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

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

Here are some thoughts of mine for the rest of the committe!
Prof Faust
Prof. Elias Tsigaridas
Prof. Klappa
Prof L.J.J

Preface

Haris Zafeiropoulos

Contents

Abstract

Περίληψη

Και στα ελληνικά

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

NGS	Next Generation Sequencing
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

Chapter 1

Introduction

1.1 Microbial ecology

1.1.1 Microbial communities: structure & function

Microbes, i.e. Bacteria, Archaea and small Eukaryotes such as protoza, are omnipresent and impact global ecosystem functions [5] through their abundance [6], versatility [7] and interactions [8].

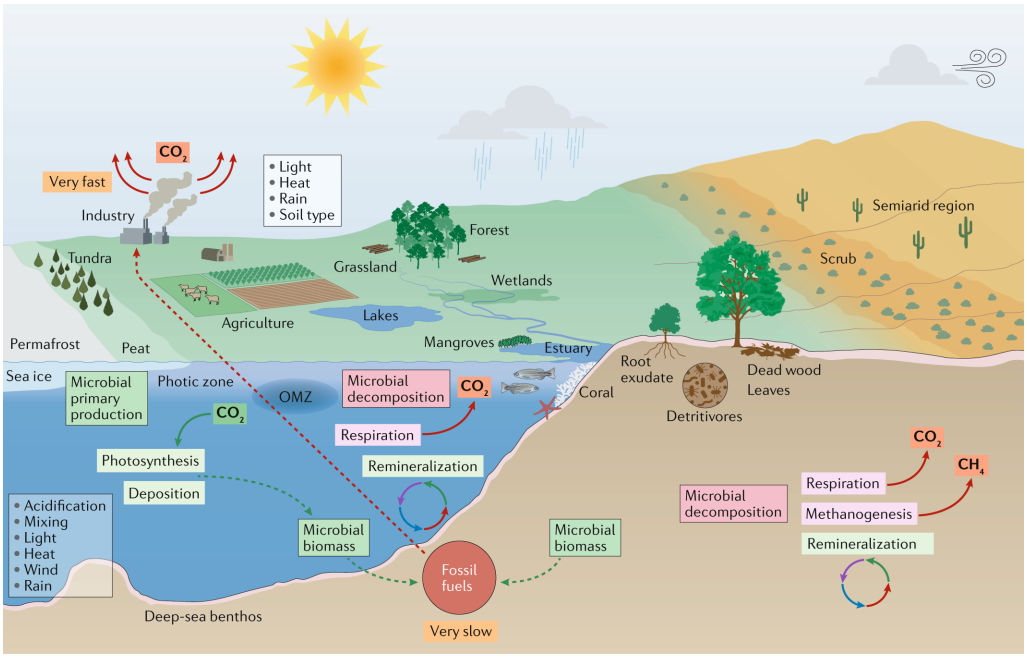

FIGURE 1.1: Marine icrobial communities contribute to CO_2 sequestration, nutrients recycle and thus to the release of CO_2 to the atmosphere. Soil microbial communities decomposers organic matter and release nutrients in the soil from [1] doi: 10.1038/s41579-019-0222-5, under Creative Commons Attribution 4.0 International License

1.1.2 The role of microbial communities in biogeochemical cycles

Microbial communities at hydrothermal vents mediate the transformation of energy and minerals produced by geological activity into organic material. Organic matter produced by autotrophic bacteria is then used to support the upper trophic levels. The hydrothermal vent fluid and the surrounding ocean water is rich in elements such as iron, manganese and various species of sulfur including sulfide, sulfite, sulfate, elemental sulfur from which they can derive energy or nutrients.[8] Microbes derive energy by oxidizing or reducing elements. Different microbial species use different chemical species of an element in their metabolic processes. For example, some microbe species oxidize sulfide to sulfate and another species will reduce sulfate to elemental sulfur. As a result, a web of chemical pathways mediated by different microbial species transform elements such as carbon, sulfur, nitrogen, and hydrogen, from one species to another. Their activity alters the original chemical composition produced by geological activity of the hydrothermal vent environment.[9]

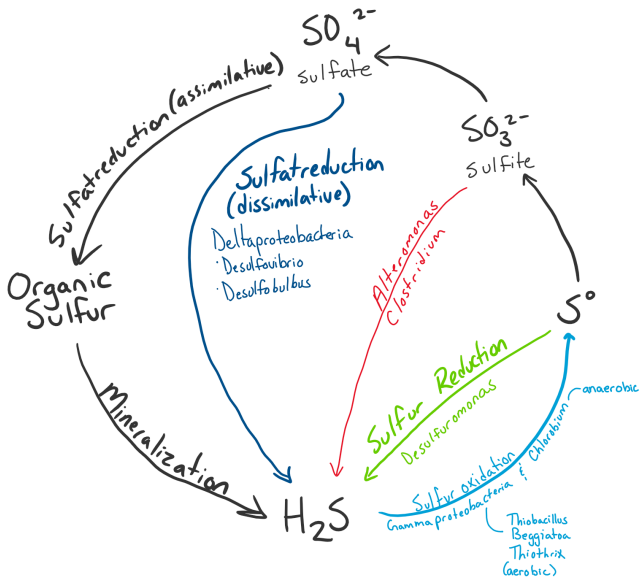

FIGURE 1.2: Sulfur cycle. Figure taken from [2]

1.1.3 Microbial interactions: unravelling the microbiome

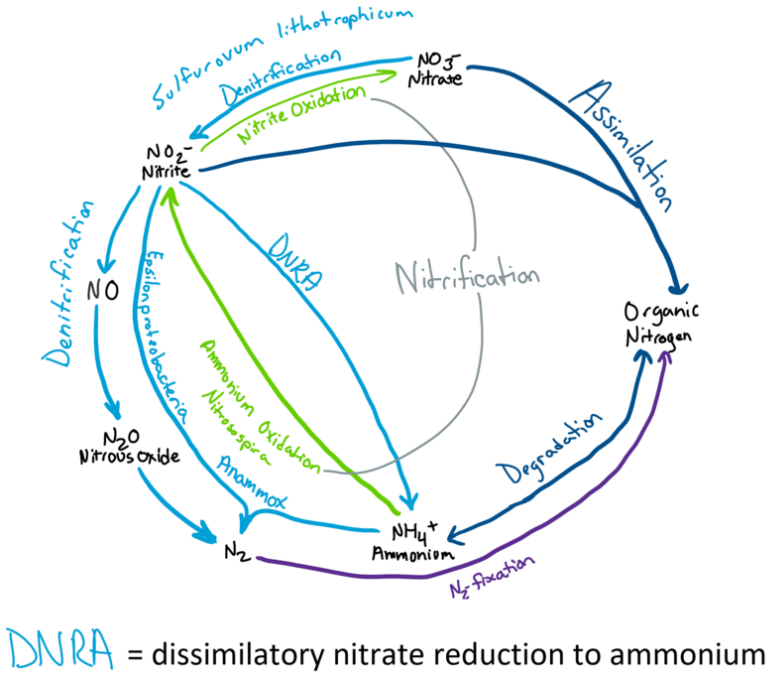

FIGURE 1.3: Nitrogen cycle. Figure taken from [3]

1.2 The era of omics

1.2.1 High Throughput Sequencing approaches

DNA metabarcoding is a rapidly evolving method that is being more frequently employed in a range of fields, such as biodiversity, biomonitoring, molecular ecology and others [9, 10]. Environmental DNA (eDNA) metabarcoding, targeting DNA directly isolated from environmental samples (e.g., water, soil or sediment, [11]), is considered a holistic approach [12] in terms of biodiversity assessment, providing high detection capacity. At the same time, it allows wide scale rapid bio-assessment [12] at a relatively low cost as compared to traditional biodiversity survey methods [13]. 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 [14].

1.2.2 Bioinformatics challenges

- need for tools

- handle the sequences

1.3 Data integration & data mining in the era of omics

1.3.1 Metadata: a key issue for the microbiome community

The Community initially focused on developing open science "best practices" for the research community. The paper "The metagenomic data life-cycle: standards and best practices" [15] provided the foundation for FAIR data management in the domain. These best practices advocated using community standards for contextual provenance and metadata at all stages of the research data life cycle.

Alongside archived sequence data, access to comprehensive metadata is important to contextualise where the data originated. On submission, submitters are given the option to provide details regarding when, where and how their samples were collected with the opportunity to align provided metadata against community developed standards where possible. However, challenges associated with metadata deposition mean submitters do not always provide comprehensive metadata - these challenges can range from: lack of training and outreach resulting in submitters not fully understanding the importance of metadata and how to comply with standards; as well as the trade-offs for the archives to provide complex and thorough validation vs simple user interfaces to ensure both compliance and submission are as easy as possible. For the ENA, extensive documentation exists on how to submit data which both encourages compliance with metadata standards and provides separate submission guidelines for different data types - usage of the documentation can mitigate common errors and often aid first-time submitters but does not reach the full user-base.

FAIR principles, to provide a multilayer set of metadata required by the different scientific communities, reflecting the inherently multi-disciplinary character of environmental microbiology. The various layers of metadata necessary for the FAIRification of MAGs should include:

1. Environmental data describing the sample of origin
2. Sequencing technology or technologies
3. Details on the computational pipeline for metagenome assembly, binning and quality assessment
4. Connection to an existing taxonomy schema

OSD's open access strategy and provenance for metadata annotation is reflected in its ENA and Pangea submissions. Among others Standardization and training are key aspects across OSD: from sampling protocols to metadata checklists and guidelines. This is inline with aims of the Elixir microbiome community (see Sections "Mobilising raw data and metadata", "Training - lack of training"); spreading the experience to other biomes can benefit such ends.

Open questions: Metadata standard definition: minimum set and formats (Some flexibility will have to be considered in sharing standards between domain-specific communities). Systems to extract the vast amount of metadata locked in the scientific literature and provide them in standard format (explored by the Biodiversity Focus Group).

Metadata associated with the raw data, the assembled data, and the workflow. The necessary scripts will be written in Python using standard libraries and Biopython. Metadata of the cleaned data Metadata associated with the data sequencing, sample collection (MIMS), and quality control analyses will be generated according to the ENA manifest to enable uploading and archiving of the data to ENA. Metadata of the assembled data Because the workflow is distributed, it is necessary for EBI-MGnify to verify the provenance of the data workflow through registration and a verification test. A unique calculated hash generated from the data and workflow code will serve as a key for verification. This metadata will be generated at this step and together with the metadata associated with the assembly, uploaded to ENA/MGnify for further downstream functional annotation. Metadata to accompany the taxonomic inventories Metadata associated with the previous two steps will be summarised for inclusion with the taxonomic inventories (biom file format and CSV) for publication on the EMBRC GOs website.

- Metadata of the cleaned data; metadata associated with the data sequencing, sample collection (MIMS), and quality control analyses
- Metadata of the assembled data
- Metadata to accompany the taxonomic inventories

1.3.2 Ontologies & databases: the corner stone of mordern biology

Databases

- GenBank, ENA
- repositories such as MGnify
- PubMed

Ontologies:

- ENVO
- NCBI Taxonomy
- Gene Ontology
- Uniprot
- KEGG

1.4 Metabolic modeling at the omics era

1.4.1 Genome-scale metabolic model analysis

The relationship between genotype and phenotype is fundamental to biology. Many levels of control are introduced when moving from one to the other. Systems biology aims at deciphering "the strategy" both at the cell and at higher levels of organization, in case of multicell species, that enables organisms to produce orderly adaptive behavior in the face of widely varying genetic and environmental conditions ([16]); the term "strategy" is used as per [17]. Systems biology approaches aim at interpreting how a system's properties emerge; from the cell to the community level.

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

1.5 The hypersaline Tristomo swamp: a case study of an extreme environment

1.6 Systems biology from a computational resources point-of-view

1.7 Aims and objectives

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 computational.

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 3**, 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

2. microbial taxa and biological processes they occur

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

In **Chapter 4**, 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 5**, 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 microbial interactions for the assemblages were infered. All the methods developed and presented in the previous chapters were used to enhance the analysis of this microbiome.

In **Chapter 6**, 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 the **Conclusions** chapter, 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 Introduction

2.1.2 Contribution

2.1.3 Methods & Implementation

2.1.4 Results & Validation

2.1.5 Discussion

[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.

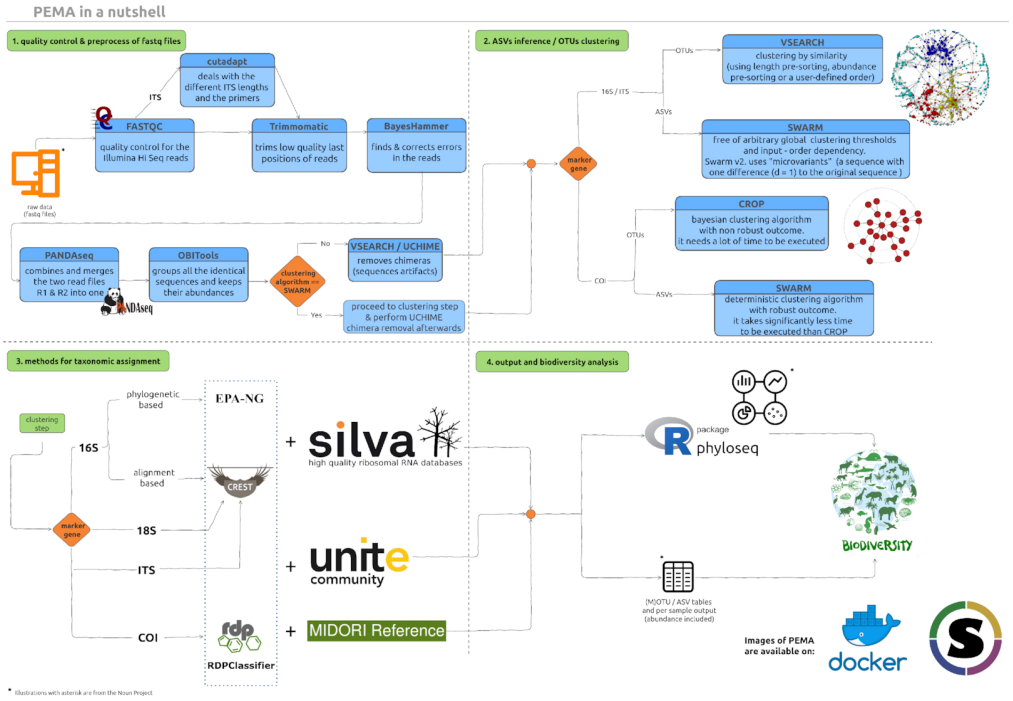

FIGURE 2.1: The PEMA workflow: figure from publication

2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data[2]

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 Introduction

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 [18]. However, it is essential that comprehensive public databases containing well curated, up-to-date sequences from voucher specimens are available [19]. 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 [20]. Nevertheless, there is not a

[2] For author contributions, please refer to the relevant section. Modified version of the published review.

truly "universal" genetic marker that is capable of being amplified for all species across different taxa [21]. Different markers have been used for different taxonomic groups [9]. 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 [22]. 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 [23]. There are cases where COI has been the standard marker for metabarcoding, such as in the assessment of freshwater macroinvertebrates [24] even though not all taxonomic groups can be differentiated to the species level using this locus [9]; for example, in case of fish other loci are widely used such as 12S rRNA gene (hereafter referred to as 12S rRNA) [25].

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 [26, 27].

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 [28]. Numts are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms [29].

Thus, as Mioduchowska et al. (2018) [30] highlight, when universal primers are used targeting the COI locus, it is possible to co-amplify both non-target numts and prokaryotes [31]. 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 [30].

Even though there are various known issues [14], COI is indeed considered as the "gold standard" for community DNA metabarcoding of bulk metazoan samples [32]; bulk is an environmental sample containing mainly organisms from the taxonomic group under study providing high quality and quantity of DNA [33]. However, as highlighted in the same study, this is not the case for eDNA samples. As Stat et al. (2017) [12] 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 [34, 35, 36]. Cold-adapted marine gammaproteobacteria are an indicative example for this case as shown by Siddall et al. (2009) [31].

2.2.2 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 [37]) 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 [31], 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., [38, 39]). 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) [40]. 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) [41] (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. [42].

2.2.3 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) [43]. 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 [44]. 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), 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 1.

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 [45]; 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 [46, 47].

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*

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

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 1). 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 [48].

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 [49] 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 [50, 51]. 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. The computational resources of the IMBBC High Performance Computing system, called Zorba [52], were exploited to address the needs of the tasks.

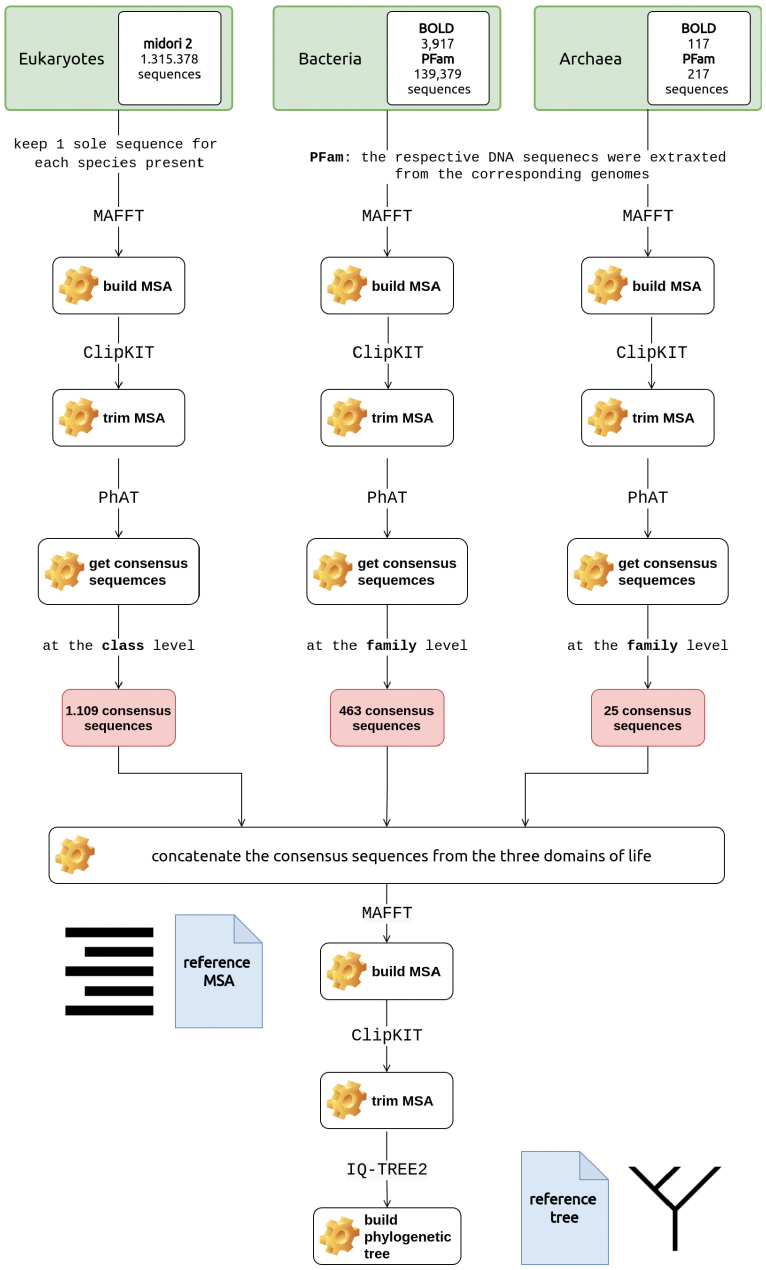

FIGURE 2.2: Overview of the approach followed to build the COI reference tree of life. Sequences were retrieved from Midori 2 (eukaryotes) and BOLD (bacteria and archaea) repositories. Consensus sequences at the family level were built for each domain specific dataset. MAFFT and consensus sequences at 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.

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 [53] 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 [54]. The query sequences are then split to build a discrete query MSA. Finally, the Evolutionary Placement Algorithm EPA-ng [41] 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 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 [55]. 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.

2.2.4 Results & Validation

Evaluation of the phylogenetic tree

The inferred phylogenetic tree is shown in Figure 2, with the bacterial (light blue) and archaeal (dark green) branches highlighted; in Suppl. material 3: Fig. 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 (Suppl. material 4: Fig. S2; the figure was built using the heat-tree module of the gappa examine tool).

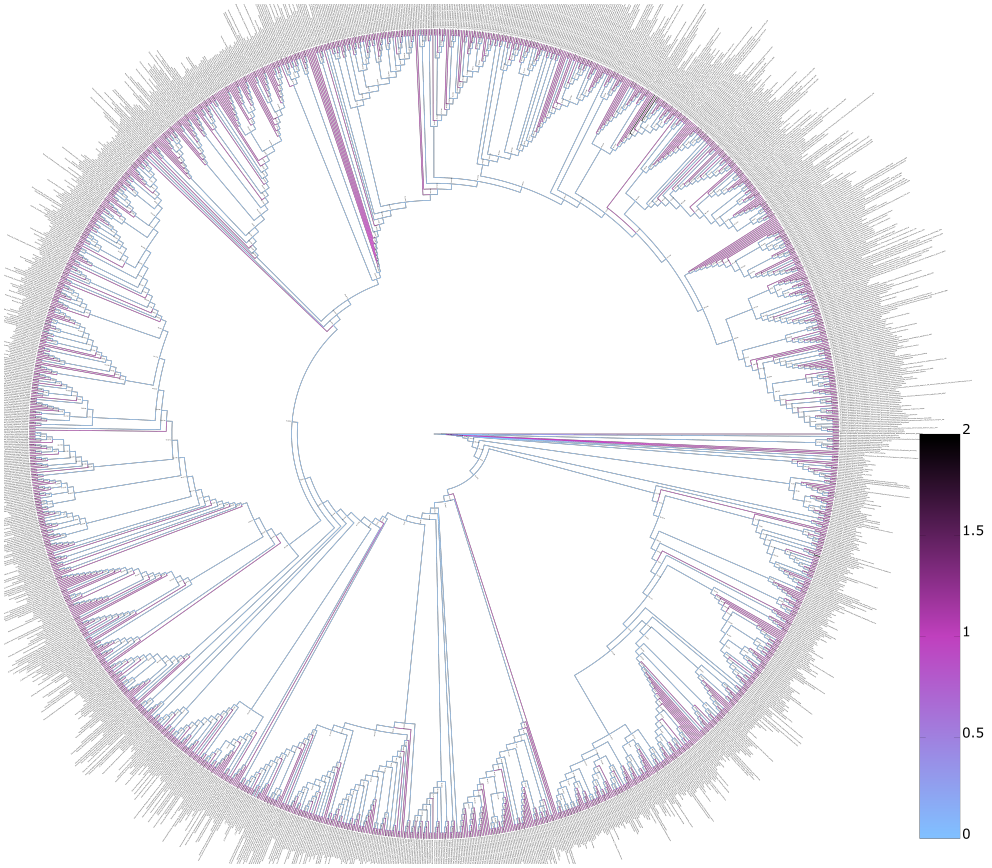

FIGURE 2.3: Phylogenetic tree of the consensus sequences retrieved; the tree that DARN makes use of. Light blue: bacterial branches. Dark green: archaeal branches. White: eukaryotic branches.

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). All bacterial sequences returned an incorrect eukaryotic assignment. Contrarily, when the phylogenetic tree was used, all the bacterial sequences were captured.

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).

More specifically, 57 marine, surface water, eDNA samples from Ireland were analysed through a. QIIME2 [56] and DADA2 [57] and, b. PEMA [58]. Similarly, 18 mangrove and 18 reef marine eDNA samples from Honduras, were analyzed using a. JAMP v0.74 and DnoisE [59] 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) [60] 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 [61] that PEMA invokes was set equal to 2 and 10 to cover a great range of different cases [62]. 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) in the different amplicon datasets used. A coherent visual representation of the DARN outcome for all the datasets is available here. 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). 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

Sample(s) accession number	Envo type	Sample type	Primer set	Amplicon length (bp)	Bioinfo pipeline(s)	# of ASVs	~% of sequence assignments per domain (if PEMA, using d = 10)			
							Eukaryotes	Bacteria	Archaea	distant
ERS6449795-ERS6449829	marine	eDNA	jgHCO2198 - jgLCO1490 and LoboF1 - LoboR1	658	QIIME2 - Dada2	13,376	11	88.0	0.02	0.003
					PEMA (d = 10)	39,454	25	75.0	0.1	0.4
ERS6463899–ERS6463901	marine reef				JAMP dada2 PEAR vsearch DnoisE	1,304	35	65.0	-	0.2
ERS6463906–ERS6463911 ERS6463913–ERS6463918 ERS6463920–ERS6463922		eDNA	mlCOIintF - jgHCO2198	313	PEMA (d = 10)	11,545	46	50.0	1	3
ERS6463744–ERS6463761	marine mangrove				JAMP dada2 PEAR vsearch DnoisE	663	40	60.0	-	0.6
					PEMA (d = 10)	5,879	49	47.0	1.0	2.0
ERR3460466	marine	bulk	mlCOIintF - jgHCO2198	313	PEMA (d = 2)	193	99	1	-	-
ERR3460467	marine	bulk				74	97	0.0	-	3
ERR3460470	marine	eDNA				184	71	28.0	0	1
ERS6488992	lentic	eDNA	fwhF2 - EPTDr2	142	PEMA (d = 10)	416	85	7	3	5
ERS6488993	lentic					315	99.2	0.4	0.4	-
ERS6488994	lentic					823	90	4	2	4
ERS6488995	lotic	eDNA	BF3 - BR2	458	PEMA (d = 10)	1,940	64	34.0	2	0.3

TABLE 2.2: 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.

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) 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 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.

2.2.5 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. [34, 37] 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 [34]. Furthermore, dark matter seems to be particularly common in eDNA as compared to bulk samples [32]. However, it should be mentioned that the high number of prokaryotic sequences in COI metabarcoding data is also reflecting known issues with contamination [63, 64, 65], incorrectly labeled reference sequences [66] and holobionts [67, 68] 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 [69, 25] 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 [70].

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.

2.3 A workflow for marine Genomic Observatories data analysis

Chapter 3

Software development to build a knowledge-base at the systems biology level

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

Publication relative to this chapter: under submission

3.1.1 Introduction

3.1.2 Contribution

3.1.3 Methods & Implementation

3.1.4 Results & Validation

3.1.5 Discussion

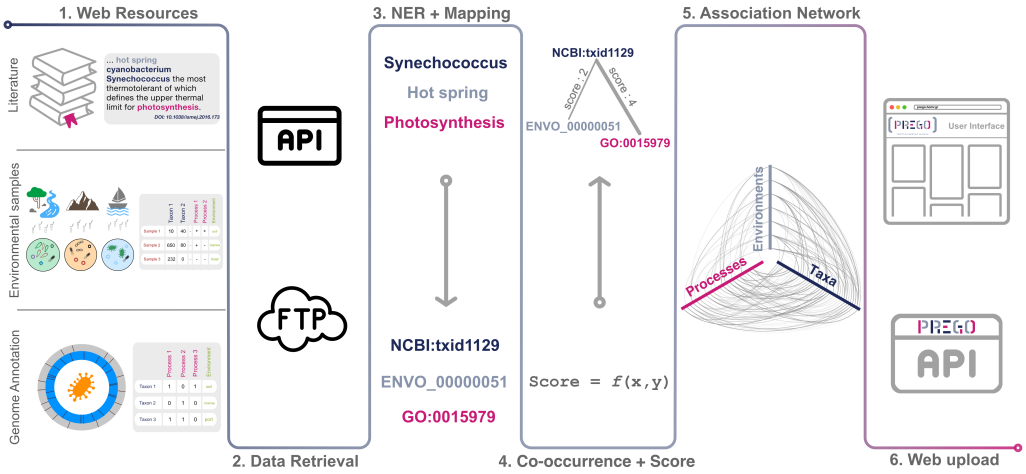

FIGURE 3.1: PREGO methodology: figure in the publication under submission

Chapter 4

Software development to establish metabolic flux sampling approaches at the community level

4.1 A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks

Publication relative to this chapter: [71]

4.1.1 Introduction

Systems Biology is a fundamental field and paradigm that represents a crucial era in Biology. Its functionality and usefulness rely on metabolic networks that model the reactions occurring inside an organism and provide the means to understand the underlying mechanisms that govern biological systems. We address the problem of sampling uniformly steady states of a metabolic network. We use a convex polytope to represent this set. However, the polytopes that result from biological data are of very high dimension (in the order of thousands) and in most, if not all, the cases are considerably skinny. Therefore, to perform uniform sampling efficiently in this setting, we need a novel algorithmic and computational framework specially tailored for the properties of metabolic networks. We present a complete software framework to handle sampling from convex polytopes that result from metabolic networks. Its backbone is a Multiphase Monte Carlo Sampling (MMCS) algorithm. We demonstrate the efficiency of our approach by performing extensive experiments on various metabolic networks. Notably, sampling on the most complicated human metabolic network accessible today, Recon3D, corresponding to a polytope of dimension 5 335, took less than 30 hours. To the best of our knowledge, that is out of reach for existing software.

But why being interested in such a task ? The genome of most bacteria are rather short to have issues like that.

However, MAGs can be brought together and build the metabolic model of a whole community!

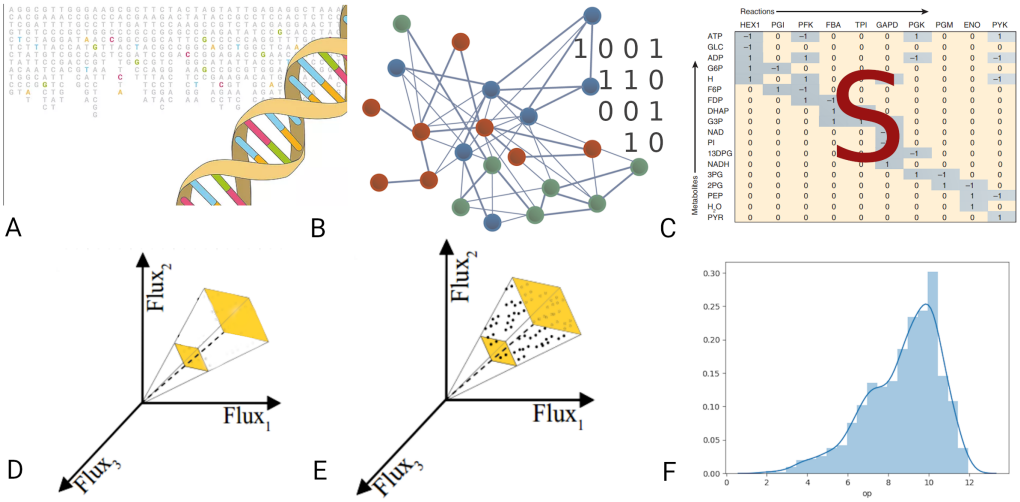

FIGURE 4.1: From DNA sequences to distributions of metabolic fluxes. (A) The genes of an organism provide us with the enzymes that it can potentially produce. Enzymes are like a blueprint for the reactions they can catalyze. (B) Using the enzymes we identify the reactions in the organism. (C) We construct the stoichiometric matrix of the metabolic model. (D) We consider the flux space under different conditions (e.g., steady states); they correspond to polytopes containing flux vectors addressing these conditions. (E) We sample from polytopes that are typically skinny and of high dimension. (F) The distribution of the flux of a reaction provides great insights to biologists.

4.1.2 Contribution

We introduce a Multi-phase Monte Carlo Sampling (MMCS) algorithm to sample from a polytope P. In particular, we split the sampling procedure in phases where, starting from P, each phase uses the sample to round the polytope and provide it as input to the next phase.

This improves the efficiency of the random walk in the next phase, see For sampling, we propose an improved variant of Billiard Walk that enjoys faster arithmetic complexity per step. We also handle efficiently the potential arithmetic inaccuracies near to the boundary, see [72] for a detailed treatment.

We accompany the MMCS algorithm with a powerful MCMC diagnostic, namely the estimation of Effective Sample Size (ESS), to identify a satisfactory convergence to the uniform distribution. However, our method is flexible and we can use any random walk and combination of MCMC diagnostics to decide convergence.

The open-source implementation of our algorithms[1] provides a complete software framework to sample efficiently in metabolic networks. We demonstrate the efficiency of our tools by performing experiments on almost all the metabolic networks that are

[1] https://github.com/GeomScale/volume_approximation/tree/v1.1.0-2

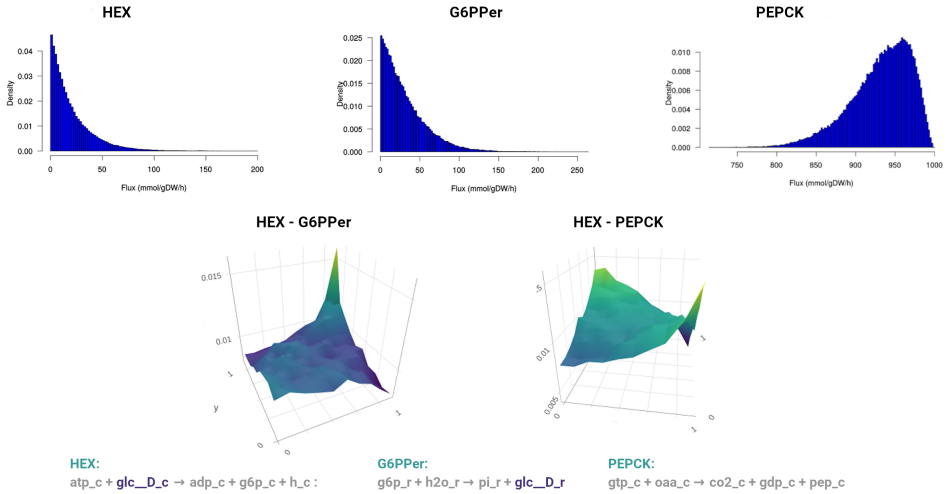

FIGURE 4.2: Flux distributions in the most recent human metabolic network Recon3D [4]. We estimate the flux distributions of the reactions catalyzed by the enzymes Hexokinase (D-Glucose:ATP) (HEX), Glucose-6-Phosphate Phosphatase, Edoplasmic Reticular (G6PPer) and Phosphoenolpyruvate carboxykinase (GTP) (PEPCK). As we sample steady states, the production rate of $glc\_\_D\_c$ should be equal to its consumption rate. Thus, in the corresponding copula, we see a positive dependency between HEX, i.e., the reaction that consumes $glc\_\_D\_c$ and G6PPer, that produces it. Furthermore, the PEPCK reaction operates when there is no $glc\_\_D\_c$ available and does not operate when the latter is present. Thus, in their copula we observe a negative dependency between HEX and PEPCK. A copula is a bivariate probability distribution for which the marginal probability distribution of each variable is uniform. It implies a positive dependency when the mass of the distribution concentrates along the up-diagonal (HEX - G6PPer) and a negative dependency when the mass is concentrated along the down-diagonal (HEX - PEPCK). The bottom line contains the reactions and their stoichiometry.

publicly available and by comparing with the state-of-the-art software packages, like cobra Our implementation is faster than cobra for low dimensional models, with a speed-up that ranges from 10 to 100 times; this gap on running times increases for bigger models We measure the quality of the sample our software produces using two widely used diagnostics, i.e., ESS and potential scale reduction factor (PSRF) [73]. The highlight of our method is the ability to sample from the most complicated human metabolic network that is accessible today, namely Recon3D. In Figure 1 we estimate marginal univariate and bivariate flux distributions in Recon3D which validate:

- the quality of the sample by confirming a mutually exclusive pair of biochemical pathways, and that
- our method indeed generates steady states.

In particular, our software can sample $1.44 \cdot 10^5$ points from a 5335-dimensional polytope in a day using modest hardware. This set of points suffices for the majority

of systems biology analytics. To our understanding this task is out of reach for existing software.

Lastly, MMCS algorithm is quite general sampling scheme and so it has the potential to also address other hard computational problems like multivariate integration and volume estimation of polytopes.

A preliminary version of this paper appeared in [71]. The current full version contains additional and more detailed experimental results, all the proofs of the various statements and theorems, the pseudocode of all the algorithms, an updated discussion of previous work, and a more detailed presentation of our approach and tools.

4.1.3 Methods & Implementation

Efficient Billiard walk

Algorithm 1: Billiard Walk(P, p, ρ, τ, W)

Require: polytope P; point $p \in P$; upper bound on the number of reflections ρ; parameter τ to adjust the length of the trajectory; walk length W.
Ensure: a point in P (uniformly distributed in P).
for $j = 1, \ldots, W$ **do**
$L \leftarrow -\tau \ln \eta$; $\eta \sim \mathcal{U}(0,1)$ { *length of the trajectory* } $i \leftarrow 0$ { *current number of reflections* } $p_0 \leftarrow p$ { *initial point of the step* } pick a uniform vector u_0 from the unit sphere { *initial direction*}
while $i \leq \rho$ **do**
$\ell \leftarrow \{p_i + t u_i, 0 \leq t \leq L\}$ { *this is a segment*}
if $\partial P \cap \ell = \emptyset$ **then**
$p_{i+1} \leftarrow p_i + L u_i$ **break**
end if
$p_{i+1} \leftarrow \partial P \cap \ell$; { *point update* }
the inner vector, s, of the tangent plane at p,
s.t. $||s|| = 1$, $L \leftarrow L - |P \cap \ell|$, $u_{i+1} \leftarrow u_i - 2(u_i^T s)s$ { *direction update* }
$i \leftarrow i + 1$
end while
if $i = \rho$ **then**
$p \leftarrow p_0$
else
$p \leftarrow p_i$
end if
end for
return p

At each step of Billiard Walk, we compute the intersection point of a ray, say $\ell := \{p + tu,\ t \in \mathbb{R}_+\}$, with the boundary of P, ∂P, and the normal vector of the tangent plane of P at the intersection point. The inner vector of the facet that the intersection point belongs to is a row of A. To compute the point $\partial P \cap \ell$ where the first reflection of a Billiard

Walk step takes place we need to compute the intersection of ℓ with all the hyperplanes that define the facets of P. This corresponds to solve (independently) the following m linear equations

$$a_j^T(p_0 + t_j u_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T u_0,\ j \in [k], \tag{4.1}$$

and keep the smallest positive t_j; a_j is the j-th row of the matrix A. We solve each equation in $\mathcal{O}(d)$ operations and so the overall complexity is $\mathcal{O}(dk)$, where k is the number of rows of A and thus an upper bound on the number of facets of P. A straightforward approach for Billiard Walk would consider that each reflection costs $\mathcal{O}(kd)$ and thus the per step cost is $\mathcal{O}(\rho kd)$. However, our improved version performs more efficiently both *point* and *direction updates* in pseudo-code by storing some computations from the previous iteration combined with a preprocessing step. The preprocessing step involves the normal vectors of the facets and takes $k^2 d$ operations. So the amortized per-step complexity of Billiard Walk becomes $\mathcal{O}((\rho + d)k)$. The pseudo-code appear in Algorithm 4.1.3.

Multiphase Monte Carlo Sampling algorithm

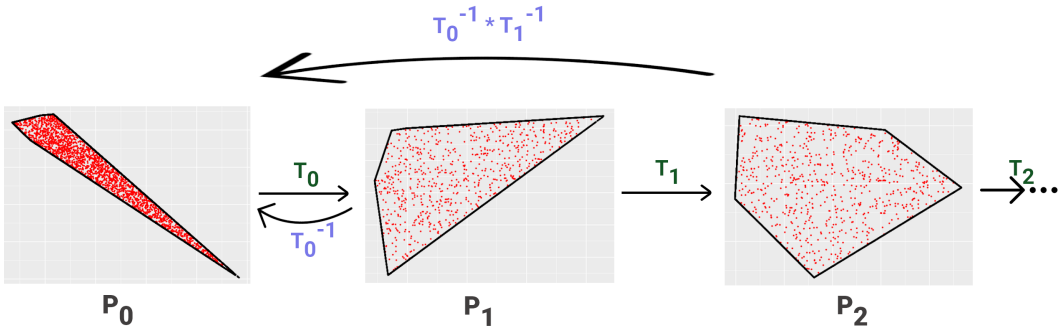

FIGURE 4.3: An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer n and starts at phase $i = 0$ sampling from P_0. In each phase it samples a maximum number of points λ. If the sum of Effective Sample Size in each phase becomes larger than n before the total number of samples in P_i reaches λ then the algorithm terminates. Otherwise, we proceed to a new phase. We map back to P_0 all the generated samples of each phase.

4.1.4 Results

4.1.5 Discussion

Flux sampling at the community level!

Chapter 5

Studying the microbiome as a whole: the way forward

Publication relative to this chapter: ongoing work, to be submitted before phd defense, probably not accepted by then though.

5.1 Microbial interactions inference in communities of a hypersaline swamp elucidate mechanisms governing taxonomic & functional profiles

5.1.1 Introduction

5.1.2 Contribution

5.1.3 Methods

darn and PEMA will be used at this point, among other software
PREGO and dingo will be used to this end

5.1.4 Results

5.1.5 Discussion

Chapter 6

An overview of the computational requirements & solutions in microbial ecology

6.1 0s and 1s in marine molecular research: a regional HPC perspective

Publication relative to this chapter: [52]

6.1.1 Introduction

6.1.2 Contribution

6.1.3 Methods

6.1.4 Results

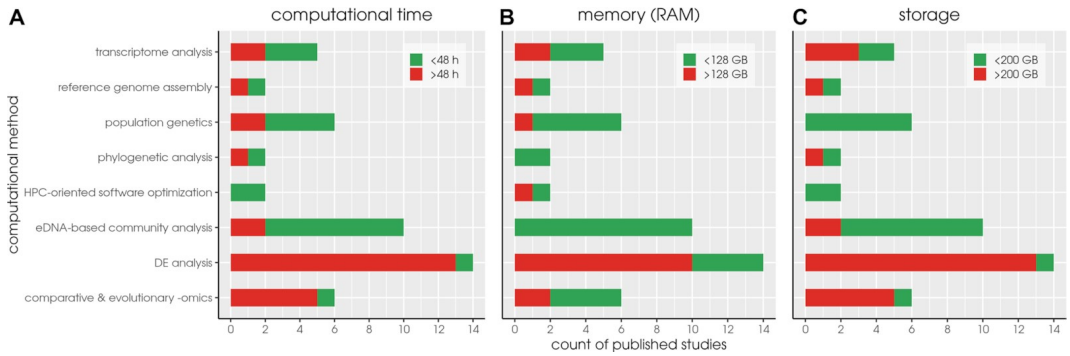

FIGURE 6.1: Computing requirements of the published studies performed on the IMBBC HPC facility over the last decade. Figure from publication.

6.1.5 Discussion

Chapter 7

Conclusions

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

Appendices

Bibliography

[1] R. Cavicchioli, W. J. Ripple, K. N. Timmis, F. Azam, L. R. Bakken, M. Baylis, M. J. Behrenfeld, A. Boetius, P. W. Boyd, A. T. Classen, *et al.*, "Scientists' warning to humanity: microorganisms and climate change," *Nature Reviews Microbiology*, vol. 17, no. 9, pp. 569–586, 2019.

[2] W. Commons, "File:sulfur cycle for hydrothermal vents.png — wikimedia commons, the free media repository," 2020. [Online; accessed 30-December-2021].

[3] W. Commons, "File:nitrogen cycle for hydrothermal vents.png — wikimedia commons, the free media repository," 2020. [Online; accessed 30-December-2021].

[4] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G. A. P. Gonzalez, M. K. Aurich, *et al.*, "Recon3D enables a three-dimensional view of gene variation in human metabolism," *Nature biotechnology*, vol. 36, no. 3, p. 272, 2018.

[5] P. G. Falkowski, T. Fenchel, and E. F. Delong, "The microbial engines that drive earth's biogeochemical cycles," *science*, vol. 320, no. 5879, pp. 1034–1039, 2008.

[6] Y. M. Bar-On, R. Phillips, and R. Milo, "The biomass distribution on earth," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. 6506–6511, 2018.

[7] H. A. Rees, A. C. Komor, W.-H. Yeh, J. Caetano-Lopes, M. Warman, A. S. Edge, and D. R. Liu, "Improving the dna specificity and applicability of base editing through protein engineering and protein delivery," *Nature communications*, vol. 8, no. 1, pp. 1–10, 2017.

[8] L. Röttjers and K. Faust, "From hairballs to hypotheses–biological insights from microbial networks," *FEMS microbiology reviews*, vol. 42, no. 6, pp. 761–780, 2018.

[9] K. Deiner, H. M. Bik, E. Mächler, M. Seymour, A. Lacoursière-Roussel, F. Altermatt, S. Creer, I. Bista, D. M. Lodge, N. De Vere, *et al.*, "Environmental dna metabarcoding: Transforming how we survey animal and plant communities," *Molecular ecology*, vol. 26, no. 21, pp. 5872–5895, 2017.

[10] K. M. Ruppert, R. J. Kline, and M. S. 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*, vol. 17, p. e00547, 2019.

[11] P. Taberlet, E. Coissac, M. Hajibabaei, and L. H. Rieseberg, "Environmental dna," 2012.

[12] M. Stat, M. J. Huggett, R. Bernasconi, J. D. DiBattista, T. E. Berry, S. J. Newman, E. S. Harvey, and M. Bunce, "Ecosystem biomonitoring with edna: metabarcoding across the tree of life in a tropical marine environment," *Scientific Reports*, vol. 7, no. 1, pp. 1–11, 2017.

[13] Y. Ji, L. Ashton, S. M. Pedley, D. P. Edwards, Y. Tang, A. Nakamura, R. Kitching, P. M. Dolman, P. Woodcock, F. A. Edwards, *et al.*, "Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding," *Ecology letters*, vol. 16, no. 10, pp. 1245–1257, 2013.

[14] B. E. Deagle, S. N. Jarman, E. Coissac, F. Pompanon, and P. Taberlet, "Dna metabarcoding and the cytochrome c oxidase subunit i marker: not a perfect match," *Biology letters*, vol. 10, no. 9, p. 20140562, 2014.

[15] P. Ten Hoopen, R. D. Finn, L. A. Bongo, E. Corre, B. Fosso, F. Meyer, A. Mitchell, E. Pelletier, G. Pesole, M. Santamaria, *et al.*, "The metagenomic data life-cycle: standards and best practices," *GigaScience*, vol. 6, no. 8, p. gix047, 2017.

[16] R. Strohman, "Maneuvering in the complex path from genotype to phenotype," *Science*, vol. 296, no. 5568, pp. 701–703, 2002.

[17] M. Polanyi, "Life's irreducible structure: Live mechanisms and information in dna are boundary conditions with a sequence of boundaries above them," *Science*, vol. 160, no. 3834, pp. 1308–1312, 1968.

[18] P. Taberlet, E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev, "Towards next-generation biodiversity assessment using dna metabarcoding," *Molecular ecology*, vol. 21, no. 8, pp. 2045–2050, 2012.

[19] T. Schenekar, M. Schletterer, L. A. Lecaudey, and S. 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*, vol. 36, no. 7, pp. 1004–1013, 2020.

[20] K. Cilleros, A. Valentini, L. Allard, T. Dejean, R. Etienne, G. Grenouillet, A. Iribar, P. Taberlet, R. Vigouroux, and S. Brosse, "Unlocking biodiversity and conservation studies in high-diversity environments using environmental dna (edna): A test with guianese freshwater fishes," *Molecular Ecology Resources*, vol. 19, no. 1, pp. 27–46, 2019.

[21] W. J. Kress, C. García-Robledo, M. Uriarte, and D. L. Erickson, "Dna barcodes for ecology, evolution, and conservation," *Trends in ecology & evolution*, vol. 30, no. 1, pp. 25–35, 2015.

[22] E. Coissac, T. Riaz, and N. Puillandre, "Bioinformatic challenges for dna metabarcoding of plants and animals," *Molecular ecology*, vol. 21, no. 8, pp. 1834–1847, 2012.

[23] P. D. Hebert, S. Ratnasingham, and J. 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*, vol. 270, no. suppl_1, pp. S96–S99, 2003.

[24] V. Elbrecht and F. Leese, "Validation and development of coi metabarcoding primers for freshwater macroinvertebrate bioassessment," *Frontiers in Environmental Science*, vol. 5, p. 11, 2017.

[25] M. Miya, R. O. Gotoh, and T. Sado, "Mifish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental dna and other samples," *Fisheries Science*, pp. 1–32, 2020.

[26] S. Ekici, G. Pawlik, E. Lohmeyer, H.-G. Koch, and F. Daldal, "Biogenesis of cbb3-type cytochrome c oxidase in rhodobacter capsulatus," *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, vol. 1817, no. 6, pp. 898–910, 2012.

[27] S. Schimo, I. Wittig, K. M. Pos, and B. Ludwig, "Cytochrome c oxidase biogenesis and metallochaperone interactions: steps in the assembly pathway of a bacterial complex," *PLoS One*, vol. 12, no. 1, p. e0170037, 2017.

[28] H. Song, J. E. Buhay, M. F. Whiting, and K. 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*, vol. 105, no. 36, pp. 13486–13491, 2008.

[29] D. Bensasson, D.-X. Zhang, D. L. Hartl, and G. M. Hewitt, "Mitochondrial pseudogenes: evolution's misplaced witnesses," *Trends in ecology & evolution*, vol. 16, no. 6, pp. 314–321, 2001.

[30] M. Mioduchowska, M. J. Czyż, B. Gołdyn, J. Kur, and J. Sell, "Instances of erroneous dna barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"?," *PLoS One*, vol. 13, no. 6, p. e0199609, 2018.

[31] M. E. Siddall, F. M. Fontanella, S. C. Watson, S. Kvist, and C. Erséus, "Barcoding bamboozled by bacteria: convergence to metazoan mitochondrial primer targets by marine microbes," *Systematic Biology*, vol. 58, no. 4, pp. 445–451, 2009.

[32] C. Andújar, P. Arribas, D. W. Yu, A. P. Vogler, and B. C. Emerson, "Why the coi barcode should be the community dna metabarcode for the metazoa," 2018.

[33] P. Taberlet, A. Bonin, L. Zinger, and E. Coissac, "Analysis of bulk samples," in *Environmental DNA*, pp. 140–143, Oxford University Press.

[34] C. Yang, Y. Ji, X. Wang, C. Yang, and W. Y. Douglas, "Testing three pipelines for 18s rdna-based metabarcoding of soil faunal diversity," *Science China Life Sciences*, vol. 56, no. 1, pp. 73–81, 2013.

[35] C. Yang, X. Wang, J. A. Miller, M. de Blécourt, Y. Ji, C. Yang, R. D. Harrison, and W. Y. Douglas, "Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator," *Ecological Indicators*, vol. 46, pp. 379–389, 2014.

[36] R. A. Collins, J. Bakker, O. S. Wangensteen, A. Z. Soto, L. Corrigan, D. W. Sims, M. J. Genner, and S. Mariani, "Non-specific amplification compromises environmental dna metabarcoding with coi," *Methods in Ecology and Evolution*, vol. 10, no. 11, pp. 1985–2001, 2019.

[37] E. Aylagas, Á. Borja, X. Irigoien, and N. Rodríguez-Ezpeleta, "Benchmarking dna metabarcoding for biodiversity-based monitoring and assessment," *Frontiers in Marine Science*, vol. 3, p. 96, 2016.

[38] F. Sinniger, J. Pawlowski, S. Harii, A. J. Gooday, H. Yamamoto, P. Chevaldonné, T. Cedhagen, G. Carvalho, and S. Creer, "Worldwide analysis of sedimentary dna reveals major gaps in taxonomic knowledge of deep-sea benthos," *Frontiers in Marine Science*, vol. 3, p. 92, 2016.

[39] Q. Haenel, O. Holovachov, U. Jondelius, P. Sundberg, and S. 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*, no. 5, 2017.

[40] G. Bernard, J. S. Pathmanathan, R. Lannes, P. Lopez, and E. Bapteste, "Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery," *Genome biology and evolution*, vol. 10, no. 3, pp. 707–715, 2018.

[41] P. Barbera, A. M. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri, and A. Stamatakis, "Epa-ng: massively parallel evolutionary placement of genetic sequences," *Systematic biology*, vol. 68, no. 2, pp. 365–369, 2019.

[42] M. Jamy, R. Foster, P. Barbera, L. Czech, A. Kozlov, A. Stamatakis, G. Bending, S. Hilton, D. Bass, and F. Burki, "Long-read metabarcoding of the eukaryotic rdna operon to phylogenetically and taxonomically resolve environmental diversity," *Molecular ecology resources*, vol. 20, no. 2, pp. 429–443, 2020.

[43] R. J. Machida, M. Leray, S.-L. Ho, and N. Knowlton, "Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples," *Scientific data*, vol. 4, no. 1, pp. 1–7, 2017.

[44] S. Ratnasingham and P. D. Hebert, "Bold: The barcode of life data system (http://www.barcodinglife. org)," *Molecular ecology notes*, vol. 7, no. 3, pp. 355–364, 2007.

[45] L. Czech, P. Barbera, and A. Stamatakis, "Methods for automatic reference trees and multilevel phylogenetic placement," *Bioinformatics*, vol. 35, no. 7, pp. 1151–1158, 2019.

[46] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform," *Nucleic acids research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[47] T. Nakamura, K. D. Yamada, K. Tomii, and K. Katoh, "Parallelization of mafft for large-scale multiple sequence alignments," *Bioinformatics*, vol. 34, no. 14, pp. 2490–2492, 2018.

[48] L. Czech, P. Barbera, and A. Stamatakis, "Genesis and gappa: processing, analyzing and visualizing phylogenetic (placement) data," *Bioinformatics*, vol. 36, no. 10, pp. 3263–3265, 2020.

[49] J. L. Steenwyk, T. J. Buida III, Y. Li, X.-X. Shen, and A. Rokas, "Clipkit: A multiple sequence alignment trimming software for accurate phylogenomic inference," *PLoS biology*, vol. 18, no. 12, p. e3001007, 2020.

[50] D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, and L. S. Vinh, "Ufboot2: improving the ultrafast bootstrap approximation," *Molecular biology and evolution*, vol. 35, no. 2, pp. 518–522, 2018.

[51] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear, "Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era," *Molecular biology and evolution*, vol. 37, no. 5, pp. 1530–1534, 2020.

[52] H. Zafeiropoulos, A. Gioti, S. Ninidakis, A. Potirakis, S. Paragkamian, N. Angelova, A. Antoniou, T. Danis, E. Kaitetzidou, P. Kasapidis, *et al.*, "0s and 1s in marine molecular research: a regional hpc perspective," *GigaScience*, vol. 10, no. 8, p. giab053, 2021.

[53] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, "Vsearch: a versatile open source tool for metagenomics," *PeerJ*, vol. 4, p. e2584, 2016.

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

[55] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a web browser," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–10, 2011.

[56] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, *et al.*, "Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science," tech. rep., PeerJ Preprints, 2018.

[57] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "Dada2: high-resolution sample inference from illumina amplicon data," *Nature methods*, vol. 13, no. 7, pp. 581–583, 2016.

[58] H. Zafeiropoulos, H. Q. Viet, K. Vasileiadou, A. Potirakis, C. Arvanitidis, P. Topalis, C. Pavloudi, and E. Pafilis, "Pema: a flexible pipeline for environmental dna metabarcoding analysis of the 16s/18s ribosomal rna, its, and coi marker genes," *GigaScience*, vol. 9, no. 3, p. giaa022, 2020.

[59] A. Antich, C. Palacin, O. S. Wangensteen, and X. Turon, "To denoise or to cluster, that is not the question: optimizing pipelines for coi metabarcoding and metaphylogeography," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–24, 2021.

[60] M. Obst, K. Exter, A. L. Allcock, C. Arvanitidis, A. Axberg, M. Bustamante, I. Cancio, D. Carreira-Flores, E. Chatzinikolaou, G. Chatzigeorgiou, *et al.*, "A marine biodiversity observation network for genetic monitoring of hard-bottom communities (arms-mbon)," *Frontiers in Marine Science*, vol. 7, p. 1031, 2020.

[61] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn, "Swarm v2: highly-scalable and high-resolution amplicon clustering," *PeerJ*, vol. 3, p. e1420, 2015.

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

[63] S. Kumar, M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter, "Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated gc-coverage plots," *Frontiers in genetics*, vol. 4, p. 237, 2013.

[64] S. M. Dittami and E. Corre, "Detection of bacterial contaminants and hybrid sequences in the genome of the kelp saccharina japonica using taxoblast," *PeerJ*, vol. 5, p. e4073, 2017.

[65] G. De Simone, A. Pasquadibisceglie, R. Proietto, F. Polticelli, S. Aime, H. JM Op den Camp, and P. Ascenzi, "Contaminations in (meta) genome data: An open issue for the scientific community," *IUBMB life*, vol. 72, no. 4, pp. 698–705, 2020.

[66] M. Steinegger and S. L. Salzberg, "Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in genbank," *Genome biology*, vol. 21, no. 1, pp. 1–12, 2020.

[67] S. F. Gilbert, J. Sapp, and A. I. Tauber, "A symbiotic view of life: we have never been individuals," *The Quarterly review of biology*, vol. 87, no. 4, pp. 325–341, 2012.

[68] E. Salvucci, "Microbiome, holobiont and the net of life," *Critical reviews in microbiology*, vol. 42, no. 3, pp. 485–494, 2016.

[69] H. Weigand, A. J. Beermann, F. Čiampor, F. O. Costa, Z. Csabai, S. Duarte, M. F. Geiger, M. Grabowski, F. Rimet, B. 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*, vol. 678, pp. 499–524, 2019.

[70] G. Huys, T. Vanhoutte, M. Joossens, A. S. Mahious, E. De Brandt, S. Vermeire, and J. Swings, "Coamplification of eukaryotic dna with 16s rrna gene-based pcr primers: possible consequences for population fingerprinting of complex microbial communities," *Current microbiology*, vol. 56, no. 6, pp. 553–557, 2008.

[71] A. Chalkis, V. Fisikopoulos, E. Tsigaridas, and H. Zafeiropoulos, "Geometric Algorithms for Sampling the Flux Space of Metabolic Networks," in *37th International Symposium on Computational Geometry (SoCG 2021)* (K. Buchin and E. Colin de Verdière, eds.), vol. 189 of *Leibniz International Proceedings in Informatics (LIPIcs)*, (Dagstuhl, Germany), pp. 21:1–21:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.

[72] A. Chevallier, S. Pion, and F. Cazals, "Hamiltonian Monte Carlo with boundary reflections, and application to polytope volume calculations," Research Report RR-9222, INRIA Sophia Antipolis, France, 2018.

[73] A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992. Publisher: Institute of Mathematical Statistics.