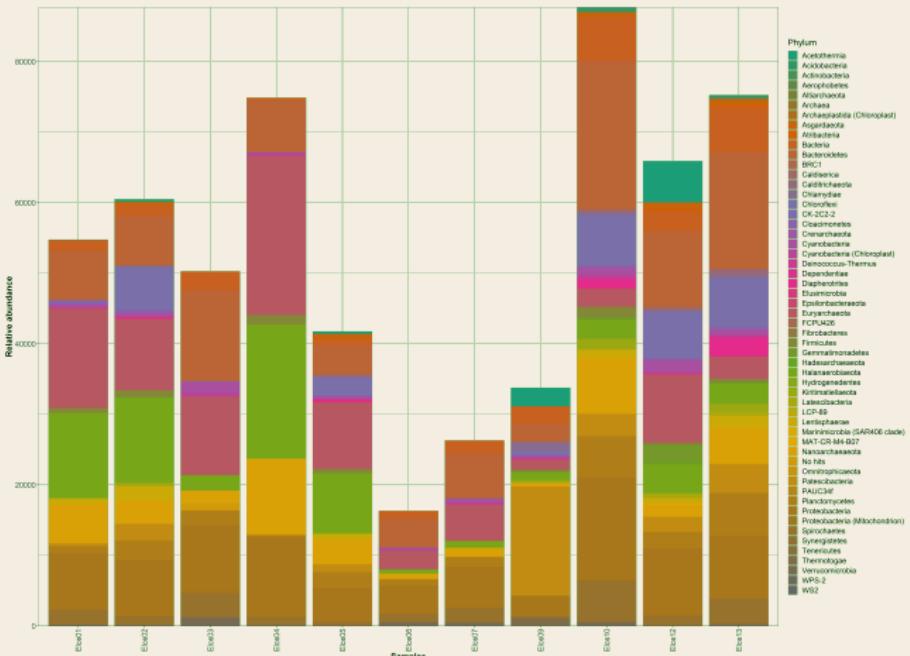
Scott, William T., et al. "Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts." Microbial cell factories 20.1 (2021): 1-15.



What about microbial interactions ?

Abundances of the main microbial taxa, at the phylum level

based on 16S rRNA amplicon data



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

