ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

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

Haris Zafeiropoulos

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Science (PhD) in Biology

Promotors:

Prof. Emmanouil Ladoukakis
Dr Evangelos Pafilis
Dr Christoforos Nikolaou

Academic year 2021 – 2022

Members of the examination committee
&
reading committee

Prof. Emmanouil Ladoukakis
Univeristy of Crete
Biology Department

Dr. Evangelos Pafilis
Hellenic Centre for Marine Research
Institute of Marine Biology, Biotechnology and Aquaculture

Dr. Christoforos Nikolaou
Biomedical Sciences Research Center “Alexander Fleming”
Institute of Bioinnovation

Prof. Konstadia (Dina) Lika
Univeristy of Crete
Biology Department

Prof. Panagiotis Sarris
University of Crete
Department of Biology

Dr. Jens Carlsson
University College Dublin
School of Biology and Environmental Science/Earth Institute

Prof. Karoline Faust
KU Leuven
Department of Microbiology and Immunology, Rega Institute

Contents

Abstract

Microbial communities are a cornerstone for most ecosystem types. To elucidate the mechanisms governing such assemblages, it is fundamental to identify the taxa present (*who*) and the processes that occur (*what*) in the various environments (*where*). Thanks to a series of technological breakthroughs vast amounts of information/data from all the various levels of the biological organization have been accumulated over the last decades. In this context, microbial ecology studies are now relying on bioinformatics methods and analyses. Therefore, a great number of challenges both from the biologist- and the computer scientist point-of-view have arisen; one among the most emerging ones being: *"what shall we do with all these pieces of information?"*. The paradigm of Systems Biology addresses this challenge by moving from reductionism to more holistic approaches attempting to interpret how the properties of a system emerge.

Aim of this PhD was to enhance microbiome data analyses by developing software addressing on-going computational challenges on the study of microbial communities. On top of that, to exploit state-of-the-art methods to identify taxa, functions and microbial interactions in assemblages of various aquatic environments. To this end, a number of publicly available data-sets were used while a swamp from the Karpathos island (Greece), was chosen as a study case for the described framework.

Environmental DNA and metabarcoding have been widely used to estimate the biodiversity (the *who*) and the structure of communities. Vast amount of sequencing data targeting certain marker genes depending the taxonomic group of interest become available thanks to High Throughput Sequencing technologies. However, the bioinformatics analysis of such data require multiple steps and parameter settings. Software workflow-oriented tools along with computing infrastructures ease this need to a great extent and PEMA was developed to this end (Chapter 2.1). However, eDNA metabarcoding has limitations too. Cytochrome c oxidase subunit I (COI) marker gene is a commonly used marker gene, especially in studies targeting eukaryotic taxa. It is well known that in COI studies a great number of the derived OTUs/ASVs get no taxonomic hits. The presence of non-eukaryotic taxa with their simultaneous absence from the most commonly-used reference databases justify this phenomenon to a great extent. DARN makes use of a COI-oriented tree of life to provide further insight to such known unknown sequences (Chapter 2.2).

Shotgun metagenomics provide further information regarding the processes that occur in a community (the *what*). Sediment and microbial mat samples as well as microbial aggregates from a hypersaline swamp in Tristomo bay (Karpathos, Greece) were analyzed. Both amplicon (16S rRNA) and shotgun sequencing data were used to characterize the

microbial structure of the communities and environmental parameters (e.g. salinity, oxygen concentration, granulometric composition) were measured at the sampling sites. Key functions supporting life in such environments were identified and metagenome-assembled genomes (MAGs) of major species found were built (Chapter **??**).

Amplicon and shotgun metagenomics approaches along with the rest of the omics technologies, have led to vast amount of data and metadata, recording the *who*, the *what* and the *where*. To enable optimal accessibility and usage of this information, a great number of databases, ontologies as well as community-standards have been developed. By exploiting data integration techniques to bring such bits of information together, as well as text mining methods to retrieve knowledge "hidden" among the billions of text lines in already published literature, the PREGO knowledge-base generates thousands of *what - where - who* potential associations (Section **??**).

The driving question though is *how* the different microbial taxa ascertain their endurance as part of a community. Metabolic interactions among the various taxa play a decisive role for the composition of such assemblages. Genome-scale metabolic networks (GEMs) enable the inference of such interactions. Random sampling on the flux space of such metabolic models, provides a representation of the flux values a model can get under various conditions. However, flux sampling is challenging from a computational point of view. To address such challenges, a Python library called dingo was developed using a Multiphase Monte Carlo Sampling algorithm (Chapter **??**). GEMs were built using the MAGs retrieved from the Tristomo swamp and metabolic interactions between them and their environment were investigated.

Similar to microbial communities, bioinformatics methods tend to build assemblages while "living" on your own is quite rare. The methods developed during this PhD project combined with state-of-the-art methods anticipate to build a framework that enables moving from the community to the species level and then back again to the one of the community.

Περίληψη

Και στα ελληνικά δουλευει σωστά

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

COI	Cytochrome c oxidase subunit I
ITS	Internal Transcribed Spacer
NGS	Next Generation Sequencing
eDNA	environmental DeoxyriboNucleic Acid
OTU	Operational Taxonomic Unit
ASV	Amplicon Sequence Variant
HPC	High Performance Computing
MCMC	Markov Chain Monte Carlo
MMCS	Multiphase Monte Carlo Sampling
PREGO	PRocess Environment OrGanism
PEMA	Pipeline for Environmental DNA Metabarcoding Analysis
DARN	Dark mAtteR iNvestigator
PSRF	Potential Scale Reduction Factor
ESS	Effective Sample Size

Symbols

$\mathbb{R}$	Set of real numbers
$\mathcal{O}$	Algorithm complexity
$\widetilde{\mathcal{O}}$	Algorithm complexity ignoring polylogarithmic factors
$\mathbb{E}$	Expected Value operator
$\mathcal{U}$	utility function....
ℓ	a ray
P	a polytope
τ	a trajectory
∂P	boundary of the P polytope
v	a flux vector
v_i	flux value of the reaction i
W	walk length
ρ	number of reflections

Chapter 1

Introduction

1.1 Microbial communities: composition , functions & interactions

1.1.1 Microbial diversity: life under extraordinary conditions

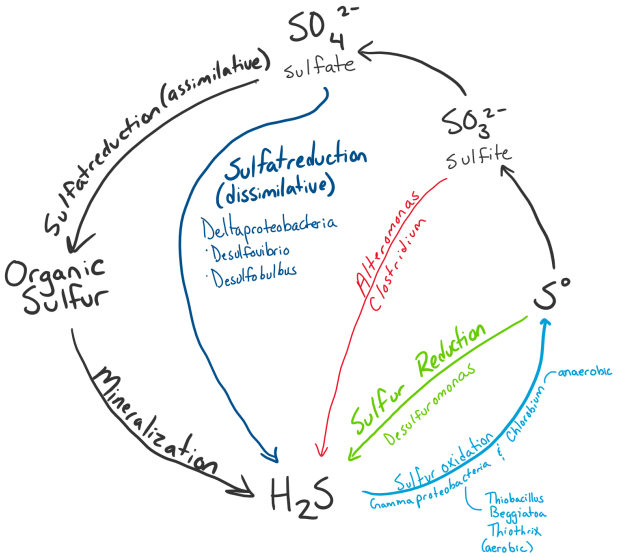
Microbes are considered to be omnipresent in the various ecosystems on Earth [1]. It was only until recently, 2019, that scientists discovered for the first time a place on Earth where no microbial forms of life are present [2]. Extremely low pH, high salt and high temperature had to be at the same place at the same time to stop microbes. However, microbes are not just abundant but exceedingly variant too. Locey and Lennon using a unified scaling law and a log-normal model of biodiversity, estimated microbial diversity at about 1 trillion species [3]. However, despite the extensive studies of the scientific community, less than 1% of the microbial species on Earth have been identified [4].

Microbes are distinguished by multiple properties. Based on their morphology microbes can be spherical (cocci), rod-shaped (bacilli), arc-shaped (vibrio), and spiral (spirochete) [5]. Based on their metabolic characteristics, microbes are further distinguished. More specifically, according to their *energy source*, a microbe can either oxidate inorganic compounds (**chemotrophs**) or sunlight (**phototrophs**). Similarly, microbes can use CO_2 (**autotrophs**) as their *carbon source*, or organic compounds (**heterotrophs**) or both (**mixotrophs**). Finally, based on their *electron source* microbes are distinguished between those using inorganic compounds (**lithotrophs**) and those using organic compounds (**organotrophs**) [6]. Microbial taxa combine combining alternatives of the aforementioned categories shape a range of microbial profile of all the possible combinations; for example **chemolithoautotrophic** bacteria, e.g. nitrifying and sulfur-oxidizing bacteria, as well as **photoautotrophic** bacteria, e.g. purple bacteria and Green sulfur bacteria. Finally, microbial taxa can also be distinguished by their various ecological distributions and activities, and by their distinct genomic structure, expression, and evolution [5].

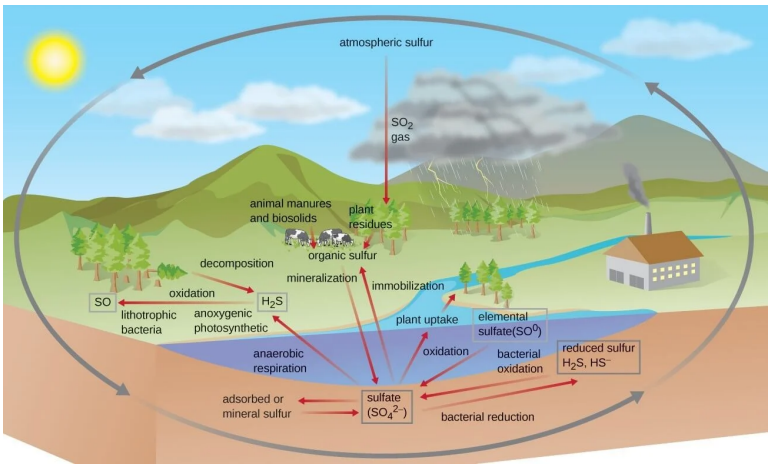
1.1.2 Functional diversity: shaping the conditions of life

However, it is not only the number of microbial taxa and their massive biomass that make the study of microbial communities essential; it is mostly their functional potentials. Life

on Earth would not be as we know it, if existed at all, if it was not for the microbes and their long contribution on ensuring life-supporting conditions. Nevertheless, these are the *biological machines responsible for planetary biogeochemical cycles* [1]; meaning that biogeochemical cycling to a global extent is powered by the metabolic processes of the microbial taxa [7]. In Figure 1.1 the contribution of microbial communities in the cycle of CO_2 is shown.



(A) Basic reactions in sulfur cycle



(B) Sulfur cycle reactions per environmnetal type

FIGURE 1.1: The cycle of sulfur (S) (up) and the contribution of microbial communities on it (down, image source: OpenStax).

The biological fluxes of most of the major elements (i.e., carbon, hydrogen, oxygen, nitrogen and sulfur) required for any biological macro-molecule, are driven largely by

microbially catalyzed, thermodynamically constrained redox reactions [1]. Phosphorus the last of the 6 fundamental elements for life, is also included in the metabolic pathways catalyzed by microbes. Thus, microbial communities consist of hundreds or even thousands of metabolically diverse strains and species [8], and their functions and determine the fitness of most organisms on Earth. In case of human health, specific microbial enzymatic pathways and molecules necessary for health promotion have been well known. Some of these "beneficial factors" are already known for probiotics and species in the human microbiome [9].

The relationship between the taxonomic and the functional profile of a microbial community has been an open question for scientists; is the *who* or the *what* more important to distinguish communities [10]? And how does each of these profiles respond to the various perturbations of an environment; Do they tend to converge [11]? Do perturbations of the taxonomic composition of a community influence the robustness of the community's functional profile [12]? divergence of each under and from an evolutionary point-of-view. Does it matter *who* is doing *what* and how does this affect the niche of a species [13]? And what about the rare taxa and their corresponding functions in an assemblage [14, 15]?

Microbial Ecology focuses on the study of the following interactions:

- those between microbial taxa and their environment
- those among the various microbial taxa present in a community, and
- those between microbial taxa and their host [4]

Microbial ecologists also investigate the role of microbial taxa in biogeochemical cycles [1] and their interaction with anthropogenic effects e.g. pollution and climate change [16].

Even though HTS has allowed a massive extension of our knowledge in specific enzymatic reactions that regulate these pathways the rules that determine the assembly, function, and evolution of these microbial communities remain unclear. Thus, both in case of environmental and human the underlying mechanisms for how microbial assemblages work and affect their environment, remain to be discovered. Understanding the underlying governing principles is central to microbial ecology [17] and crucial for designing microbial consortia for biotechnological [18] or medical applications [19].

Studies such as the one of Louca et al. have opened new frontiers in our understanding on microbial assemblages. After building metabolic functional groups and assigning more than 30,000 marine species to these groups, Louca et al. showed that the distribution of these functional groups were influenced by environmental conditions to a great extent, shaping *metabolic niches*. At the same time though, the taxonomic composition within individual functional groups were not affected by such environmental conditions [7].

1.1.3 Ecological interactions in microbial communities

Moreover, to elucidate how these assemblages work the biotic interactions have to be considered too. Microbial interactions play a fundamental role in deciphering the underlying mechanisms that govern ecosystem functioning [20, 21]. Microbes secrete costly

metabolites (called **byproducts**) to their environment, which other microbes can absorb and exploit [22]. By exchanging metabolic products, mostly as there are also other ways of interactions e.g. quorum sensing, microbial taxa establish various interactions.

The interaction between two taxa can either be neutral or positive / negative (Figure 1.2). In case of a positive interaction, there is a case where both taxa benefit one from another. This *win-win* relationship is called **mutualism** (or "cooperation") and it can be a result of *cross-feeding*, in which two species exchange metabolic products [21]. Such is the case in biofilms where multiple bacterial taxa are working together building a structure that provides them antibiotic resistance [23]. There is also the case where only one of the two taxa benefits without helping or harming the other; this interaction is called **commensalism** [21]. For example, *Nitrosomonas* oxidize ammonia (NH_3) into nitrite (NO_2^-), so *Nitrobacter* can use it to obtain energy and oxidize it into nitrate (NO_3^-) [24]. Such interactions are quite common in microbial communities.

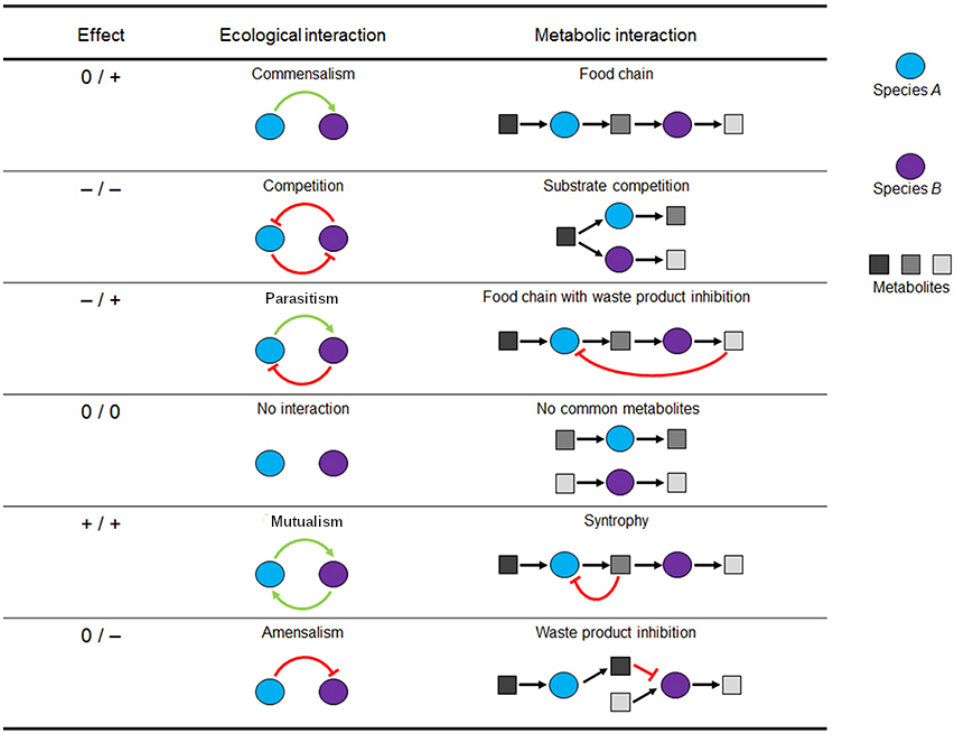


FIGURE 1.2: Microbial interaction types along with their corresponding metabolic ones. Due to certain metabolic interactions, two taxa may have a positive, a negative or a neutral effect one another. Figure based on [25]

In case of a negative interaction, can harm each other either way (**compe-tition**). That is the case between *Listeria monocytogenes* and *Lactococcus lactis* in the study of Freilich et al. where their resource competition is high enough contributing to their non-overlapping existence [26]. Moreover, similarly to commensalism, there is also the case when a taxon has a negative affect on the other without getting any harm (**amensalism**). Such is the case for *Acidithiobacillus thiooxidant* that produces sulfuric acid (H_2SO_4)

by oxidation of sulfur [27] which is responsible for lowering of pH in the culture media which inhibits the growth of most other bacteria [28]. Finally, one of the taxa may have a positive affect (host) on the other, but the latter (parasite) can be harmful to its benefactor (**parasitism**) [21]. There are multiple cases of parasitism in real-world communities; species of the genus *Bdellovibrio* for example, are parasites of other (gram-negative) bacteria [29].

However, we have a very limited understanding of such interactions and the ways that are combined to rule community-level behaviors. Thus, we cannot predict community responses to perturbations (community stability) [30]. Over the years, various methods have been used to infer such interactions. co-occurrence modells [21], time series data and causal models [31], through metabolic interactions as proxy [32]. Dynamic models, such Ordinary Differential Equations (ODEs) and the Generalized Lotka–Volterra (gLV) model [33] have been also widely used. Finally, in recent years, metabolic networks and constraint-based models have been also used to predict microbial interactions [34, 35]. This last approach allows predictions for the metabolic dynamics of the community as well as of the exact set of compounds the taxa of the community exchange [32]. Still though, microbial interactions inference is a challenging task and several questions are still open.

Apparently, the environmental conditions affect the ecological interactions to a great extent. A pair of taxa may be competitors in one case but have a neutral interaction in another one. In addition, evolutionary processes may change certain interactions; for example moving from commensalism to parasitism [36]. Both ecological and environmental interactions play a part in the composition and the functional potential of microbial assemblages. On top of that, pairwise microbial interactions can be modified by a third organism, leading to higher-order effects that influence community behaviors [37].

1.1.4 Reverse ecology: transforming ecology into a high-throughput field

For decades, *reductionism* has been the main conceptual approach in biological research [38]. Traditionally, for studies relating genetics and ecology scientists first identify an ecological adaptive phenotype and then they try to detect causal genetic variation [38]. However, as described in the previous sections, HTS data have turned the page in Biology research in numerous ways. Therefore, it is nowdays possible to *reverse* this framework and by using the genomic information retrieved, to study the ecology of a species. The **Reverse Ecology** framework uses advances in both systems biology and genomic metabolic modeling to implement community ecology studies with no a priori assumptions about the organisms under consideration [39]. Therefore, Reverse Ecology attempts to interpret HTS (genomic) data as large-scale ecological data [32].

As shown in Figure 1.3, the Reverse Ecology framework has multiple alternatives and various methods can exploit this concept. The analysis of metabolic networks (see Section **??**) plays a great part in several Reverse Ecology approaches. Most parts of this dissertation have been influenced by this, especially chapters **??**, **??** and **??**.

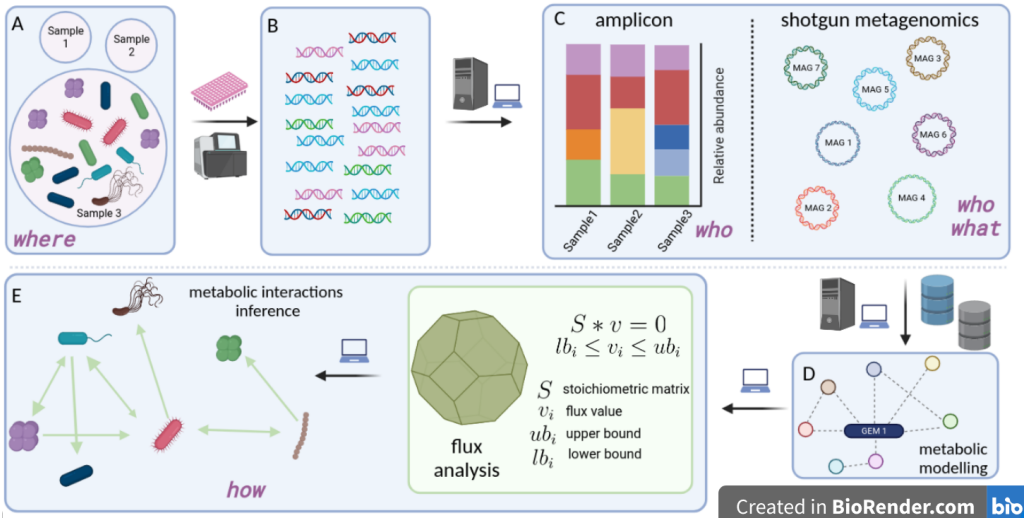


FIGURE 1.3: Without any previous knowledge of the species present in a community (A) and using HTS data (B) one can have an overview of the species present as well as in the functional profile of the community (C). Especially when the complete genome of a species has been retained (either using metagenomics (MAG) or using targeted approaches to get this (SAG)) researchers can build its corresponding GEM (D) and then infer the ecology of a taxon predicting the exogenously acquired compounds as well as ecological interactions between the taxon under study and other species present in a community (E). Both network topology - and constraint - based methods can be used to this end. Created with BioRender.com.

1.2 High Throughput Sequencing in Microbial Ecology

1.2.1 'Omics methods to access the *who* and the *what*

To discover the microbial taxa present in a sample, scientists have explored multiple ways through the years. Only a particularly limited proportion of the microbial species can be cultured [40]. Therefore, mono-cultures and enrichment cultures allow us to observe only a small fraction of the actual diversity. As a consequence, other methods for the taxonomic identification of theses species are required. Based on molecular characteristics of the microbial taxa, over the last decades, a series of methods have been developed.

Moving from single species to assemblages, molecular-based identification and functional profiling of communities has become available through marker (metabarcoding), genome (metagenomics), or transcriptome (metatranscriptomics) sequencing from environmental samples [41]. To a great extent, these methods address the problem of how to produce and get access to the information on different biological systems and molecules.

In case that the taxonomic assessment of a sample is the aim of a study, *metabarcoding* (amplicon-targeted metagenomics) and *shotgun metagenomics* can be used as alternative options. Metabarcoding studies are common, well-established, cheaper and less computationally demanding than shotgun metagenomics [42]. Its primary drawbacks are the limited information present in the short barcoding sequence and the possible taxonomic

bias arising from differential efficiency of PCR primer pairing in different species [43]. On the other hand, shotgun metagenomics offers a better taxonomic resolution at the species level by obtaining information from random sampling of virtually all genomic regions, and can address microbiome metabolic functions and entire biochemical pathways [44]. Unfortunately, it requires higher sequencing coverage and, consequently, more complex and demanding downstream bioinformatics analysis [45]. Nevertheless, it has recently been suggested that shotgun metagenomics provides a deeper characterisation of microbiome complexity that metabarcoding recently enabling to profile up to the level of strains, whose non-core genome is responsible for crucial functional differences within the same species, as the fundamental units of the community [46, 47, 48].

Targeting community composition and functional profiles in several ecological niches, microbial ecologists produce vast amount of sequencing data [49]. These approaches enable the study of ecosystems with no prior knowledge of the resident species, while at the same time a number of challenges for their management and bioinformatics analysis is rising.

1.2.2 Bioinformatics challenges in the analysis & management of HTS data

Moving from raw data to taxonomic and functional profiles of a microbial community comes with high computational costs, especially in the case of metagenome studies [50]. Sequence pre-processing, assembly, classification, and functional annotation consist of several steps the most of which a significant number of algorithms or/and software tools are available [51, 52]. Tailoring each tool's execution parameters to reflect each experiment's idiosyncrasy is vital for legitimate findings, yet it makes analyses of metagenomics data even more complex.

In addition, there are several challenges on the bioinformatics analysis *per se*. *Taxonomy assignment* in both amplicon- and shotgun metagenomics studies has several issues to meet [53]; the taxonomy of microbes is a challenge per on its own [54].

In amplicon studies, among the most major issues is the one of the abundances of the taxa found [55, 56] as well as the presence of pseudo-genes [57]. In the first case, issues such as the usually unknown number of marker gene copies per cell in the various taxa, PCR - related biases such as primer-template mismatches, length difference of amplicon, artificial base changes, chimeric molecules and library preparation - related issues such as chimera formation by the mix of amplicons from different samples makes hard for the method to have robust quantitative results [56]. Reads on the other hand resulting from pseudo-genes or/and highly divergent nuclear mitochondrial pseudo-genes (NUMTS), nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms [58], can lead either to false positive taxonomic hits or to non-hits at all, adding extra noise to the amplicon results returned.

In shotgun metagenomics studies there are also several challenges. Meta-genome *assembly* comes with a great number of challenges. Due to the uneven (and unknown) representation of the different organisms within a metagenomic mixture, simple coverage statistics can no longer be used to detect the repeats, while unrelated genomes may contain nearly-identical DNA (inter-genomic repeats) representing, for example, mobile genetic elements [59]. At the same time, *binning* is a rather tricky step too; several

algorithms have been developed to address it [60] while approaches combining the output of individual algorithms have been introduced too [61].

The vast amounts of data that come with metagenomic studies and the computational complexity for implementing multiple steps mentioned earlier imply immense computational requirements for their analysis that usually exceed the capacity of a standard personal computer [62].

For HTS data to be available to the scientific community for further exploitation, it is required to be accompanied by comprehensive metadata [63]. The potential of HTS data is revealed when they are available to the community; this way studies that could never been performed by individual researchers, labs or institutes are now possible. This way, a single researcher can now investigate how a certain environmental type reacts in response to an environmental variable by making use of hundreds of metagenomic samples that fulfill the criteria of his/her study. Finding data of interest however, can be particularly difficult. This is so because of a combination of reasons. HTS data can be particularly heterogeneous based on both the data generation and the data processing methods used. However, it is mostly the vague or even absent metadata accompanying the HTS data set several limitations in their re-usage [64].

The concept of FAIR data (Findability, Accessibility, Interoperability & Reuse) and the FAIR principles[1] along with community - driven standards and resources such as the Genomic Standards Consortium (GSC)[2], the Minimal Information about any Sequence (MIxS) [65, 66] and the National Microbiome Data Collaborative (NMDC)[3] [67] aim to address these challenges [68].

1.3 Data integration in the service of microbial ecology

1.3.1 Moving from *partial* to more *comprehensive* data interpretation

Over the last decades, based on computational and mathematical analysis and modeling, and by exploiting interdisciplinary data and knowledge, Systems Biology focuses on complex interactions within biological systems [69]. The more data becoming available from all the different levels of hierarchy of life, the more feasible for scientists to move from reductionism to more holistic approaches for interpreting how the properties of a system emerge [38].

Microbial ecology as a field would have not been the same if it was not for resources such as Integrated Microbial Genomes (IMG) and GOLD [70], SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST) [71], Pathosystems Resource Integration Center (PATRIC), [72] and many more that thousands of researches use in their every day work. All these approaches, regardless on what they focus, they are all based on data aggregation and data integration approaches. *Data aggregation* denotes the gathering of data from diverse sources in a certain scheme that will allow

[1] https://www.go-fair.org/fair-principles/
[2] https://gensc.org/
[3] https://microbiomedata.org/metadata/

them to be used as a combined data-set for further analysis [73]. In case of microbial ecology, that means that data focusing on the genetic information can be combined with phenotypical data or even with environmental and ecological data. *Data integration* on the other hand, is the process of combining everything retrieved on the data aggregation step, to get a summarization and unified view of all the accumulated data [74]. Such summarizations may lead researchers to new hypotheses that in turn, will be tested through new experiments (Figure 1.4).

Data integration comes with great challenges. Apparently, data integration methods are based on the existence of primary databases. Each of these database resources come with its own assumptions and schemas. Therefore, it is not a straight-forward task to recognize or assign and maintain the correct names of biological entities across the various databases [75]. Taxonomy is quite an indicative example. As there is no a global taxonomy system, even the species name can be a great challenge in such approaches; how to retrieve information about a species that does not have the same name on the various databases to integrate? Therefore, retrieving and mapping entities can be rather complex. Similarly to taxonomy, most biological databases are constantly changing. Thus, integration approaches need to be periodically so the always keep updated [75]. In addition to the heterogeneity of the data *per se*, further challenges that make data integration even harder in case of biological data, is the lack of unique standards [76]. In the case of HTS data, great efforts to address this challenge have been made (see Section 1.2.2).

One of the most typical examples of data integration and its potential is the STRING database[4], where multiple channels of information are combined to retrieve protein - protein interactions [77, 78]. In addition to databases of interaction experiments and others of interaction predictions, text-mining methods of the scientific literature enhance further the PPI predictions [78]. Focusing on bacterial information, BacDive[5] [79] is a great example - resource of the added value that data integration methods can provide.

Multiple integration approaches attempt to address the challenges described. The *data warehousing* approach is a widely used data integration approach and has two mains steps; first, a unified data model that can accommodate all types of information from the various source databases is schemed. Then, software is developed aiming at gathering the data from the source databases, convert them to match the unified data model and then load them into the warehouse [75]. Once these steps have been completed, further analysis of the once several bits of information - now a single data-set, can be performed. New insight may come up either from statistical analyses on the unified data-set or from their visualization [80].

[4] https://www.string-db.org/
[5] https://bacdive.dsmz.de/

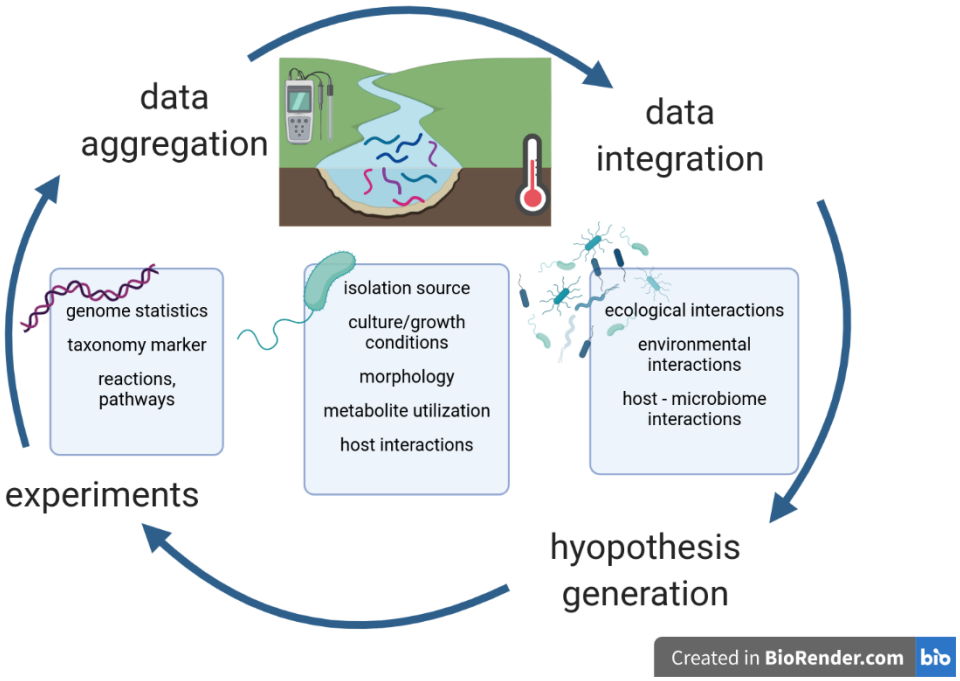


FIGURE 1.4: A data integration scheme for microbial ecology oriented data. Measurements from experiments at every level of organization of life are gathered and their summary provides researchers with new insight. Created with BioRender.com.

1.3.2 Ontologies & metadata standards: cornerstones for efficient data integration

Data integration in general, is strongly dependent by the extent that standards are used. Especially in case of vast and heterogeneous data, data integration cannot return valid results when there is not a certain way of denoting the entities included. Thus, it is dependent on the way data are distributed in the fist place as well as on whether their content follow certain principles or not. To address these challenges, several ontologies and standards have been established through the years, trying to cover all the different types of needs of the microbial ecology community.

According to Stevens et al. an *ontology* is the *"concrete form of a concepcualisation of a community's knowledge of a domain"* [81]. Ontologies attempt to capture the main concepts in a *knowledge domain*, i.e. a body of knowledge that is often associated with a specialized scientific discipline. For example, considering *where* a species live or *where* a process occurs, one need to describe the environment where the phenomenon under study takes place. Thus, the Environment Ontology (ENVO)[6] aims to provide descriptions of environments [82]. Using sets of entities, meaning entities sharing several attributes (*concepts*), descriptions of the interactions between concepts (*relations*), entities - members of a concept (*instances*) and properties of relations that aim to constrain the value a

[6] https://sites.google.com/site/environmentontology

class or an instance may get (*axioms*) aim to create an agreed vocabulary and semantic structure for exchanging information about that domain [81]. A *vocabulary* includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the [83]. Ontologies are fundamental for data integration as they ensure that the knowledge included in a text or in a data set, can be captured by both humans and computers.

Metadata are essential for most if not all types of data. Consider sampling from a set of healthy and patients. Then what if you do not know what samples are coming from each group? You data have been already degraded.

In a recent study Furner gave a list of several definitions of metadata. The one of Zeng and Qin is probably the more inclusive one: *" Structured, encoded data that describes characteristics of information-bearing entities (including individual objects, collections, or systems) to aid in the identification, discovery, assessment, management, and preservation of the described entities." Structured* are data that are highly organized and easily decipherable by machine learning algorithms while *encoded* are those that have been converted into digital signals.

Moreover, for the most efficient design and implementation but also to ensure the Interoperability of structured metadata across various computing systems and environments, a range of *standards* have been developed. *Data structure (schemas) standards* and rules for formatting the contents of metadata records along with *encoding* and *exchange* standards are combined to build up metadata.

As mentioned in Section 1.2.2, great efforts have been made on setting HTS - related metadata standards [65, 66, 67]. That is so because comprehensive metadata is the only way to ensure:

- humans will be able to contextualise where and how the data originated as well as how they were analysed
- computing systems will be able to exploit this metadata provenance further

Thus, details regarding *when*, *where* and *how* samples were collected can be provided. Moreover, these metadata may align against community developed standards where possible. For example, addressing the question of *where* a sample was collected, the answer could be *"lake"* or *ENVO:00000020*. The difference in terms of computer science is huge; it is probably trivial for a human to think that a *lake* is an aquatic environment and in fact a freshwater one. However, it is only the *relations* of an ontology that would allow a computer to come up with the same "conclusions".

Regarding the environmental metadata of a sample, ENVO [82] and MIxS [65] are working together to build a solid framework [86]. The *broad-scale environmental context* value is representing the major environmental system a sample came from; thus, *biome*[7] ENVO terms should be used as values. An ENVO *biome* term represents an ecosystem to which resident ecological communities have evolved adaptations. The *local environmental context* value, stands for entities which are in a sample's local vicinity and may have

[7] http://purl.obolibrary.org/obo/ENVO_00000428

significant causal influences on the sample; ENVO *feature* terms may be used for that. Finally, as values of the *environmental medium* category, environmental *material*[8] (one or more) immediately surrounded the sample prior to sampling. However, other resources use different schemes for describing the environment of a sample. For example in the GOLD database [87], a five-level ecosystem classification path that includes Ecosystem, Ecosystem Category, Ecosystem Type, Ecosystem Subtype and Specific Ecosystem has been adopted.

Besides the environmental metadata that describe the origin of the sample, the sequencing technology used (in case of raw data) along with metadata about the the computational steps implemented and a thorough description of the results retrieved, for example taxa found to be linked to a taxonomy scheme (in case of processed data) are required.

Figure 1.5 highlights both the potential and the challenges related to HTS - oriented metadata. Metadata describing the sample in case 1.5a are limited and neither a human nor a computer is able to capture the actual environment from where the sample was collected. In the 1.5b case, accompanying metadata are clearly more informative. Both a human and computing systems, can capture that the sample comes from an *oceanic epipelagic zone biome* (ENVO_01000035) and more specifically *oligotrophic water* (ENVO_00002223). However, two of the challenges for HTS metadata are demonstrated in this case; first, the use of *Deep Chlorophyll Maximum* denotes the need for extra terms to be added in ENVO. On top of that, the need for extra training of the community in these methods is shown as the the ENVO term denoting *oligotrophic water* should be provided as the *feature* and *Deep Chlorophyll Maximum* should be used to describe the *material*.

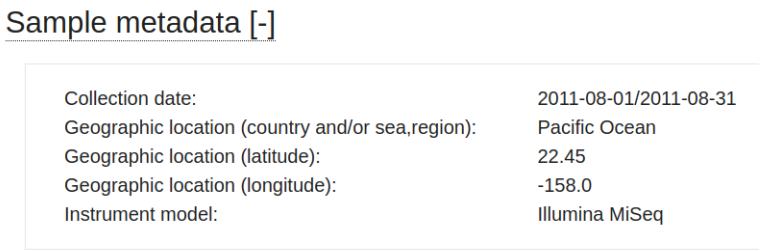
Challenges associated with metadata deposition as the one described above, mean submitters: may lack of training and outreach resulting or they do not fully realise the importance of metadata and how to comply with standards. On top of that, the non-existence of standards in many cases or the use of more than one standards lead to extra complexity. Only by a concerted effort on the part of the database providers, and with the encouragement and support of the research community, will we be able to tame the explosion of biological data [75].

1.4 Metabolic modeling: an interface for the genotype - phenotype relationship

1.4.1 Constraint-based modeling for the analysis of metabolic networks

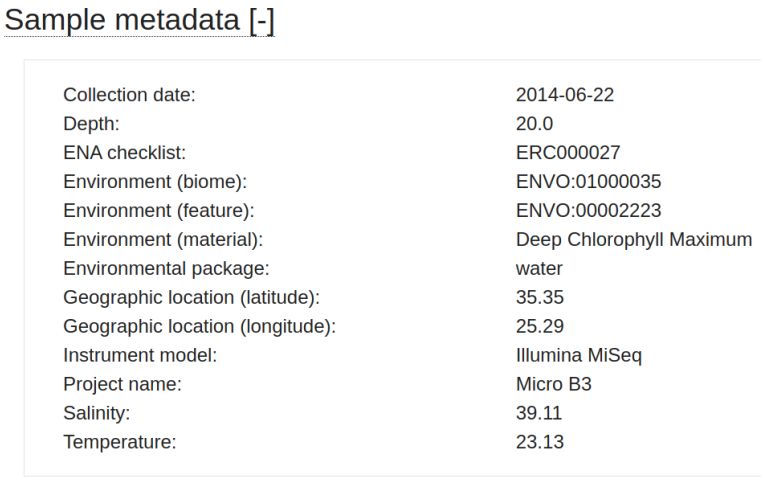
The relationship between genotype and phenotype is fundamental allowing to elucidate mechanisms that govern the physiology of a species as well as those ruling at the community level [88]. Meta- bolism penetrates most of the different levels of living entities horizontally [89] and while it reflects the genomic information it indicates what is actually going on on a cell at a certain time as a response to genetic or environmental changes [90]. One can use the *Reverse Ecology* framework (Section 1.1.4) to move all the way from genomic information to metabolism and the environment and back. To this end,

[8] http://purl.obolibrary.org/obo/ENVO_00010483


Sample metadata [-]

Collection date:	2011-08-01/2011-08-31
Geographic location (country and/or sea,region):	Pacific Ocean
Geographic location (latitude):	22.45
Geographic location (longitude):	-158.0
Instrument model:	Illumina MiSeq

(A) Poor, non machine readable metadata


Sample metadata [-]

Collection date:	2014-06-22
Depth:	20.0
ENA checklist:	ERC000027
Environment (biome):	ENVO:01000035
Environment (feature):	ENVO:00002223
Environment (material):	Deep Chlorophyll Maximum
Environmental package:	water
Geographic location (latitude):	35.35
Geographic location (longitude):	25.29
Instrument model:	Illumina MiSeq
Project name:	Micro B3
Salinity:	39.11
Temperature:	23.13

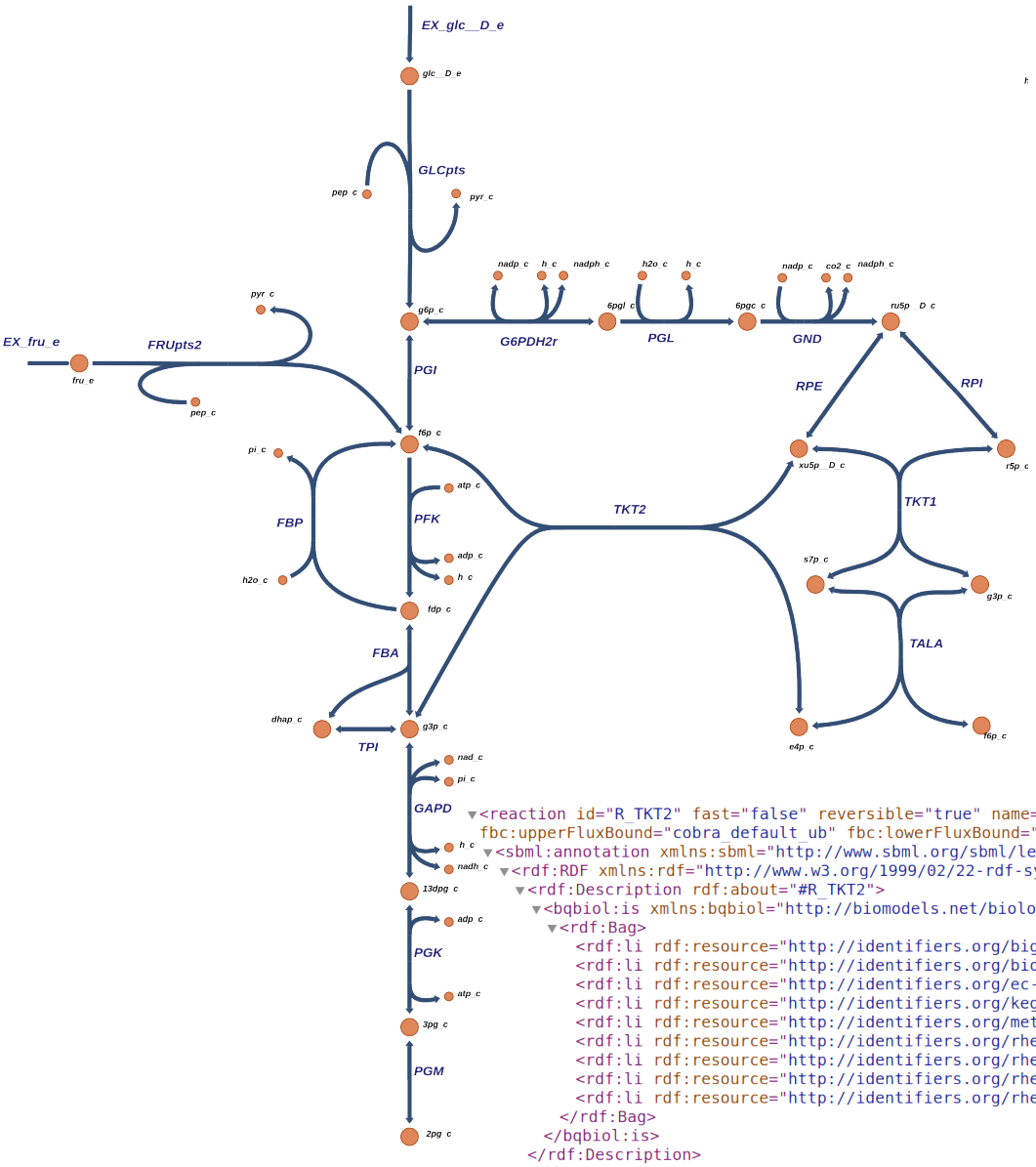
(B) Rich, partially machine readable metadata

FIGURE 1.5: Example cases of HTS - sequencing metadata. Metadata in case 1.5a fail to describe the origin of the sample both to a human and a computer. In case 1.5b further metadata have been added while most environmental metadata are provided as ENVO terms.

metabolic networks and their analysis are essential. The vast number of reaction taking place in a cell are interlinked (the product of the first acts as the substrate for the next) building up metabolic pathways, while their stoichiometry allows their mathematical representation. The rate of turnover of molecules through a metabolic reaction is called *flux*. The metabolic network of a species consists of the sum of all the reactions that take place in its cell, while *metabolic model* is its representation in a mathematical format (Figure 1.6)[9]. We call *Genome-scale metabolic models (GEMs)* incorporate the vast majority of the processes that occur in a cell or an organism in a mathematical format [91].

Once the complete genome is retrieved the enzymes and thus the potentially catalyzed by the organism reactions can be listed. However, the reconstruction of a GEM is not a straight forward task and the more the complexity of the species increases, the more effort is required for this task [92]. Thermodynamics, metabolome, physiological and labelling data as well as literature can be also integrated in such models [93].

[9]The *Escherichia coli* model of Figure 1.6 can be found at:
http://bigg.ucsd.edu/static/models/e_coli_core.xml




```
<reaction id="R_TKT2" fast="false" reversible="true" name="Transketolase" metaid="R_TKT2" sboTerm="SBO:0000176"
 fbc:upperFluxBound="cobra_default_ub" fbc:lowerFluxBound="cobra_default_lb">
  <sbml:annotation xmlns:sbml="http://www.sbml.org/sbml/level3/version1/core">
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
      <rdf:Description rdf:about="#R_TKT2">
        <bqbiol:is xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
          <rdf:Bag>
            <rdf:li rdf:resource="http://identifiers.org/bigg.reaction/TKT2"/>
            <rdf:li rdf:resource="http://identifiers.org/biocyc/META:2TRANSKETO-RXN"/>
            <rdf:li rdf:resource="http://identifiers.org/ec-code/2.2.1.1"/>
            <rdf:li rdf:resource="http://identifiers.org/kegg.reaction/R01830"/>
            <rdf:li rdf:resource="http://identifiers.org/metanetx.reaction/MNXR104869"/>
            <rdf:li rdf:resource="http://identifiers.org/rhea/27627"/>
            <rdf:li rdf:resource="http://identifiers.org/rhea/27628"/>
            <rdf:li rdf:resource="http://identifiers.org/rhea/27626"/>
            <rdf:li rdf:resource="http://identifiers.org/rhea/27629"/>
          </rdf:Bag>
        </bqbiol:is>
      </rdf:Description>
    </rdf:RDF>
  </sbml:annotation>
  <listOfReactants>
    <speciesReference species="M_e4p_c" stoichiometry="1" constant="true"/>
    <speciesReference species="M_xu5p__D_c" stoichiometry="1" constant="true"/>
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="M_f6p_c" stoichiometry="1" constant="true"/>
    <speciesReference species="M_g3p_c" stoichiometry="1" constant="true"/>
  </listOfProducts>
  <fbc:geneProductAssociation>
    <fbc:or>
      <fbc:geneProductRef fbc:geneProduct="G_b2935"/>
      <fbc:geneProductRef fbc:geneProduct="G_b2465"/>
    </fbc:or>
  </fbc:geneProductAssociation>
</reaction>
```

FIGURE 1.6: Part of the *Escherichia coli* BIGG metaboic network and the Transketolase reaction of it as integrated in the model

The analysis of GEMs has been interwoven with constraint-based modeling approachess [94]. As all compounds are finite the concentration of each metabolite is bounded [95], meaning that the models derived from the metabolic networks have constraints. Likewise, as the laws of thermodynamics need to apply in such systems, the flux of each reaction is also bounded. Therefore, the flux value of each reaction is constraint too. We call *steady state* the condition where the production rate of each metabolite equals its consumption rate [96]. Equation 1.1 represents the main concept of constraint-based modeling at a steady state.

$$\begin{aligned} S \cdot v = 0, \\ s.t. v_{lb,i} \leq v_i \leq v_{ub,i} \end{aligned} \tag{1.1}$$

where S is a $m * n$ table (m being the number of metabolites and n the number of reactions of the model) that stands for the *stoichiometric matrix* of the model. The columns of S consists of the *stoichiometric coefficients*, i.e. the number of molecules a biochemical reaction consumes and produces, of the model's reactions. $v \in \mathbb{R}^n$ is the *flux vector* that contains the fluxes of each chemical reaction of the network. As all the fluxes are bounded, for each coordinate v_i of the vector v, there are constants $v_{ub,i}$ and $v_{lb,i}$ such that $v_{lb,i} \leq v_i \leq v_{ub,i}$, for $i \in [n]$, where n is the number of reactions of the model. The *solution space* of systems such as the one of equation 1.1 "live" in a ***polytope***. Further introductory material on computational geometry can be found at Appendix A.

As discussed in [97] a great range of constraints govern the cells' operations; for a thorough overview on the constraints cells operate under, you may see [95], Chapter 16.5. (Bio)physico-chemical- (e.g., thermodynamics, nutrient uptake, oxygen availability etc.) as well as connectivity-, capacity- and rates-related constraints are applied on the functions of such a network.metabolic Each of the aforementioned constraint categories include multiple constraints, such as thermodynamics- and gene-expression-oriented constraints that add extra complexity in the model. The more constraints a model incoroporates, the more accurate the flux distributions it returns.

Using constraint-based modeling scientists can predict not only potential interactions, topology-based metabolic models are adequate for this task, but also specific metabolic dynamics in a community [32]. The most commonly used constraint-based methods for the analysis of metabolic networks are ***Flux Balance Analysis (FBA)*** [98] and *Flux Variability Analysis (FVA)* [99]. Both have been used to a great number of studies, providing fundamental isngight [100, 101]. Models estimate the minimum or the maximum of a specific (linear) ***objective function*** over the polytope. It is common for the *biomass function* of an organism to be used as the objective function. The biomass function aims at representing all metabolites needed for a cell or an organism to double. In this setting the optimization of the biomass function is like optimizing the growth of the organism itself [102]. On top of that, *dynamic FBA* approaches have tried to to study the transience of metabolism due to metabolic reprogramming [103].

1.4.2 Sampling the flux space of a metabolic model: challenges & potential

As mentioned, constraint-based approaches cover a great range of methods [94]. FBA has been proved particularly useful however, it is a *biased* method due to the selection

of the objective function. To study the global features of a metabolic network *unbiased methods* are required. On top of that, FBA is a method that addresses the question of what is the minimum or the maximum of a specific objective function, by identifying only a single optimal flux distribution. However, by construction, there is an infinite number of optimal steady states lie on a certain face of the polytope – which is also a polytope. In addition, there is no guarantee that the system under study would select the optimal steady state that FBA computes.

Using uniformly distributed steady states one could estimate the probability distribution for the flux of any reaction [104], which can lead to a deep statistical analysis of the metabolic network.

To overcome these obstacles, we sample uniformly from the set of optimal steady states and we express and quantify our uncertainty about each flux by estimating the univariate marginal probability densities [105]. Each probability density corresponds to a reaction flux. With this information at hand we can compute credible confidence intervals, estimate the average flux value, or employ other statistical methods. This procedure relies on collecting, that is sampling, a sufficient number of uniformly distributed points in the interior of the corresponding polytope.

To obtain an accurate picture of the whole solution space, once more, we sample uniformly distributed points. This way instead of a single and optimal solution, the distribution of each each reaction's flux is returned (Figure 1.7). This way, we can now investigate the properties of certain components of the whole network that potentially can lead to biological insights [95].

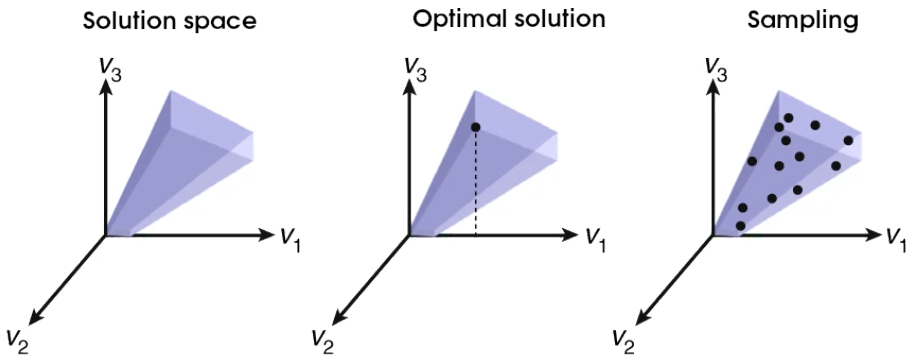


FIGURE 1.7: A visual comparison of the insight FBA ("Optimal solution") and flux sampling ("Sampling") return

Flux sampling has been proved rather valuable for a great range of applications; from design experiments and studying enzymopathies [106] to the study of metabolism under changing environmental conditions [104] and the discovery of strain-dependent differences that affect the aroma production in wine yeasts [107].

Implementations of Markov Chain Monte Carlo algorithms such as Hit-and-Run (HR) [108], the Artificial Centering Hit-and-Run (ACHR) [109] and Coordinate Hit-and-Run with Rounding (CHRR) [110] have been adopted and used to a great extent. Similarly to FBA, flux sampling algorithms, e.g. CHRR, have been integrated in cobra [111], the

most widely used software package for metabolic network analysis. Further introductory concepts of MCMC are available at Appendix **??**

On top of that, over the last few years, flux sampling has been used in computational approaches for inferring microbial interactions The Computation Of Microbial Ecosystems in Time and Space (COMETS) project and software [35] first focused on the interactions between a single species and its environment. Nowdays, metabolic modeling moves to the community level. Approaches such as the one of Diener et al. in their MICOM software [112], set a new era for the study of microbial ecology.

However, flux sampling is rather challenging from the computational point of view. The "dimensionality curse" is not a problem to a small GEM such as those of single bacterial taxa. However, to more complex species and especially in the case of community modeling the dimesion of the derived polytope can be notably high. Moreover, polytopes derived from metabolic networks are usually rather skinny, partially due to the great range the various flux values may get, making mixing hard and adding extra complexity [113, 105].

1.5 Aims and objectives

The key role of bioinformatics on microbial ecology studies was described in the previous chapters and especially when it come to HTS - oriented challenges. The potentials of addressing a subset of these challenges was also described. As HTS technologies become better and better (lower cost, higher accuracy) and HTS data become more and more available, efforts to overcome these issues are undoubtedly of great importance.

The aim of this PhD was double:

1. to enhance the analysis of microbiome data by building algorithms and software to address some of the on-going computational challenges on the field.

2. to exploit these methods to identify taxa, functions, especially related to sulfur cycle, and microbial interactions that support life in microbial community assemblages in hypersaline sediments.

All parts of this work are purely computational. Both samples and their corresponding sequencing data used in Chapter **??** have been collected and produced by Dr. Christina Pavloudi [10].

In **Chapter 2**, challenges derived from the analysis of HTS amplicon data are examined. A bioinformatics pipeline, called PEMA, for the analysis of several marker genes was developed, combinining several new technologies that allow large scale analysis of hundreds of samples. In addition, a software tool called darn, was built to investigate the unassigned sequences in amplicon data of the COI marker gene.

In **Chapter ??**, data integration, data mining and text-mining methods were exploited to build a knowledge-base, called prego, including millions of associations between:

1. microbial taxa and the environments they have been found in

[10] https://scholar.google.com/citations?user=3zs1rNkAAAAJ&hl=en&oi=sra

2. microbial taxa and biological processes they occur

3. environmental types and the biological processes that take place there

In **Chapter ??**, the challenges of flux sampling in metabolic models of high dimensions was presented along with a Multiphase Monte Carlo Sampling (MMCS) algorithm we developed.

In **Chapter ??**, sediment samples from a hypersaline swamp in Tristomo, Karpathos Greece were analysed using both amplicon and shotgun metagenomics. The taxonomic and the functional profiles of the microbial communities present there were investigated. Key metabolic processes for ensuring life at such an extreme environment were identified. Microbial interactions of the assemblages retrieved were also studied by exploiting data integration and reverse ecology approaches.

In **Chapter ??**, the history of the IMBBC-HCMR HPC facility was presented indicating the vast needs of computing resources in modern analyses in general and in microbial studies more specifically.

Finally, in **Chapter 3**, general discussion and conclusions that have derived from this research were presented.

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, and COI marker genes [1]

Citation:
Zafeiropoulos, H., Viet, H.Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C. and 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), p.giaa022,
DOI: 10.1093/gigascience/giaa022.

2.1.1 Abstract

Background: Environmental DNA and metabarcoding allow the identification of a mixture of species and launch a new era in bio- and eco-assessment. Many steps are required to obtain taxonomically assigned matrices from raw data. For most of these, a plethora of tools are available; each tool's execution parameters need to be tailored to reflect each experiment's idiosyncrasy. Adding to this complexity, the computation capacity of high-performance computing systems is frequently required for such analyses. To address the difficulties, bioinformatic pipelines need to combine state-of-the art technologies and

[1] For author contributions, please refer to the relevant section. Modified version of the published review; extra features have been added and discussed on this thesis.
You may find the Supplementary files of this study through PEMA's publication (https://academic.oup.com/gigascience/article/9/3/giaa022/5803335#supplementary-data)

algorithms with an easy to get-set-use framework, allowing researchers to tune each study. Software containerization technologies ease the sharing and running of software packages across operating systems; thus, they strongly facilitate pipeline development and usage. Likewise programming languages specialized for big data pipelines incorporate features like roll-back checkpoints and on-demand partial pipeline execution.

Findings: PEMA is a containerized assembly of key metabarcoding analysis tools that requires low effort in setting up, running, and customizing to researchers' needs. Based on third-party tools, PEMA performs read pre-processing, (molecular) operational taxonomic unit clustering, amplicon sequence variant inference, and taxonomy assignment for 16S and 18S ribosomal RNA, as well as ITS and COI marker gene data. Owing to its simplified parameterization and checkpoint support, PEMA allows users to explore alternative algorithms for specific steps of the pipeline without the need of a complete re-execution. PEMA was evaluated against both mock communities and previously published datasets and achieved results of comparable quality.

Conclusions: A high-performance computing–based approach was used to develop PEMA; however, it can be used in personal computers as well. PEMA's time-efficient performance and good results will allow it to be used for accurate environmental DNA metabarcoding analysis, thus enhancing the applicability of next-generation biodiversity assessment studies.

2.1.2 Introduction

Environmental DNA (eDNA) metabarcoding inaugurates a new era in bio- and eco-monitoring [114]. eDNA refers to genetic material obtained directly from environmental samples (soil, sediment, water, etc.) without any obvious signs of biological source material [115]. Metabarcoding is the combination of DNA taxonomy, based on taxa-specific marker genes (e.g., 16S ribosomal RNA [rRNA] for Bacteria and Archaea, cytochrome oxidase subunit 1 [COI] and 18S rRNA for Metazoa, ITS for Fungi), and high-throughput DNA sequencing technologies; thus, simultaneous identification of a mixture of organisms is attainable [116]. eDNA metabarcoding attempts to turn the page on the way biodiversity is perceived and monitored [116]. This combination is considered to be a potential holistic approach that, once standardized, allows for higher detection capacity and at a lower cost compared to conventional methods of biodiversity assessment. However, from the raw read sequence files to an amplicon study's results, the bioinformatics analysis required can be troublesome for many researchers.

Well-established pipelines are available to process metabarcoding data for the case of 16S and 18S rRNA marker genes and bacterial communities (e.g., mothur [117], QIIME 2 [118], LotuS [119]). However, certain limitations accompany each of these and occasionally they can be far from easy-to-use software. Moreover, there is a great need for similarly straightforward and benchmarked approaches for the analysis of other marker genes. With respect to the COI and ITS marker genes, a number of pipelines have been implemented, e.g., Barque [2], ScreenForBio [120], and PIPITS [121]. However, there is still

[2] https://github.com/enormandeau/barque

need for a fast, flexible, easy-to-install, and easy-to-use pipeline for both COI and ITS marker genes.

The pipelines mentioned above, although entrenched, are still hindered by a series of hurdles. Among the most prominent are technical difficulties in installation and use, strict limitations in setting parameters for the algorithms invoked, and incompetence in partial re-execution of an analysis.

Moreover, given the computational demands of such analyses, access to high - performance computing (HPC) systems might be mandatory, e.g., to process studies with a large number of samples. This is timely given the ongoing investment of national and international efforts (e.g., see European Strategy Forum on Research Infrastructures [3]) to serve the broad biological community via commonly accessible infrastructures.

2.1.3 Contribution

PEMA (Pipeline for Environmental DNA Metabarcoding Analysis) is an open source pipeline that bundles state-of-the-art bioinformatic tools for all necessary steps of amplicon analysis and aims to address the aforementioned issues. It is designed for paired-end sequencing studies and is implemented in the BDS [122] programming language. BDS's ad hoc task parallelism and task synchronization supports heavyweight computation, which PEMA inherits. In addition, BDS supports "checkpoint" files that can be used for partial re-execution and crash recovery of the pipeline. PEMA builds on this feature to serve tool and parameter exploratory customization for optimal metabarcoding analysis fine tuning. Switching effortlessly between (molecular) operational taxonomic unit ([M]OTU) clustering and amplicon sequence variant (ASV) inference algorithms is a pertinent example. Finally, via software containerization technologies such as Docker [123] and Singularity [124], with the latter being HPC-centered, PEMA is distributed in an easy to download and install fashion on a range of systems, from regular computers to cloud or HPC environments.

From the biological perspective, monitoring biodiversity at all its different levels is of great importance. Because there is not a single marker gene to detect all taxa, researchers need to use different genes targeting each great taxonomy group separately [125]. To that end, PEMA supports the metabarcoding analysis of both prokaryotic communities, based on the 16S rRNA marker gene, and eukaryotic ones, based on the ITS (for Fungi) and COI and 18S rRNA (for Metazoa) marker genes [125].

As high-throughput sequencing (HTS) data become more and more accurate, ASVs, i.e., marker gene amplified sequence reads that differ in ≥ 1 nucleotide from each other, become easier to resolve [126]. The use of ASVs instead of OTUs has been suggested [126]; however, the choice of which approach to use should be based on each study's objective(s) [127].

PEMA supports both OTU clustering and ASV inference for all marker genes (see "OTU clustering vs ASV inference" in the "Results and Discussion" section). Two clustering algorithms, VSEARCH [128] and CROP [129], are used for the clustering of reads in

[3] https://www.esfri.eu/sites/default/files/u4/ESFRI_SCRIPTA_VOL3_INNO_double_page.pdf

(M)OTUs—the former for the case of the 16S/18S rRNA marker genes, the latter for the case of COI and ITS. Swarm v2 [130] allows ASV inference in all cases.

Taxonomic assignment is performed in an alignment-based approach, making use of the CREST LCAClassifier [131] and the Silva database [132] for the case of 16S and 18S rRNA marker genes; the Unite database [133] is used for the ITS gene. In the 16S marker gene case, phylogeny-based assignment is also supported, based on RAxML-ng [134], EPA-ng [135], and Silva [132]. For the COI marker gene, the RDPClassifier [136] and the MIDORI database [137] are used for the taxonomic assignment. In addition, ecological and phylogenetic analysis are facilitated via the phyloseq R package [138].

All the pipeline- and third-party module–controlling parameters are defined in a plain "parameter-value pair" text file. Its straightforward format eases the analysis fine tuning, complementary to the aforementioned checkpoint mechanism. A tutorial about PEMA and installation guidance can be found on PEMA's GitHub repository [4].

2.1.4 Methods & Implementation

PEMA's architecture comprises 4 main parts taking place in tandem (Figure 2.1). A detailed description of the tools invoked by PEMA and their licenses is included in Additional File 1: Supplementary Methods.

Part 1: Quality control and pre-processing of raw data

First, FastQC [139] is used to obtain an overall read-quality summary; visual inspection of each sample's quality may recommend removing those insufficient quality, as well as samples with a low number of reads, and rerunning the analysis. To correct errors produced by the sequencer, PEMA incorporates a number of tools. Trimmomatic [140] implements a series of trimming steps, which either remove parts of the sequences corresponding to the adapters or the primers, trim and crop parts of the reads, or even remove a read completely, when it fails to reach the quality-filtering standards set by the user. Cutadapt [141] is used additionally for the case of ITS to address the variability in length of this marker gene (see Additional File 1: Supplementary Methods). BayesHammer [142], an algorithm of the SPAdes assembly toolkit [143], revises incorrectly called bases. PANDAseq [144] assembles the overlapping paired-end reads, and then the obiuniq program of OBITools [145] groups all the identical sequences in every sample, keeping track of their abundances. The VSEARCH package [128] is then invoked for chimera removal; however, if the Swarm v2 algorithm is selected, this step will be performed after the ASV inference (see next section).

Part 2: (M)OTU clustering and ASV inference

Quality-controlled and processed sequences are subsequently clustered into (M)OTUs or treated as input for inferring ASVs. For the case of 16S and 18S rRNA marker genes, VSEARCH [128] is used for OTU clustering, while ASVs can be identified by the Swarm v2 algorithm [130]. VSEARCH is an accurate and fast tool that can handle large datasets; at

[4] https://github.com/hariszaf/pema

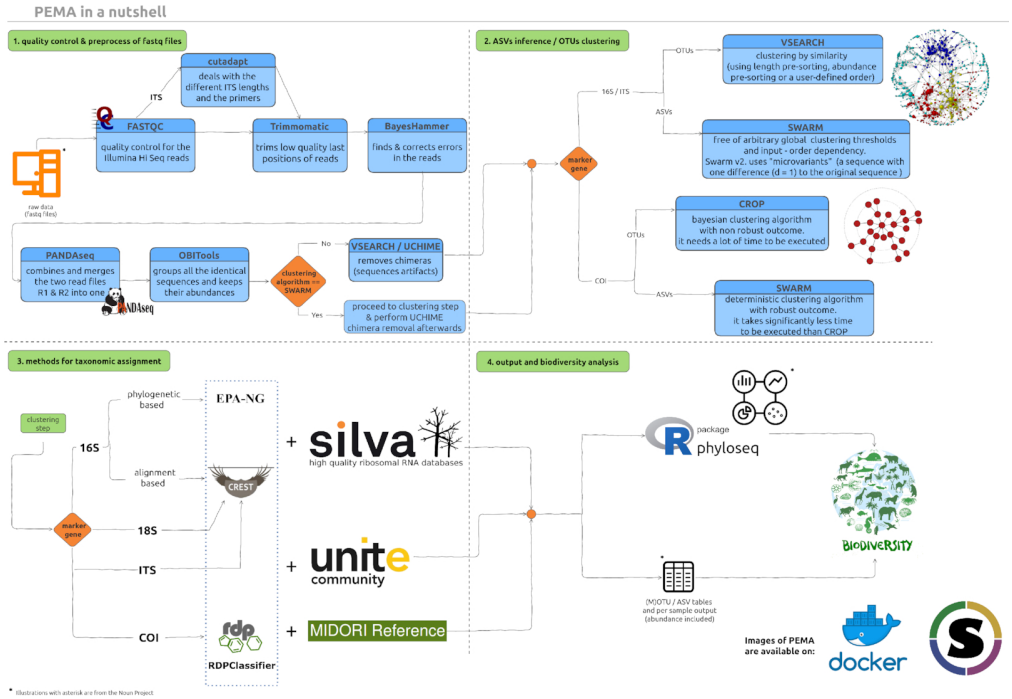


FIGURE 2.1: PEMA comprises 4 parts. The first step (top left) is the quality control and pre-processing of the Illumina sequencing reads. This step is common for both 16S rRNA and COI marker genes. The second step (top right) is the clustering of reads to (M)OTUs or their inferring to ASVs. The third step (bottom left) is the taxonomy assignment to the generated (M)OTUs/ASVs. In the fourth step (bottom right), the results of the metabarcoding analysis are provided to the user and visualized. *noun project icons by: ProSymbols (US), IconMark (PH), Nithinan Tatah (TH). clustering figure adapted from DOI: 10.7717/peerj.1420/fig-1

the same time it is a great alternative for USEARCH [146] because it is distributed under an open source license.

For the ITS and COI marker genes, CROP [129], an unsupervised probabilistic Bayesian clustering algorithm that models the clustering process using birth-death Markov chain Monte Carlo (MCMC), is used. The CROP clustering algorithm is adjusted by a series of parameters that need to be tuned by the user (namely, b, e, and z). These parameters depend on specific dataset properties such as the length and the number of reads. PEMA automatically adjusts b, e, and z by collecting such information and applying the CROP recommended parameter-setting rules [129]. ASV inference is conducted by Swarm v2 [130] in this case too.

Because the Swarm v2 algorithm is not affected by chimeras (F. Mahé, personal communication), when Swarm v2 is selected, chimera removal occurs after the clustering (see Additional File 1: Supplementary Methods: Swarm v2). This leads to a computational time gain as chimeras are sought among ASVs, instead of ungrouped reads.

Last, any singletons, i.e., sequences with only 1 read, occurring after the (M)OTU

clustering or the ASV inference may be removed according to the user's parameter settings.

Part 3: Taxonomy assignment

Alignment-based taxonomy assignment is supported for all marker gene analyses. In the case of the 16S/18S rRNA and ITS marker genes, the LCAClassifier algorithm of the CREST set of resources and tools [20] is used together with the Silva [132] and the Unite [147] database, respectively, to assign taxonomy to the OTUs. Two versions of Silva are included in PEMA: 128 (29 September 2016) and 132 (13 December 2017). Because classifiers need first to be trained for each database they use, for future Silva [132] versions new PEMA versions will be available.

For the COI marker gene, PEMA uses the RDPClassifier [136] and the MIDORI reference database [137] to assign taxonomy of the MOTUs. The MIDORI database contains quality-controlled metazoan mitochondrial gene sequences from GenBank [148].

Intended primarily for studies from less explored environments, phylogeny - based assignment is available for 16S rRNA marker gene data. PEMA maps OTUs to a custom reference tree of 1,000 Silva-derived consensus sequences (created using RAxML-ng [134] and gappa [phat algorithm] [149], Figure 2.2A). PaPaRa [150] and EPA-ng [135] combine the OTU clustering output and the reference tree to produce a phylogeny-aware alignment and map the 16S rRNA OTUs to the custom reference tree. Beyond the context of PEMA, users may visualize the output with tree viewers such as iTOL [151] (Figure 2.2B).

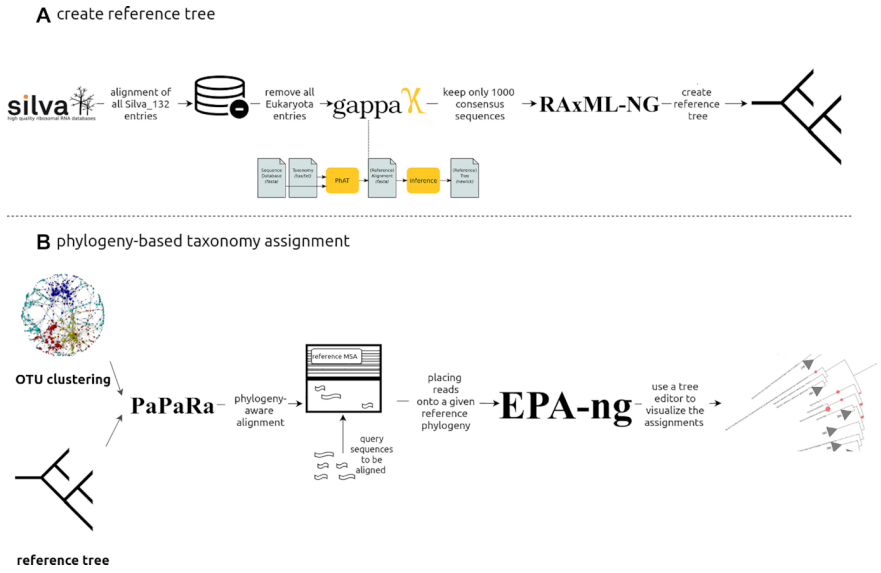


FIGURE 2.2: Phylogeny - based taxonomy assignment. A: Building a reference tree for the phylogeny-based taxonomy assignment to 16S rRNA marker gene OTUs: from the latest edition of Silva SSU, all entries referring to Bacteria and Archaea were used and using the "art" algorithm, 10,000 consensus taxa were kept. B: Using PaPaRa and the OTUs that come up from every analysis, an MSA was made and EPA - ng took over the phylogeny - based taxonomy assignment. *noun project icons by: Rockicon and A Beale.

Part 4: Ecological downstream analysis of the taxonomically assigned (M)OTU/ASV tables

PEMA's major output is either an (M)OTU or an ASV table with the assigned taxonomies and the abundances of each taxon in every sample. For each sample of the analysis, a subfolder containing statistics about the quality of its reads, as well as the taxonomies and their abundances, is also returned.

Via the phyloseq R package [138], downstream ecological analysis of the taxonomically assigned OTUs or ASVs is supported. This includes $\alpha-$ and $\beta-$diversity analysis, taxonomic composition, statistical comparisons, and calculation of correlations between samples.

When selected, in addition to the phyloseq [138] output, a multiple sequence alignment (MSA) and a phylogenetic tree of the OTU/ASVs retrieved can be returned; for the MSA, the MAFFT [152, 153] aligner is invoked while the latter is built by RAxML-ng [134].

PEMA container-based installation

An easy way of installing PEMA is via its containers. A Dockerized PEMA version is available [5]. Singularity users can *pull* the PEMA image from as described in PEMA GitHub repository [6]. Between the 2 containers, the Singularity-based one is recommended for HPC environments owing to Singularity's improved security and file accessing properties, see here [7]. PEMA can also be found in the bio.tools (id: PEMA) and SciCruch (PEMA, RRID:SCR_017676) databases. For detailed documentation, see here [8].

PEMA output

All PEMA - related files (i.e., intermediate files, final output, checkpoint files, and per - analysis parameters) are grouped in distinct (self - explanatory) subfolders per major PEMA pipeline step. In the last subfolder, i.e., subfolder 8, the results are further split into folders per sample. This eases further analysis both within the PEMA framework (e.g., partial re-execution for parameter exploration) and beyond. An extra subfolder is created when an ecological analysis via the phyloseq package has been selected.

2.1.5 Results & Validation

Evaluation

To evaluate PEMA, 2 approaches were followed. First, PEMA was benchmarked against mock community datasets. Second, PEMA was used to analyse previously published datasets. PEMA's output was then compared with the original study outcome, as well as with the output of QIIME2, LotuS, Mothur, and Barque (where applicable).

Four mock communities, 1 for each marker gene, were used. With respect to the 16S rRNA marker gene, a mock community of Gohl et al. [154] with 20 different bacterial

[5] https://hub.docker.com/r/hariszaf/pema
[6] https://github.com/hariszaf/pema
[7] https://dev.to/grokcode/singularity–a-docker-for-hpc-environments-i6p
[8] https://hariszaf.github.io/pema_documentation/

species was studied. Correspondingly, in the case of the 18S rRNA marker gene, a mock community of Bradley et al. [155] with 12 algal species was used; for the ITS, one of Bakker [156] including 19 different fungal taxa; and for the case of the COI marker gene, a mock community of Bista et al. [157] containing 14 metazoan species. More information on the mock communities, their original studies, and the results of PEMA for various combinations of parameters can be found in Additional File 2: Mock Communities.

Complementary to the mock community evaluation, 2 publicly available datasets from published studies were investigated through PEMA. For the 16S rRNA marker gene, the dataset reported by Pavloudi et al. [158] was used; the original study aimed at investigating the sediment prokaryotic diversity along a transect river–lagoon–open sea. For the COI case, the dataset of Bista et al. [159] was used; this study investigated whether eDNA can be used for the accurate detection of chironomids (a taxonomic group of macroinvertebrates) in a freshwater habitat.

In both approaches, the respective .fastq files were downloaded from the European Nucleotide Archive (ENA) of the European Bioinformatics Institute ENA-(EBI) using *ENA File Downloader version 1.2* [160] and PEMA was run on the in-house HPC cluster.

All analyses were conducted on identical Dell M630 nodes (128 GB RAM, 20 physical Intel Xeon 2.60 GHz cores).

Mock community evaluation

PEMA was tested against mock communities. An evaluation of its accuracy must capture (i) how many of PEMA's predictions are true (i.e., the percent of correctly assigned taxa among all predicted taxa) and (ii) how many of the taxa existing in the mock community were recovered successfully by PEMA. The precision statistical metric was used to assess the former, and recall, the latter. In addition, the $F1$-score was used as a combined metric of both precision and recall. Precision is calculated as the ratio of true-positive results (TP) over the total number of true- (TP) and false-positive results (FP) predicted by a model, as follows: $precision = TP/(TP+FP)$; recall is the ratio of TP over the total number of TP and false-negative results (FN): $recall = TP/(TP+FN)$. The $F1$-score is the precision and recall harmonic mean and is calculated by means of the following formula: $F1 = 2 \times (precision \times recall)/(precision + recall)$ [161].

Adequate accuracy was achieved when PEMA was used to recover the marker gene – specific mock communities at the genus level. Precision and recall scores of ~80% or more were observed, with 2 exceptions in precision but also 3 very high scores in recall. Overall the F1-scores ranged from 74% to 86%. A detailed description of the benchmark methodology and statistics analysis is given in Additional File 2: Mock Communities.

Detailed presentation of per-marker-gene–specific mock community recovery via PEMA is provided in the following sections. Several different sets of parameters were chosen for each marker gene. Each marker gene has special features (e.g., length variability, sequence variability), and each Illumina run has its own intrinsic biases (e.g., primers used, PCR protocol); thus, parameter tuning plays a crucial part in metabarcoding analyses.

In an attempt to thoroughly analyse the sequence data from the mock communities, various sets of parameters were tested on the basis of the experimental details of the published studies but also in an exploratory way. Many different parameter settings were

Marker gene	Precision	Recall	F1
16S rRNA	0.81	0.85	0.83
18S rRNA	0.75	0.90	0.82
ITS	0.79	0.94	0.86
COI	0.62	0.93	0.74

TABLE 2.1: Summary benchmark of PEMA marker - gene – specific mock community recovery (precision)

tested, especially for the steps of quality trimming of the reads and the OTU clustering/ASV inference. The differences in their output indicate how sensitive this method is, as well as the great need of a mock community in every metabarcoding study—both as a control but also as a *tuning system* for the parameter setting of the pipeline used.

16S rRNA

When PEMA was performed with the Swarm v2 algorithm (d = 3, strictness = 0.6) without removal of singletons, 18 of the 20 taxa were identified to the genus level and 3 of these even to the species level. There were 2 species that were not found in any of the PEMA runs. According to Gohl et al. [154], there was a discrepancy in the identification of those 2 species that was dependent on the amplification protocol used. It is worth mentioning that as d increases, taxa cannot be identified to species level at all; however, FP assignments decrease. Thus, when d = 30 and strictness = 0.6 for the KAPA samples, Enterococcus was not identified at all; however, PEMA finds its greatest $F1$ value (at the genus level, see Table 2.1) as the FP assignments returned are minimized. When PEMA was run using the VSEARCH clustering algorithm, high precision values were returned in all cases (>0.79). However, the recall values were decreased when using Swarm v2 (0.65–0.68).

18S rRNA

When PEMA was performed using the Swarm v2 algorithm (d = 1, strictness = 0.5), 3 of 12 community members were identified to species level (*Isochrysis galbana*, *Nannochloropsis oculata*, and *Thalassiosira pseudonana*), 6 to genus, and the remaining 3 to class; the latter were all the green algae species (Chlorophyta) of the mock community. However, a better $F1$-score (0.82) was achieved when the class of Chlorophyceae was not found at all (d = 1, strictness = 0.3) because the FPs were decreased to only 1. When the VSEARCH algorithm was used, *I. galbana* was identified only to the genus level, the *Nannochloropsis* to the order level (Eustigmatales), and the *Poterioochromonas* genus to its class (Chrysophyceae).

ITS

When PEMA was performed using the Swarm v2 algorithm (d = 20) and targeting the ITS2 region, ASVs from 5 of the 19 species of the mock community were assigned to species

level, 10 to genus, 2 to family, and 2 to class level. Contrary to the study by Bakker [156], PEMA identified the genus Chytriomyces in all 3 samples, as well as the Ustilaginaceae family. Only 1 FP assignment was recorded. When the CROP algorithm was used, PEMA's output was less accurate; the *Fusarium* species contained in the mock community were not identified further than their family (Nectriaceae). As mentioned by Bakker [156], many reads deriving from the *Fusarium spp.* were not assigned to species level because of the quality-trimming step. In addition, a manually assembled reference database for the taxonomy assignment was used in the initial study, containing only sequences of the mock community species, which biased this step, making the results not directly comparable to our case.

COI

When PEMA was performed on the Bista et al. dataset [157] and using Swarm v2 ($d = 10$), it identified 12 of the 14 species included in the mock community. The sole non - identified species were *Bithynia leachii* and *Anisus vortex*. For *B. leachii* no entry exists in the MIDORI database, version MIDORI_LONGEST_1.1. However, the existence of another species of the genus *Bithynia* was recorded. With respect to *A. vortex*, PEMA returned a high abundance ASV assigned to the *Anisus* genus but with a low confidence level. PEMA managed to identify all the members of the mock community. This includes *Physa fontinalis*, which was originally not designed to be a member of the mock community but, as Bista et al. [157] explain, was recorded owing to cross - contamination. In the case of the COI marker gene, unique sequences with low abundances (singletons or doubletons) often lead to spurious MOTUs/ASVs. Thus, as shown in Additional File 2: Mock Communities, the FP assignments are decreased when these low-abundant sequences are removed; also, the abundance of the assignments (i.e., read counts) retrieved can indicate FP assignments. Thus, TP assignments occur in greater abundance, with hundreds or even thousands of reads—contrary to most of the FP results, whose abundance is < 10 read counts. That is mostly for the case of the COI marker gene because eukaryotes are under study; eukaryotes have a great number of copies of this marker gene — different numbers of copies among the different species — and not just a single one as is almost always the case in bacteria. Therefore, assignments with such low abundances should be doubted as TP results in analyses on real datasets.

Comparison with existing software

PEMA's features were compared with those of mothur [117], QIIME 2 [118], LotuS [119] and Barque. Table 2.2 presents a detailed comparison among the 4 tools' features in terms of marker gene support, diversity and phylogeny analysis capability, parameter setting and mode of execution, operation system availability, and HPC suitability. As shown, PEMA is equally feature - rich, if not richer in certain feature categories, compared with the other software packages. In particular, PEMA's support for COI marker gene studies is distinctive; 2 methods for taxonomy assignment are supported, and PEMA's easy parameter setting, step - by - step execution, and container distribution render it user and analysis friendly.

Feature	LotuS	QIIME 2	mothur	Barque	PEMA
16S rRNA	✓	✓	✓		✓
18S rRNA	✓	✓	✓		✓
ITS	✓	✓			✓
COI				✓	✓
diversity indices		✓	✓		✓
alignment-based taxonomy assignment	✓	✓	✓	✓	✓
phylogenetic-based taxonomy assignment	✓	✓			✓
parameters assigned in the command line	✓	✓	✓		
parameters assigned through a text file	✓			✓	✓
step-by-step execution	✓	✓	✓		✓
all steps in one go possible	✓			✓	✓
available for any Operating System (Linux, OSX, Windows)		✓	✓		✓
traditional application installation	✓	✓	✓	✓	
available as a virtual machine		✓			
available as a container		✓			✓
available for HPC as a container (Singularity container)					✓

TABLE 2.2: Comparison of the basic features of the different pipelines

Evaluation on real datasets and against other tools

In the following sections, a comparative study on real datasets of the 16S rRNA and COI marker genes is presented. Analyses using PEMA and the pipelines mentioned above that support each of these 2 marker genes were performed, both with multiple sets of parameters. It is typical for pipelines to invoke a variety of established tools. In many cases, a number of tools are common among different pipelines. Therefore, it is important to stress that such comparisons should not be taken into account strictly; declaring that one pipeline is better than another is not trivial. Potentials and limitations of both the pipelines and the metabarcoding method, as well as the importance of the role of the pipeline user, are underlined in the following sections.

16S rRNA marker gene analysis evaluation

To evaluate PEMA's performance, a comparative analysis of the Pavloudi et al. [158] dataset with mothur [117], QIIME 2 [118], LotuS [119] and PEMA was conducted.

		QIIME 2				
Parameter	**LotuS**	**mothur**	**Deblur**	**DADA2**	**PEMA**	**Pavloudi et al. [158]**
No. of OTUs	9,849	142,669	517	1,023	6,028	7,050
Execution time (h)	~9	~67*	2.5	~5	~1.5	~26

TABLE 2.3: OTU predictions and execution time for the different pipelines.
* ~ 56 if the reference database is already built

It is known that the choice of parameters affects the output of each analysis; therefore, it is expected that different user choices might distort the derived outputs. For this reason and for a direct comparison of the pipelines, we have included all the commands and parameters chosen in the framework of this study in Additional File 1: Supplementary Methods. The results of the processing of the sequences by PEMA are presented in Table S1. All analyses were conducted on identical Dell M630 nodes (128 GB RAM, 20 physical Intel Xeon 2.60 GHz cores). LotuS, mothur, and QIIME 2 operated in a single-thread (core) fashion. PEMA, given the BDS intrinsic parallelization [122], operated with up to the maximum number of node cores (in this case 20).

The execution time and the reported OTU number of each tool are presented in Table 2.3. LotuS and PEMA resulted in a final number of OTUs comparable to that of Pavloudi et. al [158]. Clearly, owing to PEMA's parallel execution support, the analysis time can be significantly reduced (~ 1.5 hours in this case). The execution time depends on the parameters chosen for each software (see Additional File 1: Supplementary Methods).

Owing to the non - full overlap of the sequence reads, mothur resulted in an inflated number of OTUs; thus, it was excluded from further analyses. The results of all the pipelines were analysed with the phyloseq script that is provided with PEMA. The taxonomic assignment of the PEMA - retrieved OTUs is shown in Figure 2.3. The phyla that were found in the samples are similar to the ones that were found in the original study [158]. Although the lowest number of OTUs was found in the marine station (Kal) (Supplementary Table S3), which is not in accordance with Pavloudi et. al [158], the general trend of a decreasing number of OTUs with increasing salinity was observed as in the original study (Supplementary Figure S1). Notably, this result was not observed with the other tested pipelines (Supplementary Table S3). Furthermore, each of the pipelines resulted in a different taxonomic profile (Supplementary Figures S2–S4), with an extreme case of missing the order of Betaproteobacteriales (Supplementary Figures S5–S7).

Moreover, when the PERMANOVA analysis was run for the results of PEMA, LotuS, and DADA2, it was clear that the microbial community composition was significantly different in each of the 3 sampled habitats (i.e., river, lagoon, open sea) (PERMANOVA: F.Model = 7.0718, $P < 0.001$; F.Model = 6.5901, $P < 0.001$; F.Model = 2.2484, $P < 0.05$, respectively), which is in accordance with Pavloudi et al. [158]. However, this was not the case with Deblur (PERMANOVA: $P > 0.05$). Overall, PEMA's output is in accordance with the original study [158], and seen through this perspective PEMA performed equally well with the other tested pipelines, along with having the shortest execution time.

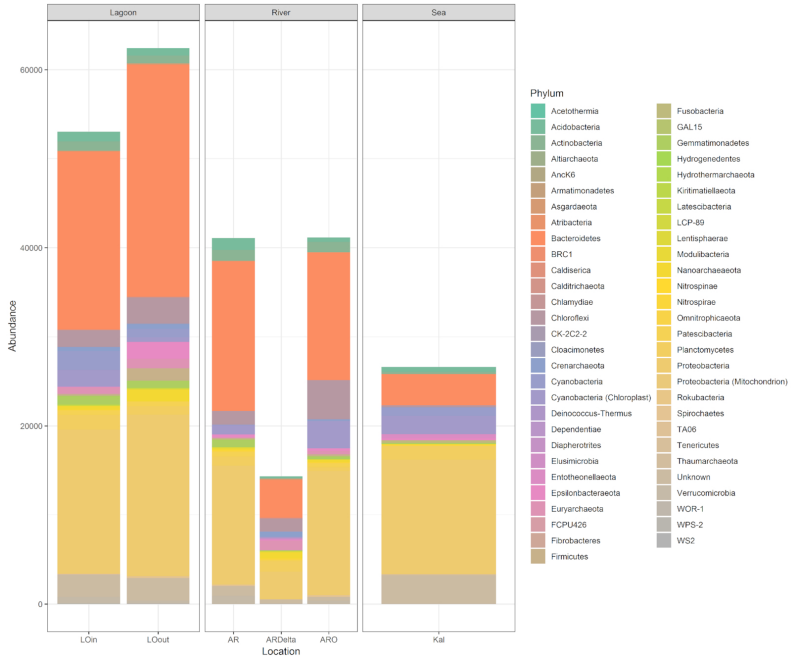


FIGURE 2.3: OTU bar plot at the phylum level. Bar plot depicting the taxonomy of the retrieved OTUs from PEMA for the dataset of Pavloudi et al. [158], at the phylum level for the case of the 16S marker gene. AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; LOin: Logarou station inside the lagoon; LOout: Logarou station in the channel connecting the lagoon to the gulf; Kal: Kalamitsi.

COI marker gene analysis evaluation

Bista et al. [159] created 2 COI libraries of different sizes: COIS (235 - bp amplicon size) and COIF (658 - bp amplicon size). The sequencing reads of COIS were selected for PEMA's evaluation; the COIF sequencing read pairs had no overlap so as to be merged and therefore were not considered appropriate for the analysis.

As previously, PEMA's performance was evaluated through a comparative analysis of the Bista et al. [159] dataset with Barque [9]; the commands and parameters chosen can be found in Additional File 1: Supplementary Methods. Regarding the creation of the MOTU table, in the Bista et al. [159] study VSEARCH [128] was used with a clustering at 97% similarity threshold. Afterwards, the BLAST+ (megablast) algorithm [162] was used against a manually created database including all NCBI GenBank COI sequences of length > 100 bp (June 2015) while excluding environmental sequences and higher taxonomic level information [159]. As discussed in the publication, this approach resulted in 138 unique MOTUs of which 73 were assigned to species level. For PEMA's evaluation, the chosen clustering algorithm was Swarm v2, using different options for the cluster radius (d) parameter (Table 2.4); according to Mahé et al. [130], this is the most important parameter because it affects the number of MOTUs that are being created. The resulting

[9] https://github.com/enormandeau/barque

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

TABLE 2.4: PEMA'sa output and execution time; PEMA's output and execution time (using a 20-core node) for different values of Swarm's d parameter.

MOTUs were classified against the MIDORI reference database [137] using RDPClassifier [136]. The results of the processing of the sequences are reported in Supplementary Table S3. For the case of Barque, the BOLD Database was used [163].

As shown in Table 2.4, PEMA resulted in 83 species-level MOTUs with a cluster radius (d) of 2, which is similar to the findings of the published study (i.e., 73 species). Although both the clustering algorithm and the taxonomy assignment methods were different between the original [159] and the present study, the results regarding the number of unique species present in the samples are in agreement to a considerable extent.

The computational time required by PEMA for the completion of the analysis is also reported in Table 2.4. Regardless of the value of the d parameter, all analyses were completed in ~ 2 hours, i.e., fast enough to allow parameter testing and customization. Regarding Barque, the analysis resulted in the identification of 51 species-level MOTUs and was concluded in 15 minutes. This difference is due to the error correction step of PEMA (BayesHammer algorithm [142]), which plays an important part in the enhanced results that PEMA returns, but it also requires a certain computational time; Barque does not have an analogous step, and therefore its overall execution time is shorter.

PEMA performed better than Barque at identifying taxa that were included in the positive control contents of the published study (Table 2.5).

2.1.6 Discussion

OTU clustering vs ASV inference

There is an ongoing discussion about whether ASVs exceed OTUs. The strongest argument to this end is that ASVs are real biological sequences. Hence, they can be compared between different studies in a straightforward way; considered as consistent labels. In comparison, de novo OTUs are constructed, or "clustered," with respect to the emergent features of each specific dataset. Therefore, OTUs defined in 2 different datasets cannot be directly compared.

However, the OTU concept is not compulsorily related to the clustering approach; it is widely used to describe results based on its biological meaning but it does not imply clustering. In addition, according to Callahan et al. [126], "ASV methods infer the biological sequences in the sample prior to the introduction of amplification and sequencing

Barque	**PEMA**	**Bista et al. [50]**
*Ablabesmyia monilis**	*Ablabesmyia monilis**	*Ablabesmyia monilis*
	*Crangonyx pseudogracilis**	*Crangonyx pseudogracilis*
	*Radix sp.**	*Radix sp.*
	Chironomidae sp.*	Chironomidae sp.
	*Ancylus sp.***	*Ancylus fluviatilis*
	*Athripsodes aterrimus, Athripsodes cinereus***	*Athripsodes albifrons*
*Chironomus anthracinus***	*Chironomus sp., Chironomus anthracinus, Chironomus pseudothummi, Chironomus riparius***	*Chironomus tentans*
*Polypedilum sordens***		*Polypedilum nubeculosum*
*Athripsodes aterrimus***		*Athripsodes albifrons*

TABLE 2.5: Comparison of the taxonomy of retrieved MOTUs among PEMA, Barque, and the positive controls of Bista et al. [159] ; * Taxonomies identical to the published study (species level), ** Taxonomies identical to the published study (genus level).

errors, and distinguish sequence variants differing by as little as one nucleotide." As a result, ASVs could be considered as OTUs of higher resolution.

It is due to this concept confusion that algorithms whose rationale is considerably closer to the variant-based approach are still considered as OTU clustering algorithms [126]. Swarm v2 produces all possible *microvariants* of an amplicon to implement an exact-string comparison [130]. Furthermore, real biological sequences, *clouds of microvariants* are produced as its output, which can be used for comparisons between different studies. Thus, Swarm v2 can be considered as an ASV-inferring algorithm.

Traditional clustering methods have certain limitations such as arbitrary global clustering thresholds and centroid selection because they depend on the input order and are time-consuming, etc. [164], which variant-based approaches manage to address. However certain algorithms for OTU clustering such as VSEARCH have been proven to be especially reliable, and they are widely used by many researchers. Furthermore, ASVs intend to improve taxonomic resolution; however, a vast number of inferred ASVs (see here [10] for more) can lead to inflation of diversity estimates, especially in the case of microbial communities, thus making the analysis even more complicated.

ASV or OTU approaches are supported by PEMA, although we have found that similar ecological results are produced by both these methods, as also suggested by Glassman and Martiny [165].

[10] http://fiererlab.org/2017/05/02/lumping-versus-splitting-is-it-time-for-microbial-ecologists-to-abandon-otus/

Beyond environmental ecology, ongoing and future work

PEMA is mainly intended to support eDNA metabarcoding analysis and be directly applicable to next - generation biodiversity / ecological assessment studies. Given that community composition analysis may also serve additional research fields, e.g., microbial pathology, the potential impact of such pipelines is expected to be much higher. Ongoing PEMA work focuses on serving a wide scientific audience and on making it applicable to more types of studies. The easy set - up and execution of PEMA allows users to work closely with national and European HPC / e - infrastructures (e.g., ELIXIR Greece [11], LifeWatch ERIC, [12] EMBRC ERIC [13]). To that end and in a mid - term perspective, a CWL version of PEMA will be explored. The aim of this effort is to reach out to a wider scientific audience and address both their ongoing as well as future analysis needs.

By supporting the analysis of the most commonly used marker genes for Bacteria and Archaea (16S rRNA), Fungi (ITS), and Metazoa (COI/18S rRNA), a holistic biodiversity assessment approach is now possible through PEMA and eDNA metabarcoding; although, from a mid-term perspective, it is our intention to allow ad hoc and in - house databases to be used as reference for the taxonomy assignment.

Conclusions

PEMA is an accurate, execution - friendly and fast pipeline for eDNA metabarcoding analysis. It provides a per - sample analysis output, different taxonomy assignment methods, and graphics - based biodiversity / ecological analysis. This way, in addition to (M)OTU/ASV calling, it provides users with both an informative study overview and detailed result snapshots.

Thanks to a nominal number of installation and execution commands required for PEMA to be set and run, it is considered essentially user friendly. In addition, PEMA's strategic choice of a single parameter file, implementation programming language, and multiple container - type distribution grant it speed (running in parallel), on - demand partial pipeline enactment, and provision for HPC - system – based sharing.

All the aforementioned features render PEMA attractive for biodiversity / ecological assessment analyses. By supporting the analysis of the most commonly used marker genes for Prokaryotes (Bacteria and Archaea), as well as Eukaryotes (Fungi and Metazoa), PEMA allows assessment of biodiversity in different levels of biodiversity. Applications may mainly concern environmental ecology, with possible extensions to such fields as microbial pathology and gut microbiome, in line with modern research needs, from low volume to big data.

2.1.7 PEMA modules added after its publication

Since its publication, PEMA has been under continuous development and testing. Custom databases can be now used to train both classifiers used in the PEMA framework, thus the

[11] https://www.elixir-greece.org/
[12] https://www.elixir-greece.org/
[13] http://www.embrc.eu

taxonomy assignment step is not limited by the reference databases included on PEMA. With the release of the v.2.1.3 [14] version, PEMA was re-architectured completely aiming at an easier way for people to contribute. On top of that, several modules have been added, mostly in an attempt to address requests from users and e-infrustructures (e.g., LifeWatch ERIC IJI [15]). On its current version (v.2.1.5) it now supports the analysis of one extra marker gene, the 12S rRNA gene, by exploiting the 12S Vertebrate Classifier v2.0.0-ref database [166]. For the case of 18S rRNA marker gene, the PR2 database [167] was integrated so now the user may select between Silva and PR2, while Silva v.138 has been also added. Furtheremore, thanks to the ncbi-taxonomist tool [168], PEMA now provides an extended OTU/ASV table where in the last column the NCBI Taxonomy Id for the taxonomic level closer to the species name rank for which there is one, is available. Last but not least, a new version of the parameters file has been made to provide a machine-readable version of it so the values set by the user can be parsed for potential errors in an automatic way.

[14] https://github.com/hariszaf/pema/releases/tag/v.2.1.3
[15] https://www.lifewatch.eu/internal-joint-initiative/

2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data[16]

Citation:
Zafeiropoulos, H., Gargan, L., Hintikka, S., Pavloudi, C. and Carlsson, J., 2021. The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data. Metabarcoding and Metagenomics, 5, p.e69657,
DOI: 10.3897/mbmg.5.69657

2.2.1 Abstract

The mitochondrial cytochrome C oxidase subunit I gene (COI) is commonly used in environmental DNA (eDNA) metabarcoding studies, especially for assessing metazoan diversity. Yet, a great number of COI operational taxonomic units (OTUs) or/and amplicon sequence variants (ASVs) retrieved from such studies do not get a taxonomic assignment with a reference sequence. To assess and investigate such sequences, we have developed the Dark mAtteR iNvestigator (DARN) software tool. For this purpose, a reference COI-oriented phylogenetic tree was built from 1,593 consensus sequences covering all the three domains of life. With respect to eukaryotes, consensus sequences at the family level were constructed from 183,330 sequences retrieved from the Midori reference 2 database, which represented 70% of the initial number of reference sequences. Similarly, sequences from 431 bacterial and 15 archaeal taxa at the family level (29% and 1% of the initial number of reference sequences respectively) were retrieved from the BOLD and the PFam databases. DARN makes use of this phylogenetic tree to investigate COI pre-processed sequences of amplicon samples to provide both a tabular and a graphical overview of their phylogenetic assignments. To evaluate DARN, both environmental and bulk metabarcoding samples from different aquatic environments using various primer sets were analysed. We demonstrate that a large proportion of non-target prokaryotic organisms, such as bacteria and archaea, are also amplified in eDNA samples and we suggest prokaryotic COI sequences to be included in the reference databases used for the taxonomy assignment to allow for further analyses of dark matter. DARN source code is available on GitHub at https://github.com/hariszaf/darn and as a Docker image at https://hub.docker.com/r/hariszaf/darn.

2.2.2 Introduction

Metabarcoding: concept and caveats

DNA metabarcoding is a rapidly evolving method that is being more frequently employed i n a range of fields, such as biodiversity, biomonitoring, molecular ecology and others [169, 170]. Environmental DNA (eDNA) metabarcoding, targeting DNA directly isolated from environmental samples (e.g., water, soil or sediment, [171]), is considered a holistic approach (Stat et al. 2017) in terms of biodiversity assessment, providing high detection

[16] For author contributions and supplementary material please refer to the relevant sections. Modified version of the published review.

capacity. At the same time, it allows wide-scale rapid bio-assessment [172] at a relatively low cost as compared to traditional biodiversity survey methods [116].

The underlying idea of the method is to take advantage of genetic markers, i.e. marker loci, using primers anchored in conserved regions. These universal markers should have enough sequence variability to allow distinction among related taxa and be flanked by conserved regions allowing for universal or semi-universal primer design [173]. In the case of eukaryotes, the target is most commonly mitochondrial due to higher copy numbers than nuclear DNA and the potential for species level identification. Furthermore, mitochondria are nearly universally present in eukaryotic organisms, especially in case of metazoa, and can be easily sequenced and used for identification of the species composition of a sample [174]. However, it is essential that comprehensive public databases containing well curated, up-to-date sequences from voucher specimens are available [175]. This way, sequences generated by universal primers can be compared with the ones in reference databases, assessing sample OTU composition. The taxonomy assignment step of the eDNA metabarcoding method and thus, the identification via DNA-barcoding, is only as good and accurate as the reference databases [176].

Nevertheless, there is not a truly "universal" genetic marker that is capable of being amplified for all species across different taxa [177]. Different markers have been used for different taxonomic groups [169]. While bacterial and archaeal diversity is often based on the 16S rRNA gene, for eukaryotes a diverse set of loci is used from the analogous eukaryotic rRNA gene array (e.g., ITS, 18S or 28S rRNA), chloroplast genes (for plants) and mitochondrial DNA (for eukaryotes) in an attempt for species - specific resolution [125]. The mitochondrial cytochrome c oxidase subunit I (COI) marker gene has been widely used for the barcoding of the Animalia kingdom for almost two decades [178]. There are cases where COI has been the standard marker for metabarcoding, such as in the assessment of freshwater macroinvertebrates [179] even though not all taxonomic groups can be differentiated to the species level using this locus [169]; for example, in case of fish other loci are widely used such as 12S rRNA gene (hereafter referred to as 12S rRNA) [180].

The COI locus

The mitochondrial cytochrome c oxidase subunit I (also called cox1 or/and COI) is a gene fragment of 700 bp, widely used for metazoan diversity assessment. Here we present some of the reasons that microbial eukaryotes and prokaryotes are also amplified in such studies, raising the issue of the known unknown sequences. COI is a fundamental part of the heme aa3-type mitochondrial cytochrome c oxidase complex: the terminal electron acceptor in the respiratory chain. Even if aa3-type Cox have been found in bacteria, there are also other cytochrome c oxidase (Cox) groups, such as the cbb3-type cytochrome c oxidases (cbb3-Cox) and the cytochrome ba3 [181, 182].

Furthermore, the presence of highly divergent nuclear mitochondrial pseudogenes (numts) has been a widely known issue on the use of COI in barcoding and metabarcoding studies, leading to overestimates of the number of taxa present in a sample [57]. Numts are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms [58].

Thus, as Mioduchowska et al. (2018) [183] highlight, when universal primers are used targeting the COI locus, it is possible to co-amplify both non-target numts and prokaryotes [184]. This has led to multiple erroneous DNA barcoding cases and it is now not rare to encounter bacterial sequences described as metazoan in databases such as GenBank [183].

Even though there are various known issues [173], COI is indeed considered as the "gold standard" for community DNA metabarcoding of bulk metazoan samples [185]; bulk is an environmental sample containing mainly organisms from the taxonomic group under study providing high quality and quantity of DNA [186]. However, as highlighted in the same study, this is not the case for eDNA samples. As Stat et al. (2017) [172] state, in the case of eDNA samples, the target region for metazoa is found in general at considerably lower concentrations compared to those from prokaryotes because most primers targeting the COI region amplify large proportions of prokaryotes at the same time [187, 188, 189]. Cold-adapted marine gammaproteobacteria are an indicative example for this case as shown by Siddall et al. (2009) [184].

2.2.3 Contribution

The co-amplification of prokaryotes explained above, is a major reason for why many Operational Taxonomic Units (OTUs) and/or Amplicon Sequence Variants (ASVs) in eDNA metabarcoding studies cannot get taxonomy assignments when metazoan reference databases are used (c.f. Aylagas et al. 2016 [190]) or they are assigned to metazoan taxa but with very low confidence estimates. Despite the presence of such OTUs/ASVs to a varying degree in metabarcoding studies using the COI marker gene [184], to the best of our knowledge, there has not been a thorough investigation of the origin for these sequences. Although unassignable sequences could be informative, there have been few attempts to further investigate this dark matter (e.g., [191, 192]).

The aim of this study was to build a framework for extracting such 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) [193]. We argue that the vast majority of these sequences represent microbial taxa, such as bacteria and archaea.

More specifically, based on the previously described methodology by Barbera et al. (2019) [135] (see also full stack example of the EPA-ng algorithm) for large-scale phylogenetic placements, we built a framework to estimate to what extent the OTUs/ASVs retrieved in an environmental sample represent target taxa or not. That is, to evaluate the taxonomy assignment step in a metabarcoding analysis, by checking the phylogenetic placement of dark matter sequences. Similar studies have provided great insight into other marker genes, e.g. [194].

2.2.4 Methods & Implementation

Building the COI tree of life

Sequences for the COI region from all the three domains of life were retrieved from curated databases. Eukaryotic sequences were retrieved from the Midori reference 2 database

(version: GB239) [137]. Initially, 1,315,378 sequences were retrieved corresponding to 183,330 unique species from all eukaryotic taxa. With respect to bacteria and archaea, 3,917 bacterial COI sequences were obtained from the BOLD database [163]. Similarly, 117 sequences from archaea were obtained from BOLD. In addition, for all the PFam protein sequences related to the accession number for COX1 (PF00115 [17]), the respective DNA sequences were extracted from their corresponding genomes. This way an additional 217 archaeal and 9,154 bacterial sequences were obtained (see Table 1). In total, sequences from 15 archaeal, 371 bacterial families and 60 taxonomic groups of higher level not assigned in the family level, were gathered. An overview of the approach that was followed is presented in Figure 2.4.

The large number of obtained sequences effectively prevents a phylogenetic tree construction encompassing their total number in terms of building a single phylogenetic tree covering all of the three domains of life (archaea, bacteria, eukaryota). Therefore, consensus representative sequences from each of the three datasets were constructed using the PhAT algorithm [149]; based on the entropy of a set of sequences, PhAT groups sequences into a given target number of groups so they reflect the diversity of all the sequences in the dataset. As PhAT uses a multiple sequence alignment (MSA) as input, all the three domain-specific datasets were aligned using the MAFFT alignment software tool v7.453 [152, 153].

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

TABLE 2.6: Number of sequences and taxonomic species per domain of life and resources. The (#) symbols stands for "number".

In the case of Eukaryotes, the alignment of the corresponding sequences would be impractically long because of their large number (183K sequences). To address this challenge, a two-step procedure was followed; a sequence subset of 500 sequences (*reference set*) was selected and aligned and then used as a backbone for the alignment of all the remaining eukaryotic COI sequences. All sequences were considered reliable as they were retrieved from curated databases (Midori2 and BOLD). To build the reference set, a number (n) of the longest sequences from each of the various phyla were chosen, proportionally to the number (m) of sequences of each phylum (see Supplementary Table 2.6). The –min-tax-level parameter of the PhAT algorithm corresponded to the class level, for the case of eukaryotes and to the family level for archaea and bacteria. This parameter forced the PhAT algorithm to build at least one consensus sequence for each class and family respectively. The taxonomy level was not the same for the case of eukaryotes sequence dataset and those of bacteria and archaea, as the number of unique eukaryotic families was one order of magnitude higher. The PhAT algorithm was invoked through the gappa v0.6.1 collection of algorithms [195].

[17] http://www.ncbi.nlm.nih.gov/nuccore/PF00115

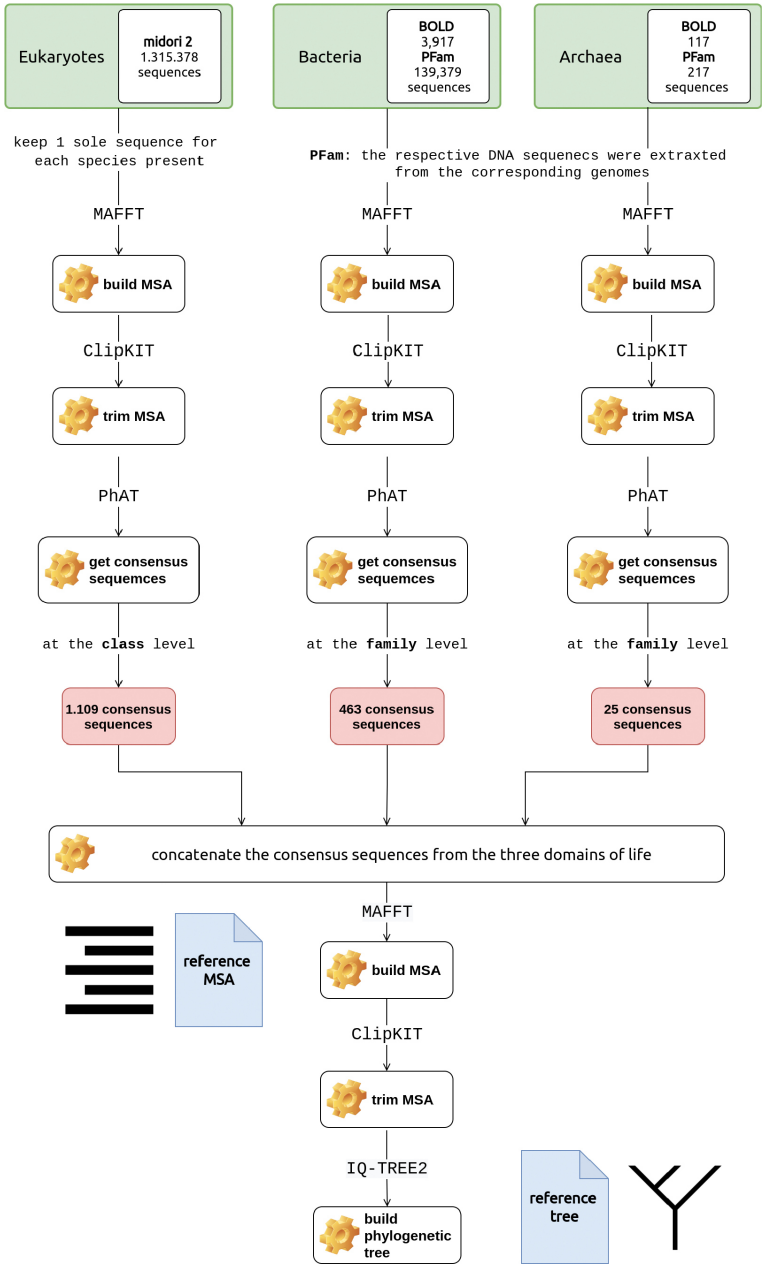


FIGURE 2.4: Overview of the approach followed to build the COI reference tree of life. Sequences were retrieved from Midori 2 (eukaryotes) and BOLD (bacteria and archaea) repositories. Consensus sequences at the family level were built for each domain specific dataset. MAFFT and consensus sequences at the family level were built using the PhAT algorithm. The COI reference tree was finally built using IQ-TREE2. Noun project icons by Arthur Slain and A. Beale.

A total of 1,109 consensus sequences (70% of total consensus sequences) were built covering the eukaryotic taxa, while 463 (29%) bacterial and 21 (1%) archaeal consensus sequences were included. The per-domain, consensus sequences returned can be found under the consensus_seqs directory on the GitHub repository (see _consensus.fasta files). These sequences were then merged as a single dataset and aligned to build a reference MSA; this time MAFFT was set to return using the –globalpair algorithm and the –maxiterate parameter equal to 1,000. The MSA returned was then trimmed with the ClipKIT software package [196] to keep only phylogenetically informative sites. The final MSA is available on GitHub, see trimmed_all_consensus_aligned_adjust_dir.aln.

The reference tree was then built based on this trimmed MSA using the IQ-TREE2 software [197, 198]. ModelFinder was invoked through IQ-TREE2 and the GTR+F+R10 model was chosen based on the Bayesian Information Criterion (BIC) among 286 models that were tested. The phylogenetic tree was then built using 1,000 bootstrap replicates (-B 1,000) and 1,000 bootstrap replicates for Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) (1,000 1000).

In the .iqtree file there are the branch support values; SH-aLRT support (%) / ultrafast bootstrap support (%).

A thorough description of all the implementation steps for building the reference tree is presented in this Google Collab Notebook [18] . The computational resources of the IMBBC High Performance Computing system, called Zorba [199], were exploited to address the needs of the tasks.

Investigating COI dark matter

The COI reference tree was subsequently used to build and implement the Dark mAtteR iNvestigator (DARN) software tool. DARN uses a .fasta file with DNA sequences as input and returns an overview of sequence assignments per domain (eukaryotes, bacteria, archaea) after placing the query sequences of the sample on the branches of the reference tree. Sequences that are not assigned to a domain are grouped as "distant". It is necessary for the input sequences to represent the proper strand of the locus, i.e. input reads should have forward orientation. Optionally, DARN invokes the orient module of the vsearch package [128] to implement this step, in case the user is not sure about the orientation of the sequences to be analysed.

The focal query sequences are aligned with respect to the reference MSA using the PaPaRa 2.0 algorithm [150]. The query sequences are then split to build a discrete query MSA. Finally, the Evolutionary Placement Algorithm EPA-ng [135] is used to assign the query sequences to the reference tree.

To visualise the query sequence assignments, a two-step method was developed. First, DARN invokes the gappa examine assign tool which taxonomically assigns placed query sequences by making use of the likelihood weight ratio (LWR) that was assigned to this exact taxonomic path. In the DARN framework, by making use of the –per-query-results and –best-hit flags, the gappa assign software assigns the LWR of each placement of the query sequences to a taxonomic rank that was built based on the taxonomies included in

[18] https://colab.research.google.com/drive/1XorHsBm1uqx5TTZsH7SeVRkUA2SS8dnY

the reference tree. The first flag ensures that the gappa assign tool will return a tabular file containing one assignment profile per input query while the latter will only return the assignment with the highest LWR. DARN automatically parses this output of gappa assign to build two input Krona profile files based on

- the LWR values of each query sequence and
- an adjustive approach where all the best hits get the same value in a binary approach (presence - absence)

In the final_outcome directory that DARN creates, two .html files, one for each of the Krona plots; Krona plots are built using the ktImportText command of KronaTools [200]. In addition four .fasta files are generated including the sequences of the sample that have been assigned to each domain or as "distant". A .json file with the metadata of the analysis is also returned including the identities of the sequences assigned to each domain.

DARN also runs the gappa assign tool with the –per-query-results flag only. This way, the user can have a thorough overview of each sample's sequence assignments, as a sequence may be assigned to more than one branch of the reference tree, sometimes even to different domains. However, in cases with sequences assigned to multiple branches, the likelihood scores are most typically up to 100-fold to 1000-fold different.

DARN source code as well as all data sequences and scripts for building the reference phylogenetic tree are available on GitHub [19].

2.2.5 Results & Validation

Evaluation of the phylogenetic tree

The inferred phylogenetic tree is shown in Figure 2.5, with the bacterial (light blue) and archaeal (dark green) branches highlighted; in Supplementary material 3: Figure S1 the distribution of the eukaryotic phyla on the tree is presented. As shown, bacteria and archaea can be distinguished from eukaryotes. Scattered bacterial branches that are present among eukaryotic ones represent the diversity of the COI locus. To evaluate the phylogenetic tree, the set of consensus sequences were placed on it using the EPA-ng algorithm. The placements (see .jplace through a phylogenetic tree viewer, e.g. iTOL) verified that the phylogenetic tree built is valid, as the consensus sequences have been placed in their corresponding taxonomic branches (Supplementary material 4: Figure S2; the figure was built using the heat-tree module of the gappa examine tool).

[19] https://github.com/hariszaf/darn

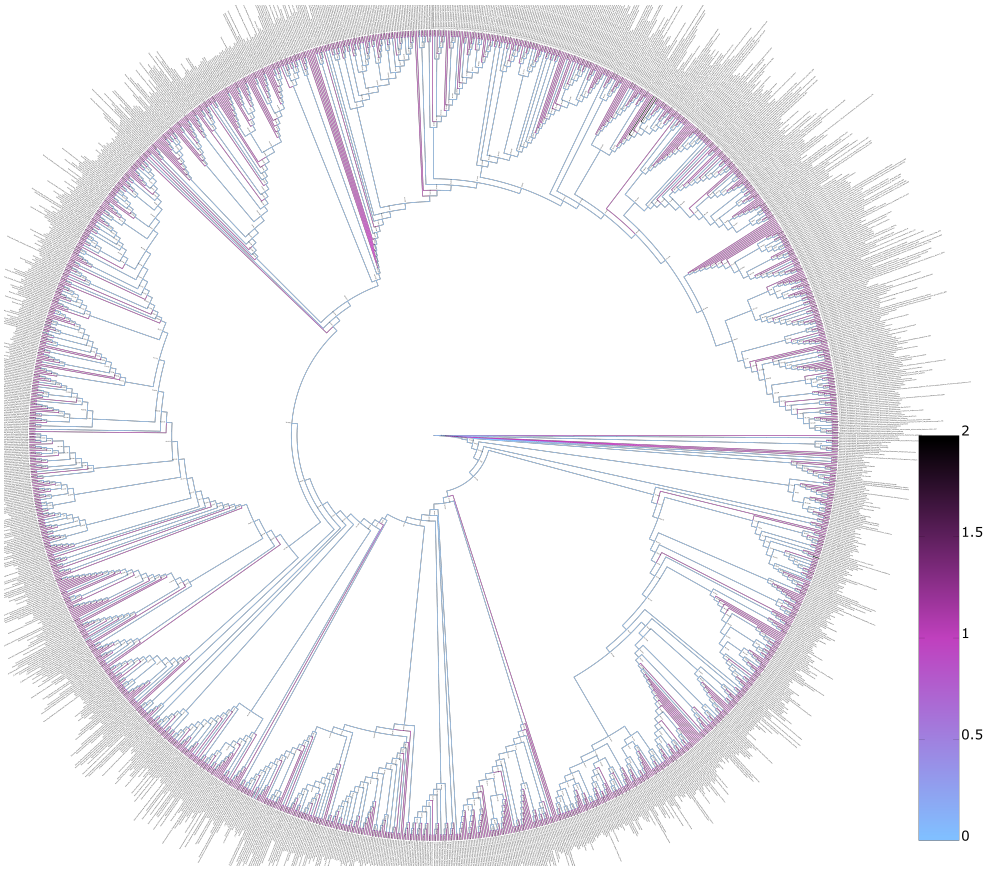


FIGURE 2.5: Phylogenetic tree of the consensus sequences retrieved; the tree that DARN makes use of. Light blue: bacterial branches. Dark green: archaeal branches. White: eukaryotic branches.

DARN using mock community data

To examine whether the phylogenetic-based taxonomy assignment addresses a real-world issue, a local blast database was built using the total number of the consensus sequences retrieved. As expected, when the consensus sequences were blasted against this local blastdb, all were matched with their corresponding sequences. However, when a mock dataset was used to evaluate the two approaches (blastdb and the phylogenetic tree) none of the bacterial sequences were captured as bacteria after blastn against the local blastdb (see output file here [20]). All bacterial sequences returned an incorrect eukaryotic assignment. Contrarily, when the phylogenetic tree was used, all the bacterial sequences were captured.

[20] https://github.com/hariszaf/darn/blob/pfam/evaluation/consensus_blast_assignments.txt

DARN using real community data

To evaluate DARN on the presence of dark matter we analysed a wide range of cases to show the ability of DARN to detect and estimate dark matter under various conditions. Both eDNA and bulk samples, from marine, lotic and lentic environments, were selected to reflect various combinations of primer and amplicon lengths, PCR protocols and bioinformatics analyses (Table 2.7).

More specifically, 57 marine, surface water, eDNA samples from Ireland were analysed through a. QIIME2 [118] and DADA2 [201] and, b. PEMA [202]. Similarly, 18 mangrove and 18 reef marine eDNA samples from Honduras, were analyzed using a. JAMP v0.74 [21] and DnoisE [203] and b. PEMA Furthermore, a sediment sample and two samples from Autonomous Reef Monitoring Structures (ARMS) one conserved in DMSO and another in ethanol from the Obst et al. (2020) [204] dataset were analysed using PEMA. In addition, one lotic and two lentic samples from Norway were analysed using PEMA. For the case of the lentic samples, multiple parameter sets regarding the ASVs inference step were implemented; i.e the d parameter of the Swarm v2 [130] that PEMA invokes was set equal to 2 and 10 to cover a great range of different cases [205]. DARN was then executed using the ASVs retrieved in each case as input. All the DARN analyses and the PEMA runs were performed on an Intel(R) Xeon(R) CPU E5649 @ 2.53GHz server of 24 CPUs and 142 GB RAM in the Area52 Research Group at the University College Dublin.

The number of sequences returned, using various bioinformatic analyses, ranged from circa 3k to 214k (Table 2.7) in the different amplicon datasets used. A coherent visual representation of the DARN outcome for all the datasets is available here[22]. The visual and interactive properties of the Krona plot allow the user to navigate through the taxonomy. Furthermore, DARN also supports a thorough investigation per OTU/ASV, as it returns a .json file with all the OTUs/ASVs ids that have been assigned in each of the four categories (Bacteria, Archaea, Eukaryotes and distant).

Significant proportions of non-eukaryote DARN assignments were observed in all marine eDNA samples (Table 2.7). Bacterial assignments made up the largest proportion of the non-eukaryotic assignments (35.3% on average and more than 75% of the OTUs/ASVs in some cases), however, archaeal assignments were also detected to a great extent as well (18.4% on average). The lentic samples were those with the shortest amplicon length among those analysed (142 bp); hence, for their orientation a database with only the shortest consensus sequences (< 700 bp) was used, as otherwise a great number of sequences did not have sufficient number of hits and was discarded (see Suppl. material 2: Table S2). It is worth mentioning that in this case, the initial number of raw reads ranged from 53,000 (ERS6488992, ERS6488993) to 88,000 (ERS6488993) while the number of ASVs returned (using Swarm with d parameter equal to 10) ranged from 365 (ERS6488993) to 823 (ERS6488993). This relatively low number of ASVs could indicate that targeting such small COI regions could decrease the co-amplification of non-targeted sequences. In the case of bulk samples (Table 2.7) only a low proportion of the sequences were not assigned as Eukaryotes, suggesting that non-eukaryotic sequences are more abundant in environmental samples. This could be expected since prokaryotes are amplified as whole

[21] https://github.com/VascoElbrecht/JAMP
[22] https://hariszaf.github.io/darn/

Sample(s) accession number	Sample type	Primer set	Amplicon length (bp)	Bioinfo pipeline(s)	# of ASVs	~% of sequence assignments per domain (if PEMA[a])			
						Eukaryotes	Bacteria	Archaea	distant
ERS6449795-ERS6449829	eDNA	jgHCO2198 - jgLCO1490 & LoboF1 - LoboR1	658	QIIME2 - Dada2	13,376	11	88.0	0.02	0.003
				PEMA	39,454	25	75.0	0.1	0.4
ERS6463899–ERS6463901	eDNA	mlCOIintF - jgHCO2198	313	JAMP dada2 PEAR vsearch DnoisE	1,304	35	65.0	-	0.2
ERS6463906–ERS6463911 ERS6463913–ERS6463918 ERS6463920–ERS6463922				PEMA	11,545	46	50.0	1	3
ERS6463744–ERS6463761				JAMP dada2 PEAR vsearch DnoisE	663	40	60.0	-	0.6
				PEMA	5,879	49	47.0	1.0	2.0
ERR3460466	bulk	mlCOIintF - jgHCO2198	313	PEMA (d = 2)	193	99	1	-	-
ERR3460467	bulk				74	97	0.0	-	3
ERR3460470	eDNA				184	71	28.0	0	1
ERS6488992	eDNA	fwhF2 - EPTDr2	142	PEMA	416	85	7	3	5
ERS6488993					315	99.2	0.4	0.4	-
ERS6488994					823	90	4	2	4
ERS6488995	eDNA	BF3 - BR2	458	PEMA	1,940	64	34.0	2	0.3

TABLE 2.7: DARN outcome over the samples or set of samples. Assignment fractions of the sequences per domain per sample in the DARN results over the samples.

[a]The *d* parameter equals 10 except mentioned otherwise

organisms from environmental samples, while metazoa that are usually the targeted taxa in COI studies, are amplified from DNA traces or/and other parts of biological source material.

2.2.6 Discussion

By making use of a COI - oriented reference phylogenetic tree built from 1,593 consensus sequences, to phylogenetically place sequences from COI metabarcoding samples onto it, the surmise for including bacteria, algae, fungi etc. [187, 190] was verified. Our results demonstrate that standard metabarcoding approaches based on the COI gene region of the mitochondrial genome will not only amplify eukaryotes, but also a large proportion of non-target prokaryotic organisms, such as bacteria and archaea. Clearly, dark matter, and especially bacteria, make up a significant proportion of sequences generated in COI based eDNA metabarcoding datasets. The large proportion of prokaryotes observed in the present study is corroborated by the findings of [187]. Furthermore, dark matter seems to be particularly common in eDNA as compared to bulk samples [185]. However, it should be mentioned that the high number of prokaryotic sequences in COI metabarcoding data is also reflecting known issues with contamination [206, 207, 208], incorrectly labeled reference sequences [209] and holobionts [210, 211] in eukaryotic genomes.

As publicly available bacterial COI sequences are far too few to represent the bacterial and archaeal diversity, their reliable taxonomic identification is not currently possible. This way, bacterial, i.e. non-target, sequences that were amplified during the library preparation have at least the possibility of a taxonomy assignment. Our implementations using DARN indicate that it is essential both for global reference databases (e.g., BOLD, Midori etc) and custom reference databases which are commonly used, to also include non-eukaryotic sequences.

While our approach specifically addressed the COI gene, DARN can be adapted to analyse any locus fragment. For instance, metabarcoding of environmental samples for the 12S rRNA mitochondrial region is often employed to assess fish biodiversity [212, 180] and the approach presented here could be adjusted to allow further analyses of the 12S rRNA data. In addition, our approach can be used to identify non-target eukaryotes when the target is bacterial taxa [213].

The approaches implemented in DARN can benefit both bulk and eDNA metabarcoding studies, by allowing quality control and further investigation of the unassigned OTUs/ASVs. The approach is also adaptable to other markers than COI. Moreover, the approach presented here allows researchers to better understand the known unknowns and shed light on the dark matter of their metabarcoding sequence data.

Chapter 3

Conclusions

3.1 Microbial diversity assesment using HTS methods

Main goal of this PhD project was to address on-going challenges related to the bioinformatics analysis of HTS-oriented studies as well as to provide ways for the optimal exploitation of such data and of the current knowledge that is linked to them.

The 16S rRNA gene has been used for decades as the golden standard for the study of microbial communities. It has been shown that the full-length 16S sequence combined with appropriate treatment of the intragenomic copy variants has the potential to provide taxonomic resolution of bacterial communities even at the strain level [214]. However, when the region is chosen carefully and a thorough alignment procedure is applied, even short short reads may return phylogenetic information comparable with the one fron full-length 16S rRNA reads [215]. This was also shown in Chapter **??** as the 16S rRNA amplicon analysis was in line with the taxonomy assignment of the shotgun reads.

Even if amplicon studies have proven themselves essential for the assessment of microbial diversity, the bioinformatics analaysis in such studies, usually comes with several issues; with the lack of parameter tuning being among the most crucial ones. As shown in Chapter 2.1 where mock communities were used to validate the PEMA results, it is parameter tuning that determines the precision and recall scores in such analyses. Sequencing mock communities along with the rest of the samples allows the tuning of the bioinformatics analysis based on a known assemblage and thus, it enables parameter tuning based on the idiosyncracy of each particular experiment/study [216].

When studying a microbial community, non-prokaryotic species need to be considered too. In that case, 16S rRNA is not the best marker to use; instead, several markers have been used for different taxonomic groups. Thus, several studies aiming at the biodiversity assessment of environmental samples, make use of several markers and apparently, workflows supporting their analysis are vital. As shown in Chapter 2.1, the PEMA approach attempts to address this challenge by supporting the analysis of several markers but also by supporting the semi-automatic analysis of any marker since training of the classifiers invoked with any local database is possible.

Moreover, it is also commonly known that pseudogenes as well as nuclear mitochondrial pseudogenes (numts) can lead to several biases in such studies [57]. To address this

challenge multiple computational efforts have been implemented [217] This issue also applies for the case of Bacteria and Archaea and the 16S rRNA gene [218] even if it has been shown that bacterial pseudogenes have a great chance of being removed almost directly after their formation; so fast that to be governed by a strictly neutral model of stochastic loss [219]. As shown in Chapter 2.2, a great part of the OTUs/ASVs retrieved from COI amplicon data may actually come from bacterial and/or archaeal taxa. Such approaches need to be merged in amplicon studies as an extra quality control stop but also to enable further investigation of the unassigned OTUs/ASVs. In Chapter 2.2 is also shown the need for reference databases to also include non-target sequences so they can distinguish actual hits.

However, there is still a major question regarding the microbial diversity assessment; how could HTS methods be used to recognise novel taxa? As shown in Chapter **??**, the reconstruction of MAGs from shotgun metagenomics data may play a great role in the description of unknown and currently uncultivated taxa. Such studies and their corresponding MAGs have enriched our knowledge on the tree of life to a great extent over the last few years, uncovering several prokaryotic phyla, leading to radical challenges on their taxonomy and the taxonomy scheme [220]. Long-read sequencing technologies such as Nanopore and PacBio, have improved their accuracy to a great extent, offering high-quality, cutting-edge alternatives for testing hypotheses about microbiome structure and functioning as well as assembly of eukaryote genomes from complex environmental DNA samples [221].

3.2 Gaining insight from literature and metadata mining

biological insight: example from the paper
value of metadata
provenance

3.3 e-infrastructures can provide both capacity and reproducibility

3.4 Flux sampling can pr

1. Role of technologies such as containerization.
2. Trends for reproducible pipelines and role of infrastuctures

3.5 Future work

to understand the patterns of biodiversity found in most natural habitats, it is crucial to understand the evolution, distribution and diversity of bacterial nutri- tional preferences and metabolic strategies across the tree of life [52] [222]

As already discussed, metabolic models at the community level. to infer microbial interactions but also to study the fitness of the community.

Eco-evolutionary dynamics of complex social strategies in microbial communities [223]

By encapsulating all software and its dependencies in an isolated and easy to reinstall environment (container) containerization addresses this challenge. In addition, packaging a software per container simplifies management of the software requirements but also facilitates the creation and management of standardized workflows/pipelines. Workflow tools such as Common Workflow Language (CWL), Snakemake and Nextflow have been proven of high value in building such pipelines as they support the connection of multiple independent software. Another route of access to metagenomics analysis datasets is the Metagenome Exchange Registry which contains mappings between a number of well-established metagenome analysis platforms and their raw data in INSDC. Its aim is to aid comparison and benchmarking of tools and services as well as to help users explore metagenomics data in INSDC that are analysed by third party services. The registry is available as an API with plans to release a user interface in future.

Thus, it is fundamental for the community to make apprehend the value of this and comply to the standards []

Flux sampling as the future of microbial interaction inference along with techniques such Raman spectometry etc.

steady-state does not consider kinetics or regulatory events Integrating stoichiometric approaches with machine learning and more [224]

Acknowledgements

This dissertation has recieved funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant aggrement No 241 (PREGO project).

Appendices

Appendix A

Computational Geometry

A.1 Definitions & concepts

A *hyperplane* is a set of the form

$$H = \{x \in \mathbb{R}^n : p^T \cdot x = t\} \tag{A.1}$$

and defines two closed *halfspaces*. Such a halfspaces would be denoted as

$$\begin{aligned} H_- &= \{x \in \mathbb{R}^n : p^T \cdot x \leq t\} \\ H_+ &= \{x \in \mathbb{R}^n : p^T \cdot x \geq t\} \end{aligned} \tag{A.2}$$

The intersection of a finite number of halfspaces builds a *polyhedron*. A system of inequalities arises:

$$a_i^t \cdot x \leq b_i, i \in \{1, .., m\} \tag{A.3}$$

where m is the number of halfspaces. Thus, a polyhedron can be denoted as:

$$P = \{x \in \mathbb{R}^n : A \cdot x \leq b\} \tag{A.4}$$

where A is a $m * n$ matrix with m being the number of halfspaces and n the dimension of the space. Finally, b is a vector of the right side of the inequalities (b_i).

A *bounded* polyhedron meaning, $\exists M > 0$ such that $\|x\| \leq M$ for all $x \in P$, is called a ***polytope***.

Some of the inequalities in A though can be geometrically redundant, meaning that if these are removed P remains the same. The *dimension* of a polytope P is equal to $n - r(P)$ where $r(P)$ is the maximum number of linearly independent defining hyperplanes containing P.

We call *defining hyperplanes* the **total** hyperplanes defined from the system, meaning those coming from the $A \cdot x \leq 0$ plus those coming from any constraints. For example, maybe we would have $x \geq 0$ then these hyperplanes would be also considered as defining hyperplanes.

We consider P as a ***fully dimensional*** polytope if and only if $dim(P) = n$. In other words, a d-polytope is full-dimensional in d-space. Each (nonredundant) inequality corresponds to a facet of the polytope

In case that our system has only inequalities, then the polytope derived is always full dimensional. However, in case that extra cosntraints as equalities are included, then the polytope derived could be full-dimensional or not. If the space defined by the equalities intersects the one defined by the inequalities, then the polytope is not full-dimensional.

A *face* is a set of points $F \subseteq P$ that belongs to the intersection of a nonempty set of defining hyperplanes To show that a valid inequality is a face we just need to find a point in the intersection of the hyperplane it defines and our polytope. To show that a face is a ***facet***, i.e. a face of dimension $n-1$, we need to show that it belongs to exactly one defining hyperplane. If it belongs to more, then it is no longer a facet.

Facets are necessary and sufficient for the complete description of a polytope in terms of valid inequalities.

If P is full-dimensional then it has a unique minimal description:

$$P = \{x \in \mathbb{R}^n : a_i^T \cdot x \leq b_i, i = \{1, ..m\}\} \tag{A.5}$$

where each of the m inequalities is unique to within a positive multiple.

Points $x_1, x_2, ..., x_k \in \mathbb{R}^n$ are *affinely independent* if the $k-1$ directions: $x_i - x_1, i \in \{2, k\}$ are linearly independent. The maximum number of affinely independent points in P is denoted as $i(P)$. Now the dimension of P can be defined as: $dim(P) = i(P) - 1$

To show that P is full-dimensional we just need to show that it has exactly $n+1$ affinely independent points.

A matrix is said to have full rank if its rank equals the largest possible for a matrix of the same dimensions, which is the lesser of the number of rows and columns.

Markov Chain Monte Carlo

Definition 1. *A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event*

Markov Chain Monte Carlo (MCMC) methods are algorithms that sample from a probabilistic distribution.

Appendix B

PREGO

B.1 Mappings

PREGO produces entity identifiers either by Named Entity Recognition (NER) with the EXTRACT tagger or by mapping retrieved identifiers to the selected ones. PREGO adopted NCBI taxonomy identifiers for taxa, Environmental Ontology for environments and Gene Ontology as a structure knowledge scheme for Processes (GObp) and Molecular Functions (GOmfs). The latter was for reasons that are two-fold, first Gene Ontology has a Creative Commons Attribution 4.0 License and second there are many resources that have mapped their identifiers to Gene Ontology. MG-RAST metagenomes and JGI/IMG isolates annotations come with KEGG orthology (KO) terms; Struo-oriented genome annotations, on the other hand, have Uniprot50 ids. The mapping from KO to GOmf and Uniprot50 to GOmf is implemented via UniProtKB mapping files of their FTP server (see idmapping.dat and idmapping_selected.tab files). By using the 3-column mapping file, the initial annotations were mapped to GOmf. As a complement, a list of metabolism-oriented KEGG ORTHOLOGY (KO) terms has been built (see *prego_mappings* in the Availability of Supporting Source Codes section). Finally, as STRUO annotations refer to GTDB genomes, publicly available mappings (accessed on 24 December 2021) were used to link the genomes used with their corresponding NCBI Taxonomy entries.

B.2 Daemons

An important component PREGO approach (Figure A1) is the regular updates which keep PREGO in line with the literature and microbiology data advances. The updates are implemented with custom scripts called daemons that are executed regularly spanning from once a month up to six-month cycles. This variation occurs because of the API requirements of each web resource as well as the computational intensity of the association extraction from the retrieved data.

Each Daemon is attached to a resource because its data retrieval methods (API, FTP) and following steps, shown in Figure A1, require special handling and multiple scripts (see *prego_daemons* in the Availability of Supporting Source Codes section).

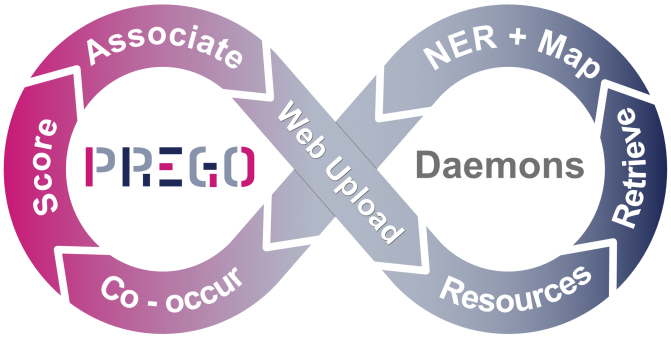


FIGURE B.1: Software daemons perform all steps of the PREGO methodology in a continuous manner similar to the Continuous Development and Continuous Integration method.

B.3 Scoring

Scoring in PREGO is used to answer the questions:

- Which associations are more thrustworthy?
- Which associations are more relevant to the user's query?

Relevant, informative, and probable associations are presented to the user through the three channels that were discussed previously. Each channel has its own scoring scheme for the associations it contains and all of them are fit in the interval $(0,5]$ to maintain consistency. The values of the score are visually shown as stars. The Genome Annotation and Isolates channel has fixed values of scores depending on the resource because Genome Annotation is straightforward, and the microbe id is known a priori. On the other hand, Environmental Samples channel data are based on samples, which contain metagenomes and OTU tables. Thus, it has two levels of organization, microbes with metadata, and sample identifiers. Each association of two entities is scored based on the number of samples they co-occur. A Literature channel scoring scheme is based on the co-mention of a pair of entities in each document, paragraph, and sentence. The differences in the nature of data require different scoring schemes in these channels. The contingency table (Table B.1) of two random variables, X and Y are the starting point for the calculation of scores. The term $X = 1$ might be a specific NCBI id and $Y = 1$ a ENVO term. The $c_{1,1}$ is the number of instances that two terms of $X = 1$ and $Y = 1$ are co-occurring, i.e., the joint frequency. The marginals are the $c_{1,.}$ and $c_{.,1}$ for x and y, respectively, which are the backgrounds for each entity type. Different handling of these frequencies leads to different measures. There is not a perfect scoring scheme, just the one that works best on a particular instance. Consequently, scoring attributes require testing different measures and their parameters.

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

TABLE B.1: Contingency table of co-occurrences between entities $X = x$ and $Y = y$. This is the basic structure for all scoring schemes. $c_{x,y}$ is the count of the co-occurrence of these entities. $c_{x,.}$ is the count of the x with all the entities of Y type (e.g., Molecular function). Conversely, $c_{.,y}$ is the count of y with all the entities of X type (e.g., taxonomy

Literature Channel

Scoring in the Literature channel is implemented as in STRING 9.1 [225] and COMPARTMENTS [226], where the text mining method uses a three-step scoring scheme. First, for each co-mention/co-occurrence between entities (e.g., Methanosarcina mazei with Sulfur carrier activity), a weighted count is calculated because of the complexity of the text.

$$c_{x,y} = \sum_{k=1}^{n} w_d \delta_{dk}(x,y) + w_p \delta_{p,k}(x,y) + w_s \delta_{sk}(x,y) \tag{B.1}$$

Different weights are used for each part of the document (k) for which both entities have been co-mentioned, $w_d = 1$ for the weight for the whole document level, $w_p = 2$ for the weight of the paragraph level, and $w_s = 0.2$ for the same sentence weight. Additionally, the delta functions are one (Equation B.1) in cases the co-mention exists, zero otherwise. Thus, the weighted count becomes higher as the entities are mentioned in the same paragraph and even higher when in the same sentence. Subsequently, the co-occurrence score is calculated as follows:

$$score_{x,y} = c_{x,y}^{a} \left(\frac{c_{x,y} c_{.,.}}{c_{x,.} c_{.,y}}\right)^{1-a} \tag{B.2}$$

where $a = 0.6$ is a weighting factor, and the $c_{x,.}$, $c_{.,.}$, $c_{.,y}$ are the weighted counts as shown in Table B.1 estimated using the same Equation B.2. This value of the weighting factor has been chosen because it has been optimized and benchmarked in various applications of text mining [34,70,71]. The value of Equation B.2 is sensitive to the increasing size of the number of documents (MEDLINE PubMed—PMC OA). Therefore, to obtain a more robust measure, the value of the score is transformed to z-score. This transformation is elaborated in detail in the COMPARTMENTS resource [226]. Finally, the confidence score is the z-score divided by two. Cases in which the scores exceed the (0,4] interval are capped to a maximum of 4 to reflect the uncertainty of the text mining pipeline.

Environmental Samples Channel

Data from environmental samples are OTU tables and metagenomes. Thus, for each entity x, the number of samples is calculated as the background and a number of samples

of the associated entity (metadata background) c.,y (see Table A1). Each association between entities x, y has a number of samples, cx,y that they co-occur. Note that each resource is independent and the scoring scheme is applied to its entities. This means that the same association can appear in multiple resources with different scores. The score is calculated with the following formula:

$$score_{x,y} = 2.0 * \sqrt{\frac{c_{x,y}}{c_{.,y}}}^{a} \tag{B.3}$$

This score is asymmetric because the denominator is the marginal of the associated entity. Thus, the score decreases as the marginal of y is increasing, i.e., the number of samples that y is found. On the other hand, it promotes associations in which the number of samples of the association are similar to the marginal of y. The exponents on the numerator and denominator equal to 0.5 and to 0.1, respectively, in order to reduce the rapid increase of score. Lastly, the value of the score is capped in the range $(0,4]$.

B.4 Bulk download

Users can also download programmatically all associations per channel through the links that are shown in Table B.2. The data are compressed to reduce the download size and md5sum files are provided as well for a sanity check of each download.

Channel	Link	md5sum	Size (in GB)
Literature	literature.tar.gz	literature.tar.gz.md5	5.4
Environmental Samples	environmental_samples.tar.gz	environmental_samples.tar.gz.md5	0.69
Annotated genomes and isolates	annotated_genomes_isolates.tar.gz	annotated_genomes_isolates.tar.gz.md5	0.26

TABLE B.2: Bulk download links and md5sum files.

Bibliography

[1] Paul G Falkowski, Tom Fenchel, and Edward F Delong. The microbial engines that drive earth's biogeochemical cycles. *science*, 320(5879):1034–1039, 2008.

[2] Jodie Belilla, David Moreira, Ludwig Jardillier, Guillaume Reboul, Karim Benzerara, José M López-García, Paola Bertolino, Ana I López-Archilla, and Purificación López-García. Hyperdiverse archaea near life limits at the polyextreme geothermal dallol area. *Nature ecology & evolution*, 3(11):1552–1561, 2019.

[3] Kenneth J Locey and Jay T Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016.

[4] What is microbial ecology? URL https://www.isme-microbes.org/what-microbial-ecology.

[5] Paul V Dunlap. Microbial diversity. 2001.

[6] MT Madigan, KS Bender, DH Buckley, WM Sattley, and DA Stahl. Brock biology of microorganisms. 15th global edition. *Boston, US: Benjamin Cummins*, 2018.

[7] Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016.

[8] Gabriel E Leventhal, Carles Boix, Urs Kuechler, Tim N Enke, Elzbieta Sliwerska, Christof Holliger, and Otto X Cordero. Strain-level diversity drives alternative community types in millimetre-scale granular biofilms. *Nature microbiology*, 3(11): 1295–1303, 2018.

[9] Maria L Marco. Defining how microorganisms benefit human health. *Microbial Biotechnology*, 14(1):35–40, 2021.

[10] Zhenjiang Xu, Daniel Malmer, Morgan GI Langille, Samuel F Way, and Rob Knight. Which is more important for classifying microbial communities: who's there or what they can do? *The ISME journal*, 8(12):2357–2359, 2014.

[11] Sylvie Estrela, Jean CC Vila, Nanxi Lu, Djordje Bajić, Maria Rebolleda-Gómez, Chang-Yu Chang, Joshua E Goldford, Alicia Sanchez-Gorostiaga, and Álvaro Sánchez. Functional attractors in microbial community assembly. *Cell Systems*, 13(1):29–42, 2022.

[12] Alexander Eng and Elhanan Borenstein. Taxa-function robustness in microbial communities. *Microbiome*, 6(1):1–19, 2018.

[13] Stilianos Louca, Martin F Polz, Florent Mazel, Michaeline BN Albright, Julie A Huber, Mary I O'Connor, Martin Ackermann, Aria S Hahn, Diane S Srivastava, Sean A Crowe, et al. Function and functional redundancy in microbial systems. *Nature ecology & evolution*, 2(6):936–943, 2018.

[14] Qing-Lin Chen, Jing Ding, Dong Zhu, Hang-Wei Hu, Manuel Delgado-Baquerizo, Yi-Bing Ma, Ji-Zheng He, and Yong-Guan Zhu. Rare microbial taxa as the major drivers of ecosystem multifunctionality in long-term fertilized soils. *Soil Biology and Biochemistry*, 141:107686, 2020.

[15] Alexandre Jousset, Christina Bienhold, Antonis Chatzinotas, Laure Gallien, Angélique Gobet, Viola Kurm, Kirsten Küsel, Matthias C Rillig, Damian W Rivett, Joana F Salles, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME journal*, 11(4):853–862, 2017.

[16] Ricardo Cavicchioli, William J Ripple, Kenneth N Timmis, Farooq Azam, Lars R Bakken, Matthew Baylis, Michael J Behrenfeld, Antje Boetius, Philip W Boyd, Aimée T Classen, et al. Scientists' warning to humanity: microorganisms and climate change. *Nature Reviews Microbiology*, 17(9):569–586, 2019.

[17] Samir Giri, Leonardo Oña, Silvio Waschina, Shraddha Shitut, Ghada Yousif, Christoph Kaleta, and Christian Kost. Metabolic dissimilarity determines the establishment of cross-feeding interactions in bacteria. *Current Biology*, 31(24): 5547–5557, 2021.

[18] Samir Giri, Shraddha Shitut, and Christian Kost. Harnessing ecological and evolutionary principles to guide the design of microbial production consortia. *Current Opinion in Biotechnology*, 62:228–238, 2020.

[19] Wentao Kong, David R Meldgin, James J Collins, and Ting Lu. Designing microbial consortia with defined social interactions. *Nature Chemical Biology*, 14(8):821–829, 2018.

[20] Raíssa Mesquita Braga, Manuella Nóbrega Dourado, and Welington Luiz Araújo. Microbial interactions: ecology in a molecular perspective. *Brazilian Journal of Microbiology*, 47:86–98, 2016.

[21] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012.

[22] Alan R Pacheco, Mauricio Moel, and Daniel Segrè. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature communications*, 10(1):1–12, 2019.

[23] Alfonso Santos-Lopez, Christopher W Marshall, Michelle R Scribner, Daniel J Snyder, and Vaughn S Cooper. Evolutionary pathways to antibiotic resistance are dependent upon environmental structure and bacterial lifestyle. *Elife*, 8:e47612, 2019.

[24] Hendrikus J Laanbroek, Marie-Josee Baar-Gilissen, and Hans L Hoogveld. Nitrite as a stimulus for ammonia-starved nitrosomonas europaea. *Applied and Environmental Microbiology*, 68(3):1454–1457, 2002.

[25] Octavio Perez-Garcia, Gavin Lear, and Naresh Singhal. Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Frontiers in microbiology*, 7:673, 2016.

[26] Shiri Freilich, Anat Kreimer, Isacc Meilijson, Uri Gophna, Roded Sharan, and Eytan Ruppin. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic acids research*, 38(12):3857–3868, 2010.

[27] Roberto A Bobadilla Fazzini, Maria Paz Cortés, Leandro Padilla, Daniel Maturana, Marko Budinich, Alejandro Maass, and Pilar Parada. Stoichiometric modeling of oxidation of reduced inorganic sulfur compounds (riscs) in acidithiobacillus thiooxidans. *Biotechnology and Bioengineering*, 110(8):2242–2251, 2013.

[28] Qusheng Jin and Matthew F Kirk. ph as a primary control in environmental microbiology: 1. thermodynamic perspective. *Frontiers in Environmental Science*, 6:21, 2018.

[29] H Stolp. Interactions between bdellovibrio and its host cell. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 204(1155):211–217, 1979.

[30] Ophelia S Venturelli, Alex V Carr, Garth Fisher, Ryan H Hsu, Rebecca Lau, Benjamin P Bowen, Susan Hromada, Trent Northen, and Adam P Arkin. Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular systems biology*, 14(6):e8157, 2018.

[31] Kumar Mainali, Sharon Bewick, Briana Vecchio-Pagan, David Karig, and William F Fagan. Detecting interaction networks in the human microbiome with conditional granger causality. *PLoS computational biology*, 15(5):e1007037, 2019.

[32] Roie Levy and Elhanan Borenstein. Reverse ecology: from systems to environments and back. In *Evolutionary systems biology*, pages 329–345. Springer, 2012.

[33] Didier Gonze, Katharine Z Coyte, Leo Lahti, and Karoline Faust. Microbial communities as dynamical systems. *Current opinion in microbiology*, 44:41–49, 2018.

[34] Almut Heinken, Arianna Basile, and Ines Thiele. Advances in constraint-based modelling of microbial communities. *Current Opinion in Systems Biology*, 27: 100346, 2021.

[35] Ilija Dukovski, Djordje Bajić, Jeremy M Chacón, Michael Quintin, Jean CC Vila, Snorre Sulheim, Alan R Pacheco, David B Bernstein, William J Riehl, Kirill S Korolev, et al. A metabolic modeling platform for the computation of microbial ecosystems in time and space (comets). *Nature protocols*, 16(11):5030–5082, 2021.

[36] Eric Parmentier, Déborah Lanterbecq, and Igor Eeckhaut. From commensalism to parasitism in carapidae (ophidiiformes): heterochronic modes of development? *PeerJ*, 4:e1786, 2016.

[37] Eyal Bairey, Eric D Kelsic, and Roy Kishony. High-order species interactions shape ecosystem diversity. *Nature communications*, 7(1):1–7, 2016.

[38] Denis Noble. *The music of life: biology beyond genes*. Oxford University Press, 2008.

[39] Yang Cao, Yuanyuan Wang, Xiaofei Zheng, Fei Li, and Xiaochen Bo. Revecor: an r package for the reverse ecology analysis of microbiomes. *BMC bioinformatics*, 17 (1):1–6, 2016.

[40] Andrew D Steen, Alexander Crits-Christoph, Paul Carini, Kristen M DeAngelis, Noah Fierer, Karen G Lloyd, and J Cameron Thrash. High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal*, 13(12): 3126–3130, 2019.

[41] Joshua E Goldford, Nanxi Lu, Djordje Bajić, Sylvie Estrela, Mikhail Tikhonov, Alicia Sanchez-Gorostiaga, Daniel Segrè, Pankaj Mehta, and Alvaro Sanchez. Emergent simplicity in microbial community assembly. *Science*, 361(6401):469–474, 2018.

[42] Karen L Bell, Robert A Petit III, Anya Cutler, Emily K Dobbs, J Michael Macpherson, Timothy D Read, Kevin S Burgess, and Berry J Brosi. Comparing whole-genome shotgun sequencing and dna metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and evolution*, 11(22):16082–16098, 2021.

[43] Steven J Blazewicz, Romain L Barnard, Rebecca A Daly, and Mary K Firestone. Evaluating rrna as an indicator of microbial activity in environmental communities: limitations and uses. *The ISME journal*, 7(11):2061–2068, 2013.

[44] Thomas J Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, 5:209, 2014.

[45] Ilaria Laudadio, Valerio Fulci, Francesca Palone, Laura Stronati, Salvatore Cucchiara, and Claudia Carissimi. Quantitative assessment of shotgun metagenomics and 16s rdna amplicon sequencing in the study of human gut microbiome. *Omics: a journal of integrative biology*, 22(4):248–254, 2018.

[46] Sonia Dávila-Ramos, Hugo G Castelán-Sánchez, Liliana Martínez-Ávila, María del Rayo Sánchez-Carbente, Raúl Peralta, Armando Hernández-Mendoza, Alan DW Dobson, Ramón A Gonzalez, Nina Pastor, and Ramón Alberto Batista-García. A

review on viral metagenomics in extreme environments. *Frontiers in microbiology*, page 2403, 2019.

[47] Adam G Clooney, Fiona Fouhy, Roy D Sleator, Aisling O'Driscoll, Catherine Stanton, Paul D Cotter, and Marcus J Claesson. Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis. *PloS one*, 11(2): e0148028, 2016.

[48] Nicola Segata. On the road to strain-resolved comparative metagenomics. *MSystems*, 3(2):e00190–17, 2018.

[49] Peter W Harrison, Alisha Ahamed, Raheela Aslam, Blaise TF Alako, Josephine Burgin, Nicola Buso, Mélanie Courtot, Jun Fan, Dipayan Gupta, Muhammad Haseeb, et al. The european nucleotide archive in 2020. *Nucleic acids research*, 49(D1):D82–D85, 2021.

[50] Chao Yang, Debajyoti Chowdhury, Zhenmiao Zhang, William K Cheung, Aiping Lu, Zhaoxiang Bian, and Lu Zhang. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, 19:6301–6314, 2021.

[51] Florian P Breitwieser, Jennifer Lu, and Steven L Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, 20(4):1125–1136, 2019.

[52] Despoina D Roumpeka, R John Wallace, Frank Escalettes, Ian Fotheringham, and Mick Watson. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in genetics*, 8:23, 2017.

[53] H Ye Simon, Katherine J Siddle, Daniel J Park, and Pardis C Sabeti. Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794, 2019.

[54] Donovan H Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J Mussig, and Philip Hugenholtz. A complete domain-to-species taxonomy for bacteria and archaea. *Nature biotechnology*, 38(9):1079–1086, 2020.

[55] Vera G Fonseca. Pitfalls in relative abundance estimation using edna metabarcoding, 2018.

[56] Miklós Bálint, Mohammad Bahram, A Murat Eren, Karoline Faust, Jed A Fuhrman, Björn Lindahl, Robert B O'Hara, Maarja Öpik, Mitchell L Sogin, Martin Unterseher, et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS microbiology reviews*, 40(5):686–700, 2016.

[57] Hojun Song, Jennifer E Buhay, Michael F Whiting, and Keith A Crandall. Many species in one: Dna barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the national academy of sciences*, 105(36):13486–13491, 2008.

[58] Douda Bensasson, De-Xing Zhang, Daniel L Hartl, and Godfrey M Hewitt. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in ecology & evolution*, 16(6):314–321, 2001.

[59] Jay S Ghurye, Victoria Cepeda-Espinoza, and Mihai Pop. Focus: microbiome: metagenomic assembly: overview, challenges and applications. *The Yale journal of biology and medicine*, 89(3):353, 2016.

[60] Yi Yue, Hao Huang, Zhao Qi, Hui-Min Dou, Xin-Yi Liu, Tian-Fei Han, Yue Chen, Xiang-Jun Song, You-Hua Zhang, and Jian Tu. Evaluating metagenomics tools for genome binning with real metagenomic datasets and cami datasets. *BMC bioinformatics*, 21(1):1–15, 2020.

[61] Wei-Zhi Song and Torsten Thomas. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics*, 33(12): 1873–1875, 2017.

[62] Ivan Merelli, Horacio Pérez-Sánchez, Sandra Gesing, and Daniele D'Agostino. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. *BioMed research international*, 2014, 2014.

[63] Pajau Vangay, Josephine Burgin, Anjanette Johnston, Kristen L Beck, Daniel C Berrios, Kai Blumberg, Shane Canon, Patrick Chain, John-Marc Chandonia, Danielle Christianson, et al. Microbiome metadata standards: Report of the national microbiome data collaborative's workshop and follow-on activities. *Msystems*, 6(1): e01194–20, 2021.

[64] Bin Hu, Shane Canon, Emiley A Eloe-Fadrosh, Michal Babinski, Yuri Corilo, Karen Davenport, William D Duncan, Kjiersten Fagnan, Mark Flynn, Brian Foster, et al. Challenges in bioinformatics workflows for processing microbiome omics data at scale. *Frontiers in Bioinformatics*, 1, 2022.

[65] Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, et al. Minimum information about a marker gene sequence (mimarks) and minimum information about any (x) sequence (mixs) specifications. *Nature biotechnology*, 29(5):415–420, 2011.

[66] Pelin Yilmaz, Jack A Gilbert, Rob Knight, Linda Amaral-Zettler, Ilene Karsch-Mizrachi, Guy Cochrane, Yasukazu Nakamura, Susanna-Assunta Sansone, Frank Oliver Gloeckner, and Dawn Field. The genomic standards consortium: bringing standards to life for microbial ecology. *The ISME journal*, 5(10):1565–1567, 2011.

[67] Elisha M Wood-Charlson, Deanna Auberry, Hannah Blanco, Mark I Borkum, Yuri E Corilo, Karen W Davenport, Shweta Deshpande, Ranjeet Devarakonda, Meghan Drake, William D Duncan, et al. The national microbiome data collaborative: enabling microbiome science. *Nature Reviews Microbiology*, 18(6):313–314, 2020.

[68] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[69] Iman Tavassoly, Joseph Goldfarb, and Ravi Iyengar. Systems biology primer: the basic methods and approaches. *Essays in biochemistry*, 62(4):487–500, 2018.

[70] I-Min A Chen, Ken Chu, Krishnaveni Palaniappan, Anna Ratner, Jinghua Huang, Marcel Huntemann, Patrick Hajek, Stephan Ritter, Neha Varghese, Rekha Seshadri, et al. The img/m data management and analysis system v. 6.0: new tools and advanced capabilities. *Nucleic acids research*, 49(D1):D751–D763, 2021.

[71] Ross Overbeek, Robert Olson, Gordon D Pusch, Gary J Olsen, James J Davis, Terry Disz, Robert A Edwards, Svetlana Gerdes, Bruce Parrello, Maulik Shukla, et al. The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic acids research*, 42(D1):D206–D214, 2014.

[72] Igor B Zhulin. Databases for microbiologists. *Journal of bacteriology*, 197(15): 2458–2467, 2015.

[73] Andrew Simpson, Mark Slaymaker, and David Gavaghan. On the secure sharing and aggregation of data to support systems biology research. In *International Conference on Data Integration in the Life Sciences*, pages 58–73. Springer, 2010.

[74] Maria Victoria Schneider and Rafael C Jimenez. Teaching the fundamentals of biological data integration using classroom games. *PLoS computational biology*, 8 (12):e1002789, 2012.

[75] Lincoln D Stein. Integrating biological databases. *Nature Reviews Genetics*, 4(5): 337–345, 2003.

[76] Thomas Triplet and Gregory Butler. Systems biology warehousing: challenges and strategies toward effective data integration. In *Proc. 3rd International Conference on Advances in Databases, Knowledge, and Data Applications, St. Maarten. IARIA*, pages 34–40, 2011.

[77] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

[78] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[79] Lorenz Christian Reimer, Anna Vetcininova, Joaquim Sardà Carbasse, Carola Söhngen, Dorothea Gleim, Christian Ebeling, and Jörg Overmann. Bac dive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. *Nucleic acids research*, 47(D1):D631–D636, 2019.

[80] Sabina Leonelli. Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 44(4): 503–514, 2013.

[81] Robert Stevens, Carole A Goble, and Sean Bechhofer. Ontology-based knowledge representation for bioinformatics. *Briefings in bioinformatics*, 1(4):398–414, 2000.

[82] Pier Luigi Buttigieg, Evangelos Pafilis, Suzanna E Lewis, Mark P Schildhauer, Ramona L Walls, and Christopher J Mungall. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of biomedical semantics*, 7(1):1–12, 2016.

[83] Mike Uschold, Martin King, Stuart Moralee, and Yannis Zorgios. The enterprise ontology. *The knowledge engineering review*, 13(1):31–89, 1998.

[84] Jonathan Furner. Definitions of "metadata": A brief survey of international standards. *Journal of the Association for Information Science and Technology*, 71(6): E33–E42, 2020.

[85] Marcia Lei Zeng and Jian Qin. *Metadata, Second Edition*. Chicago : Neal-Schuman, 2016.

[86] Pier Luigi Buttigieg EnvironmentOntology. Using envo with mixs · environmentontology/envo wiki, Apr 2021. URL https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS.

[87] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A Chen, Nikos C Kyrpides, and TBK Reddy. Genomes online database (gold) v. 8: overview and updates. *Nucleic Acids Research*, 49(D1):D723–D733, 2021.

[88] Andrew Morris, Kyle Meyer, and Brendan Bohannan. Linking microbial communities to ecosystem functions: what we can learn from genotype–phenotype mapping in organisms. *Philosophical Transactions of the Royal Society B*, 375(1798):20190244, 2020.

[89] John R Schramski, Anthony I Dell, John M Grady, Richard M Sibly, and James H Brown. Metabolic theory predicts whole-ecosystem properties. *Proceedings of the National Academy of Sciences*, 112(8):2617–2622, 2015.

[90] Cassio Lima, Howbeer Muhamadali, and Royston Goodacre. The role of raman spectroscopy within quantitative metabolomics. *Annual Review of Analytical Chemistry*, 14:323–345, 2021.

[91] Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2009.

[92] Ines Thiele and Bernhard Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93, 2010.

[93] Joana Saldida, Anna Paolo Muntoni, Daniele de Martino, Georg Hubmann, Bastian Niebel, A Mareike Schmidt, Alfredo Braunstein, Andreas Milias-Argeits, and Matthias Heinemann. Unbiased metabolic flux inference through combined thermodynamic and 13c flux analysis. *bioRxiv*, 2020.

[94] Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305, 2012.

[95] Bernhard Ø. Palsson. *Systems biology*. Cambridge university press, 2015.

[96] Ali Cakmak, Xinjian Qi, A Ercument Cicek, Ilya Bederman, Leigh Henderson, Mitchell Drumm, and Gultekin Ozsoyoglu. A new metabolomics analysis technique: steady-state metabolic network dynamics analysis. *Journal of bioinformatics and computational biology*, 10(01):1240003, 2012.

[97] Jennifer L Reed. Shrinking the metabolic solution space using experimental datasets. 2012.

[98] Jeffrey D Orth, Ines Thiele, and Bernhard Ø. Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[99] Steinn Gudmundsson and Ines Thiele. Computationally efficient flux variability analysis. *BMC bioinformatics*, 11(1):1–3, 2010.

[100] Avantika A Shastri and John A Morgan. Flux balance analysis of photoautotrophic metabolism. *Biotechnology progress*, 21(6):1617–1626, 2005.

[101] Stephen Philip Chapman, Caroline Mary Paget, Giles Nicholas Johnson, and Jean-Marc Schwartz. Flux balance analysis reveals acetate metabolism modulates cyclic electron flow and alternative glycolytic pathways in chlamydomonas reinhardtii. *Frontiers in plant science*, 6:474, 2015.

[102] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–349, 2010.

[103] Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle III. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical journal*, 83 (3):1331–1340, 2002.

[104] Helena A Herrmann, Beth C Dyson, Lucy Vass, Giles N Johnson, and Jean-Marc Schwartz. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *NPJ systems biology and applications*, 5(1):1–8, 2019.

[105] Jan Schellenberger and Bernhard Ø. Palsson. Use of randomized sampling for analysis of metabolic networks. *Journal of biological chemistry*, 284(9):5457–5461, 2009.

[106] Nathan D Price, Jan Schellenberger, and Bernhard O Palsson. Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. *Biophysical journal*, 87(4):2172–2186, 2004.

[107] William T Scott, Eddy J Smid, David E Block, and Richard A Notebaart. Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts. *Microbial cell factories*, 20(1):1–15, 2021.

[108] Robert L. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984. ISSN 0030364X, 15265463.

[109] David E Kaufman and Robert L Smith. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1):84–95, 1998.

[110] Hulda S Haraldsdóttir, Ben Cousins, Ines Thiele, Ronan MT Fleming, and Santosh Vempala. CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743, 2017.

[111] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdóttir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019.

[112] Christian Diener, Sean M Gibbons, and Osbaldo Resendis-Antonio. Micom: metagenome-scale modeling to infer metabolic interactions in the gut microbiota. *MSystems*, 5(1):e00606–19, 2020.

[113] Hulda S Haraldsdóttir, Ben Cousins, Ines Thiele, Ronan M.T Fleming, and Santosh Vempala. CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743, 01 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx052.

[114] A Pavan-Kumar, P Gireesh-Babu, and WS Lakra. Dna metabarcoding: a new approach for rapid biodiversity assessment. *J Cell Sci Mol Biol*, 2(1):111, 2015.

[115] Philip Francis Thomsen and Eske Willerslev. Environmental dna–an emerging tool in conservation for monitoring past and present biodiversity. *Biological conservation*, 183:4–18, 2015.

[116] Yinqiu Ji, Louise Ashton, Scott M Pedley, David P Edwards, Yong Tang, Akihiro Nakamura, Roger Kitching, Paul M Dolman, Paul Woodcock, Felicity A Edwards, et al. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology letters*, 16(10):1245–1257, 2013.

[117] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

[118] Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhiyan Arumugam, Francesco Asnicar, et al. Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science. Technical report, PeerJ Preprints, 2018.

[119] Falk Hildebrand, Raul Tadeo, Anita Yvonne Voigt, Peer Bork, and Jeroen Raes. Lotus: an efficient and user-friendly otu processing pipeline. *Microbiome*, 2(1):1–7, 2014.

[120] Jan Axtner, Alex Crampton-Platt, Lisa A Hörig, Azlan Mohamed, Charles CY Xu, Douglas W Yu, and Andreas Wilting. An efficient and robust laboratory workflow and tetrapod database for larger scale environmental dna studies. *GigaScience*, 8 (4):giz029, 2019.

[121] Hyun S Gweon, Anna Oliver, Joanne Taylor, Tim Booth, Melanie Gibbs, Daniel S Read, Robert I Griffiths, and Karsten Schonrogge. Pipits: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the i llumina sequencing platform. *Methods in ecology and evolution*, 6(8):973–980, 2015.

[122] Pablo Cingolani, Rob Sladek, and Mathieu Blanchette. Bigdatascript: a scripting language for data pipelines. *Bioinformatics*, 31(1):10–16, 2015.

[123] Babak Bashari Rad, Harrison John Bhatti, and Mohammad Ahmadi. An introduction to docker and analysis of its performance. *International Journal of Computer Science and Network Security (IJCSNS)*, 17(3):228, 2017.

[124] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459, 2017.

[125] Eric Coissac, Tiayyba Riaz, and Nicolas Puillandre. Bioinformatic challenges for dna metabarcoding of plants and animals. *Molecular ecology*, 21(8):1834–1847, 2012.

[126] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639–2643, 2017.

[127] Charlie Pauvert, Marc Buée, Valérie Laval, Véronique Edel-Hermann, Laure Fauchery, Angélique Gautier, Isabelle Lesur, Jessica Vallance, and Corinne Vacher. Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, 41:23–33, 2019.

[128] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

[129] Xiaolin Hao, Rui Jiang, and Ting Chen. Clustering 16s rrna for otu prediction: a method of unsupervised bayesian clustering. *Bioinformatics*, 27(5):611–618, 2011.

[130] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420, 2015.

[131] Anders Lanzén, Steffen L Jørgensen, Daniel H Huson, Markus Gorfer, Svenn Helge Grindhaug, Inge Jonassen, Lise Øvreås, and Tim Urich. Crest–classification resources for environmental sequence tags. *PloS one*, 7(11):e49334, 2012.

[132] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.

[133] Matthias C Rillig, Masahiro Ryo, Anika Lehmann, Carlos A Aguilar-Trigueros, Sabine Buchert, Anja Wulf, Aiko Iwasaki, Julien Roy, and Gaowen Yang. The role of multiple global change factors in driving soil functions and microbial biodiversity. *Science*, 366(6467):886–890, 2019.

[134] Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455, 2019.

[135] Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, and Alexandros Stamatakis. Epa-ng: massively parallel evolutionary placement of genetic sequences. *Systematic biology*, 68(2):365–369, 2019.

[136] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

[137] Ryuji J Machida, Matthieu Leray, Shian-Lei Ho, and Nancy Knowlton. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific data*, 4(1):1–7, 2017.

[138] Paul J McMurdie and Susan Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217, 2013.

[139] Fastqc, Jun 2015. URL https://qubeshub.org/resources/fastqc.

[140] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[141] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200. URL https://journal.embnet.org/index.php/embnetjournal/article/view/200. Number: 1.

[142] Sergey I Nikolenko, Anton I Korobeynikov, and Max A Alekseyev. Bayeshammer: Bayesian clustering for error correction in single-cell sequencing. In *BMC genomics*, volume 14, pages 1–11. Springer, 2013.

[143] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[144] Andre P Masella, Andrea K Bartram, Jakub M Truszkowski, Daniel G Brown, and Josh D Neufeld. Pandaseq: paired-end assembler for illumina sequences. *BMC bioinformatics*, 13(1):1–7, 2012.

[145] Frédéric Boyer, Céline Mercier, Aurélie Bonin, Yvan Le Bras, Pierre Taberlet, and Eric Coissac. obitools: A unix-inspired software package for dna metabarcoding. *Molecular ecology resources*, 16(1):176–182, 2016.

[146] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[147] Rolf Henrik Nilsson, Karl-Henrik Larsson, Andy F S Taylor, Johan Bengtsson-Palme, Thomas S Jeppesen, Dmitry Schigel, Peter Kennedy, Kathryn Picard, Frank Oliver Glöckner, Leho Tedersoo, et al. The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic acids research*, 47(D1):D259–D264, 2019.

[148] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D Pruitt, and Eric W Sayers. Genbank. *Nucleic acids research*, 46(D1): D41–D47, 2018.

[149] Lucas Czech, Pierre Barbera, and Alexandros Stamatakis. Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics*, 35(7):1151–1158, 2019.

[150] Simon A Berger and Alexandros Stamatakis. Papara 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension. *Heidelberg Institute for Theoretical Studies*, 2012.

[151] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research*, 49(W1):W293–W296, 2021.

[152] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[153] Tsukasa Nakamura, Kazunori D Yamada, Kentaro Tomii, and Kazutaka Katoh. Parallelization of mafft for large-scale multiple sequence alignments. *Bioinformatics*, 34(14):2490–2492, 2018.

[154] Daryl M Gohl, Pajau Vangay, John Garbe, Allison MacLean, Adam Hauge, Aaron Becker, Trevor J Gould, Jonathan B Clayton, Timothy J Johnson, Ryan Hunter, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature biotechnology*, 34(9):942–949, 2016.

[155] Ian M. Bradley, Ameet J. Pinto, Jeremy S. Guest, and G. Voordouw. Design and evaluation of illumina miseq-compatible, 18s rrna gene-specific primers for improved characterization of mixed phototrophic communities. *Applied and Environmental Microbiology*, 82(19):5878–5891, 2016. doi: 10.1128/AEM.01630-16.

[156] Matthew G Bakker. A fungal mock community control for amplicon sequencing experiments. *Molecular ecology resources*, 18(3):541–556, 2018.

[157] Iliana Bista, Gary R Carvalho, Min Tang, Kerry Walsh, Xin Zhou, Mehrdad Hajibabaei, Shadi Shokralla, Mathew Seymour, David Bradley, Shanlin Liu, et al. Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, 18(5):1020–1034, 2018.

[158] Christina Pavloudi, Jon B Kristoffersen, Anastasis Oulas, Marleen De Troch, and Christos Arvanitidis. Sediment microbial taxonomic and functional diversity in a natural salinity gradient challenge remane's "species minimum" concept. *PeerJ*, 5: e3687, 2017.

[159] I Bista, GR Carvalho, K Walsh, M Seymour, M Hajibabaei, D Lallias, M Christmas, and S Creer. Annual time-series analysis of aqueous edna reveals ecologically relevant dynamics of lake ecosystem biodiversity. nat. commun. 8, 14087, 2017.

[160] Peter W Harrison, Blaise Alako, Clara Amid, Ana Cerdeño-Tárraga, Iain Cleland, Sam Holt, Abdulrahman Hussein, Suran Jayathilaka, Simon Kay, Thomas Keane, et al. The european nucleotide archive in 2018. *Nucleic acids research*, 47(D1): D84–D88, 2019.

[161] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[162] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):1–9, 2009.

[163] Sujeevan Ratnasingham and Paul DN Hebert. Bold: The barcode of life data system (http://www. barcodinglife. org). *Molecular ecology notes*, 7(3):355–364, 2007.

[164] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593, 2014.

[165] Sydney I Glassman and Jennifer BH Martiny. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *MSphere*, 3(4):e00148–18, 2018.

[166] Teresita M. Porter. terrimporter/12SvertebrateClassifier: 12S Vertebrate Classifier v2.0.0-ref, August 2021. URL https://doi.org/10.5281/zenodo.5157047.

[167] Laure Guillou, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, Gaétan Burgaud, Colomban De Vargas, Johan Decelle, et al. The protist ribosomal reference database (pr2): a catalog of unicellular eukaryote small sub-unit rrna sequences with curated taxonomy. *Nucleic acids research*, 41(D1):D597–D604, 2012.

[168] Jan P Buchmann and Edward C Holmes. Collecting and managing taxonomic data with ncbi-taxonomist. *Bioinformatics*, 36(22-23):5548–5550, 2020.

[169] Kristy Deiner, Holly M Bik, Elvira Mächler, Mathew Seymour, Anaïs Lacoursière-Roussel, Florian Altermatt, Simon Creer, Iliana Bista, David M Lodge, Natasha De Vere, et al. Environmental dna metabarcoding: Transforming how we survey animal and plant communities. *Molecular ecology*, 26(21):5872–5895, 2017.

[170] Krista M Ruppert, Richard J Kline, and Md Saydur Rahman. Past, present, and future perspectives of environmental dna (edna) metabarcoding: A systematic review in methods, monitoring, and applications of global edna. *Global Ecology and Conservation*, 17:e00547, 2019.

[171] Pierre Taberlet, Eric Coissac, Mehrdad Hajibabaei, and Loren H Rieseberg. Environmental dna, 2012.

[172] Michael Stat, Megan J Huggett, Rachele Bernasconi, Joseph D DiBattista, Tina E Berry, Stephen J Newman, Euan S Harvey, and Michael Bunce. Ecosystem biomonitoring with edna: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, 7(1):1–11, 2017.

[173] Bruce E Deagle, Simon N Jarman, Eric Coissac, François Pompanon, and Pierre Taberlet. Dna metabarcoding and the cytochrome c oxidase subunit i marker: not a perfect match. *Biology letters*, 10(9):20140562, 2014.

[174] Pierre Taberlet, Eric Coissac, François Pompanon, Christian Brochmann, and Eske Willerslev. Towards next-generation biodiversity assessment using dna metabarcoding. *Molecular ecology*, 21(8):2045–2050, 2012.

[175] Tamara Schenekar, Martin Schletterer, Laurène A Lecaudey, and Steven J Weiss. Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an edna fish assessment in the volga headwaters. *River Research and Applications*, 36(7):1004–1013, 2020.

[176] Kevin Cilleros, Alice Valentini, Luc Allard, Tony Dejean, Roselyne Etienne, Gael Grenouillet, Amaia Iribar, Pierre Taberlet, Regis Vigouroux, and Sebastien Brosse. Unlocking biodiversity and conservation studies in high-diversity environments using environmental dna (edna): A test with guianese freshwater fishes. *Molecular Ecology Resources*, 19(1):27–46, 2019.

[177] W John Kress, Carlos García-Robledo, Maria Uriarte, and David L Erickson. Dna barcodes for ecology, evolution, and conservation. *Trends in ecology & evolution*, 30(1):25–35, 2015.

[178] Paul DN Hebert, Sujeevan Ratnasingham, and Jeremy R De Waard. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270 (suppl_1):S96–S99, 2003.

[179] Vasco Elbrecht and Florian Leese. Validation and development of coi metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers in Environmental Science*, 5:11, 2017.

[180] Masaki Miya, Ryo O Gotoh, and Tetsuya Sado. Mifish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental dna and other samples. *Fisheries Science*, pages 1–32, 2020.

[181] Seda Ekici, Grzegorz Pawlik, Eva Lohmeyer, Hans-Georg Koch, and Fevzi Daldal. Biogenesis of cbb3-type cytochrome c oxidase in rhodobacter capsulatus. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1817(6):898–910, 2012.

[182] Sonja Schimo, Ilka Wittig, Klaas M Pos, and Bernd Ludwig. Cytochrome c oxidase biogenesis and metallochaperone interactions: steps in the assembly pathway of a bacterial complex. *PLoS One*, 12(1):e0170037, 2017.

[183] Monika Mioduchowska, Michał Jan Czyż, Bartłomiej Gołdyn, Jarosław Kur, and Jerzy Sell. Instances of erroneous dna barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? *PLoS One*, 13(6):e0199609, 2018.

[184] Mark E Siddall, Frank M Fontanella, Sara C Watson, Sebastian Kvist, and Christer Erséus. Barcoding bamboozled by bacteria: convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, 58(4):445–451, 2009.

[185] Carmelo Andújar, Paula Arribas, Douglas W Yu, Alfried P Vogler, and Brent C Emerson. Why the coi barcode should be the community dna metabarcode for the metazoa, 2018.

[186] Pierre Taberlet, Aurélie Bonin, Lucie Zinger, and Eric Coissac. Analysis of bulk samples. In *Environmental DNA*, pages 140–143. Oxford University Press.

[187] ChenXue Yang, YingQiu Ji, XiaoYang Wang, ChunYang Yang, and W Yu Douglas. Testing three pipelines for 18s rdna-based metabarcoding of soil faunal diversity. *Science China Life Sciences*, 56(1):73–81, 2013.

[188] Chenxue Yang, Xiaoyang Wang, Jeremy A Miller, Marleen de Blécourt, Yinqiu Ji, Chunyan Yang, Rhett D Harrison, and W Yu Douglas. Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, 46:379–389, 2014.

[189] Rupert A Collins, Judith Bakker, Owen S Wangensteen, Ana Z Soto, Laura Corrigan, David W Sims, Martin J Genner, and Stefano Mariani. Non-specific amplification compromises environmental dna metabarcoding with coi. *Methods in Ecology and Evolution*, 10(11):1985–2001, 2019.

[190] Eva Aylagas, Ángel Borja, Xabier Irigoien, and Naiara Rodríguez-Ezpeleta. Benchmarking dna metabarcoding for biodiversity-based monitoring and assessment. *Frontiers in Marine Science*, 3:96, 2016.

[191] Frédéric Sinniger, Jan Pawlowski, Saki Harii, Andrew J Gooday, Hiroyuki Yamamoto, Pierre Chevaldonné, Tomas Cedhagen, Gary Carvalho, and Simon Creer. Worldwide analysis of sedimentary dna reveals major gaps in taxonomic knowledge of deep-sea benthos. *Frontiers in Marine Science*, 3:92, 2016.

[192] Quiterie Haenel, Oleksandr Holovachov, Ulf Jondelius, Per Sundberg, and Sarah J Bourlat. Ngs-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from hållö island, smögen, and soft mud from gullmarn fjord, sweden. *Biodiversity data journal*, (5), 2017.

[193] Guillaume Bernard, Jananan S Pathmanathan, Romain Lannes, Philippe Lopez, and Eric Bapteste. Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. *Genome biology and evolution*, 10(3):707–715, 2018.

[194] Mahwash Jamy, Rachel Foster, Pierre Barbera, Lucas Czech, Alexey Kozlov, Alexandros Stamatakis, Gary Bending, Sally Hilton, David Bass, and Fabien Burki. Long-read metabarcoding of the eukaryotic rdna operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular ecology resources*, 20(2): 429–443, 2020.

[195] Lucas Czech, Pierre Barbera, and Alexandros Stamatakis. Genesis and gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics*, 36(10):3263–3265, 2020.

[196] Jacob L Steenwyk, Thomas J Buida III, Yuanning Li, Xing-Xing Shen, and Antonis Rokas. Clipkit: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS biology*, 18(12):e3001007, 2020.

[197] Diep Thi Hoang, Olga Chernomor, Arndt Von Haeseler, Bui Quang Minh, and Le Sy Vinh. Ufboot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution*, 35(2):518–522, 2018.

[198] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.

[199] Haris Zafeiropoulos, Anastasia Gioti, Stelios Ninidakis, Antonis Potirakis, Savvas Paragkamian, Nelina Angelova, Aglaia Antoniou, Theodoros Danis, Eliza Kaitetzidou, Panagiotis Kasapidis, et al. 0s and 1s in marine molecular research: a regional hpc perspective. *GigaScience*, 10(8):giab053, 2021.

[200] Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a web browser. *BMC bioinformatics*, 12(1):1–10, 2011.

[201] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581–583, 2016.

[202] Haris Zafeiropoulos, Ha Quoc Viet, Katerina Vasileiadou, Antonis Potirakis, Christos Arvanitidis, Pantelis Topalis, Christina Pavloudi, and Evangelos Pafilis. Pema: a flexible pipeline for environmental dna metabarcoding analysis of the 16s/18s ribosomal rna, its, and coi marker genes. *GigaScience*, 9(3):giaa022, 2020.

[203] Adrià Antich, Creu Palacin, Owen S Wangensteen, and Xavier Turon. To denoise or to cluster, that is not the question: optimizing pipelines for coi metabarcoding and metaphylogeography. *BMC bioinformatics*, 22(1):1–24, 2021.

[204] Matthias Obst, Katrina Exter, A Louise Allcock, Christos Arvanitidis, Alizz Axberg, Maria Bustamante, Ibon Cancio, Diego Carreira-Flores, Eva Chatzinikolaou, Giorgos Chatzigeorgiou, et al. A marine biodiversity observation network for genetic monitoring of hard-bottom communities (arms-mbon). *Frontiers in Marine Science*, 7:1031, 2020.

[205] S Kamenova. A flexible pipeline combining clustering and correction tools for prokaryotic and eukaryotic metabarcoding. peer community in ecology 1: 100043, 2020.

[206] Sujai Kumar, Martin Jones, Georgios Koutsovoulos, Michael Clarke, and Mark Blaxter. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated gc-coverage plots. *Frontiers in genetics*, 4:237, 2013.

[207] Simon M Dittami and Erwan Corre. Detection of bacterial contaminants and hybrid sequences in the genome of the kelp saccharina japonica using taxoblast. *PeerJ*, 5: e4073, 2017.

[208] Giovanna De Simone, Andrea Pasquadibisceglie, Roberta Proietto, Fabio Polticelli, Silvio Aime, Huub JM Op den Camp, and Paolo Ascenzi. Contaminations in (meta) genome data: An open issue for the scientific community. *IUBMB life*, 72(4):698–705, 2020.

[209] Martin Steinegger and Steven L Salzberg. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in genbank. *Genome biology*, 21(1):1–12, 2020.

[210] Scott F Gilbert, Jan Sapp, and Alfred I Tauber. A symbiotic view of life: we have never been individuals. *The Quarterly review of biology*, 87(4):325–341, 2012.

[211] Emiliano Salvucci. Microbiome, holobiont and the net of life. *Critical reviews in microbiology*, 42(3):485–494, 2016.

[212] Hannah Weigand, Arne J Beermann, Fedor Čiampor, Filipe O Costa, Zoltán Csabai, Sofia Duarte, Matthias F Geiger, Michał Grabowski, Frédéric Rimet, Björn Rulik, et al. Dna barcode reference libraries for the monitoring of aquatic biota in europe: Gap-analysis and recommendations for future work. *Science of the Total Environment*, 678:499–524, 2019.

[213] Geert Huys, Tom Vanhoutte, Marie Joossens, Amal S Mahious, Evie De Brandt, Severine Vermeire, and Jean Swings. Coamplification of eukaryotic dna with 16s rrna gene-based pcr primers: possible consequences for population fingerprinting of complex microbial communities. *Current microbiology*, 56(6):553–557, 2008.

[214] Jethro S Johnson, Daniel J Spakowicz, Bo-Young Hong, Lauren M Petersen, Patrick Demkowicz, Lei Chen, Shana R Leopold, Blake M Hanson, Hanako O Agresta, Mark Gerstein, et al. Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature communications*, 10(1):1–11, 2019.

[215] Patricio Jeraldo, Nicholas Chia, and Nigel Goldenfeld. On the suitability of short reads of 16s rrna for phylogeny-based analyses in environmental surveys. *Environmental microbiology*, 13(11):3000–3009, 2011.

[216] Nicholas A Bokulich, Michal Ziemski, Michael S Robeson II, and Benjamin D Kaehler. Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. *Computational and Structural Biotechnology Journal*, 18:4048–4062, 2020.

[217] Teresita M Porter and Mehrdad Hajibabaei. Profile hidden markov model sequence analysis can help remove putative pseudogenes from dna barcoding and metabarcoding datasets. *BMC bioinformatics*, 22(1):1–20, 2021.

[218] Anna Y Pei, William E Oberdorf, Carlos W Nossa, Ankush Agarwal, Pooja Chokshi, Erika A Gerz, Zhida Jin, Peng Lee, Liying Yang, Michael Poles, et al. Diversity of 16s rrna genes within individual prokaryotic genomes. *Applied and environmental microbiology*, 76(12):3886–3897, 2010.

[219] Chih-Horng Kuo and Howard Ochman. The extinction dynamics of bacterial pseudogenes. *PLoS genetics*, 6(8):e1001050, 2010.

[220] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab776. URL https://doi.org/10.1093/nar/gkab776.

[221] Leho Tedersoo, Mads Albertsen, Sten Anslan, and Benjamin Callahan. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied and Environmental Microbiology*, 87(17):e00626–21, 2021.

[222] Djordje Bajic and Alvaro Sanchez. The ecology and evolution of microbial metabolic strategies. *Current opinion in biotechnology*, 62:123–128, 2020.

[223] Kyle I Harrington and Alvaro Sanchez. Eco-evolutionary dynamics of complex social strategies in microbial communities. *Communicative & integrative biology*, 7 (1):e28230, 2014.

[224] Ankur Sahu, Mary-Ann Blätke, Jędrzej Jakub Szymański, and Nadine Töpfer. Advances in flux balance analysis by integrating machine learning and mechanism-based models. *Computational and Structural Biotechnology Journal*, 19:4626–4640, 2021.

[225] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2012.

[226] Janos X Binder, Sune Pletscher-Frankild, Kalliopi Tsafou, Christian Stolte, Seán I O'Donoghue, Reinhard Schneider, and Lars Juhl Jensen. Compartments: unification and visualization of protein subcellular localization evidence. *Database*, 2014, 2014.

Short CV

Education

- **Doctor of Philosophy** (2018 – 2022), University of Crete, Biology Department
  Thesis: Microbial communities through the lens of high throughput sequencing, data integration and metabolic networks analysis
  Thesis conducted at IMBBC - HCMR

- **M.Sc. in Bioinformatics** (2016 – 2018), University of Crete, School of Medicine
  Thesis: eDNA metabarcoding for biodiversity assessment: Algorithm design and bioinformatics analysis pipeline implementation
  Thesis conducted at IMBBC - HCMR

- **B.Sc. in Biology** (2011 – 2016), National and Kapodistrian University of Athens, department of Biology
  Thesis: Morphology, morphometry and anatomy of species of the genus *Pseudamnicola* in Greece

Research projects - working Experience

- **A workflow for marine Genomic Observatories data analysis** (2021 - ongoing)
  Role: scientific responsiblse & developer
  This EOSC-Life funded project aims at developing a workflow for the analysis of EMBRC's Genomic Observatories (GOs) data, allowing researchers to deal better with this increasing amount of the data and make them more easily interpretable.

- **PREGO: Process, environment, organism (PREGO)** (2019 - 2021)
  Role: PhD candidate
  PREGO is a systems-biology approach to elucidate ecosystem function at the microbial dimension.

- **ELIXIR-GR** (2019 - 2021)
  Role: technical support
  ELIXIR-GR is the Greek National Node of the ESFRI European RI ELIXIR, a distributed e-Infrastructure aiming at the construction of a sustainable European infrastructure for biological information.

- **RECONNECT** (2018 - 2020)
  Role: technical support
  RECONNECT is an Interreg V-B "Balkan-Mediterranean 2014-2020" project. It aims to develop strategies for sustainable management of Marine Protected Areas (MPAs) and Natura 2000 sites.

Publications

- **PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types.**
  Zafeiropoulos, H., Paragkamian S.[2], Stelios Ninidakis, Georgios A. Pavlopoulos, Lars Juhl Jensen, and Evangelos Pafilis. *Microorganisms* 10, no. 2 (2022): 293., DOI: 10.3390/microorganisms10020293

- **The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data**
  Zafeiropoulos H., Gargan L., Hintikka S., Pavloudi C., & Carlsson J. *Metabarcoding and Metagenomics*, 5, p.e69657, 2021, DOI: 10.3897/mbmg.5.69657

- **0s & 1s in marine molecular research: a regional HPC perspective.**
  Zafeiropoulos H., Gioti A., Ninidakis S., Potirakis A., ..., & Pafilis E. *GigaScience*, 9(3), p.giab053, 2021 DOI: 10.1093/gigascience/giab053

- **Geometric Algorithms for Sampling the Flux Space of Metabolic Networks**
  Chalkis, A., Fisikopoulos, V., Tsigaridas, E. & **Zafeiropoulos, H.** *37th International Symposium on Computational Geometry (SoCG 2021)*, 21:1–21:16, 189, 2021 DOI: 10.4230/LIPIcs.SoCG.2021.21

- **The Santorini Volcanic Complex as a Valuable Source of Enzymes for Bioenergy**
  Polymenakou, P.N., Nomikou, P., **Zafeiropoulos, H.**, Mandalakis, M., Anastasiou, T.I., Kilias, S., Kyrpides, N.C., Kotoulas, G. & Magoulas,A. *Energies*, 14(5), p.1414, 2021 DOI: 10.3390/en14051414

- **PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes**
  Zafeiropoulos, H., Viet, H.Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C. & Pafilis, E. *GigaScience*, 9(3), p.giaa022, 2020 DOI: 10.1093/gigascience/giaa022

In preparation

- dingo: a Python library for metabolic networks analysis
- Deciphering the functional potential of a hypersaline swamp microbial mat community

[2] ZH and PS contibuted equally in this study

Awards

- **European Molecular Biology Organization Short-Term Fellowship** (2022)
  Project title: Exploiting data integration, text-mining and computational geometry to enhance microbial interactions inference from co-occurrence networks
  https://hariszaf.github.io/microbetag/

- **Mikrobiokosmos travel grant in memorium of Prof. Kostas Drainas** (2021)

- **Google Summer of Code** (2021)
  Project title: From DNA sequences to metabolic interactions: building a pipeline to extract key metabolic processes
  Report, GSoC archive

- **Federation of European Microbiological Societies Meeting Attendance Grant** (2020)
  for joining the *Metagenomics, Metatranscript- omics and multi 'omics for microbial community studies* Physalia course

- **Short Term Scientific Mission (STSM) - DNAqua-net COST action** (2019)
  Project title: A comparison of bioinformatic pipelines and sampling techniques to enable benchmarking of DNA metabarcoding
  Report

- **Best Poster Award @ Hellenic Bioinformatics conference** (2018)
  for *PEMA: a Pipeline for Environmental DNA Metabarcoding Analysis*

Selected presentations

- **Bioinformatics Open Source Conference - BOSC2021** (2021)
  dingo: A python library for metabolic networks sampling & analysis, video poster - video

- **1st DNAQUA International Conference** (2021)
  PEMA v2: addressing metabarcoding bioinformatics analysis challenges, oral talk - video

- **Federation of European Microbiological Societies - FEMS2020** (2020)
  "Mining literature and -omics (meta)data to associate microorganisms, biological processes and environment types" - video poster

- **PyData Global PyData2020**
  "Geometric and statistical methods in systems biology: the case of metabolic networks", oral talk - video

- **8th International Barcode of Life Conference** - 2019
  "P.E.M.A.: a pipeline for environmental DNA metabarcoding analysis" (flashtalk)

Participation in proposal writing

- "Climate Change Metagenomic Record Index (CCMRI)" project: submitted at the 2nd Call for H.F.R.I Research Projects to Support Faculty Members & Researchers (June 2020). Approved for funding
- "A workflow for marine Genomic Observatories data analysis" project: submitted at the second Training Open Call of EOSC-Life (November 2020). Approved for funding

Contact

Personal website: https://hariszaf.github.io/
GitHub account: https://github.com/hariszaf
Twitter account: @haris_zaf
Account in ResearchGate
e-mail: haris.zafr@gmail.com