



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

<b>Preface</b>	<b>i</b>
<b>Contents</b>	<b>ii</b>
<b>Abstract</b>	<b>iv</b>
<b>Περίληψη</b>	<b>v</b>
<b>List of Figures and Tables</b>	<b>vi</b>
<b>List of Abbreviations and Symbols</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Microbial ecology . . . . .	1
1.1.1 Microbial communities: structure & function . . . . .	1
1.1.2 The role of microbial communities in biogeochemical cycles . . . . .	2
1.1.3 Microbial interactions: unravelling the microbiome . . . . .	2
1.2 The era of omics . . . . .	3
1.2.1 High Throughput Sequencing approaches . . . . .	3
1.2.2 Bioinformatics challenges . . . . .	3
1.3 Data integration & data mining in the era of omics . . . . .	4
1.3.1 Metadata: a key issue for the microbiome community . . . . .	4
1.3.2 Ontologies & databases: the corner stone of modern biology . . . . .	5
1.4 Metabolic modeling at the omics era . . . . .	6
1.4.1 Genome-scale metabolic model analysis . . . . .	6
1.4.2 Sampling the flux space of a metabolic model: challenges & potential . . . . .	6
1.5 The hypersaline Tristomo swamp: a case study of an extreme environment . . . . .	6
1.6 Systems biology from a computational resources point-of-view . . . . .	6
1.7 Aims and objectives . . . . .	6
<b>2 Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment</b>	<b>9</b>
2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes . . . . .	9
2.1.1 Introduction . . . . .	9
2.1.2 Contribution . . . . .	10
2.1.3 Methods & Implementation . . . . .	11
2.1.4 Results & Validation . . . . .	15
2.1.5 Discussion . . . . .	22

2.2	The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data . . . . .	25
2.2.1	Introduction . . . . .	25
2.2.2	Contribution . . . . .	26
2.2.3	Methods & Implementation . . . . .	26
2.2.4	Results & Validation . . . . .	30
2.2.5	Discussion . . . . .	34
2.3	A workflow for marine Genomic Observatories data analysis . . . . .	35
<b>3</b>	<b>Software development to build a knowledge-base at the systems biology level</b>	<b>37</b>
3.1	PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types . . . . .	37
3.1.1	Introduction . . . . .	37
3.1.2	Contribution . . . . .	37
3.1.3	Methods & Implementation . . . . .	37
3.1.4	Results & Validation . . . . .	37
3.1.5	Discussion . . . . .	37
<b>4</b>	<b>Software development to establish metabolic flux sampling approaches at the community level</b>	<b>39</b>
4.1	A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks	39
4.1.1	Introduction . . . . .	39
4.1.2	Contribution . . . . .	40
4.1.3	Methods & Implementation . . . . .	42
4.1.4	Results . . . . .	43
4.1.5	Discussion . . . . .	43
<b>5</b>	<b>Studying the microbiome as a whole: the way forward</b>	<b>45</b>
5.1	Microbial interactions inference in communities of a hypersaline swamp elucidate mechanisms governing taxonomic & functional profiles . . . . .	45
5.1.1	Introduction . . . . .	45
5.1.2	Contribution . . . . .	45
5.1.3	Methods . . . . .	45
5.1.4	Results . . . . .	45
5.1.5	Discussion . . . . .	45
<b>6</b>	<b>An overview of the computational requirements &amp; solutions in microbial ecology</b>	<b>47</b>
6.1	0s and 1s in marine molecular research: a regional HPC perspective . . . . .	47
6.1.1	Introduction . . . . .	47
6.1.2	Contribution . . . . .	47
6.1.3	Methods . . . . .	47
6.1.4	Results . . . . .	47
6.1.5	Discussion . . . . .	48
<b>7</b>	<b>Conclusions</b>	<b>49</b>
<b>Bibliography</b>		<b>53</b>

# **Abstract**

# Περίληψη

Και στα ελληνικά

# List of Figures and Tables

## List of Figures

1.1	Marine microbial communities contribute to CO <sub>2</sub> sequestration, nutrients recycle and thus to the release of CO <sub>2</sub> 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	Sulfur cycle. Figure taken from [2] . . . . .	2
1.3	Nitrogen cycle. Figure taken from [3] . . . . .	3
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 . . . . .	12
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. . . . .	14
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. [4], 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. . . . .	20

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. . . . .	28
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. . . . .	31
3.1	PREGO methodology: figure in the publication under submission . . . . .	38
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. . . . .	40
4.2	Flux distributions in the most recent human metabolic network Recon3D [5]. 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 <i>glc_D_c</i> 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 <i>glc_D_c</i> and G6PPer, that produces it. Furthermore, the PEPCK reaction operates when there is no <i>glc_D_c</i> 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. . . . .	41
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 $\lambda$ . If the sum of Effective Sample Size in each phase becomes larger than $n$ before the total number of samples in $P_i$ reaches $\lambda$ then the algorithm terminates. Otherwise, we proceed to a new phase. We map back to $P_0$ all the generated samples of each phase. . . . .	43

6.1 Computing requirements of the published studies performed on the IMBBC HPC facility over the last decade. Figure from publication. . . . .	47
--	----

## List of Tables

2.1 Summary benchmark of PEMA marker - gene – specific mock community recovery (precision) . . . . .	16
2.2 Comparison of the basic features of the different pipelines . . . . .	18
2.3 ~56 . . . . .	19
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. . . . .	21
2.5 Comparison of the taxonomy of retrieved MOTUs among PEMA, Barque, and the positive controls of Bista et al. [6] ; * Taxonomies identical to the published study (species level), ** Taxonomies identical to the published study (genus level). . . . .	22
2.6 Number of sequences and taxonomic species per domain of life and resources. The (#) symbols stands for "number". . . . .	27
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. . . . .	33

# **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 protozoa, are omnipresent and impact global ecosystem functions [7] through their abundance [8], versatility [9] and interactions [10].

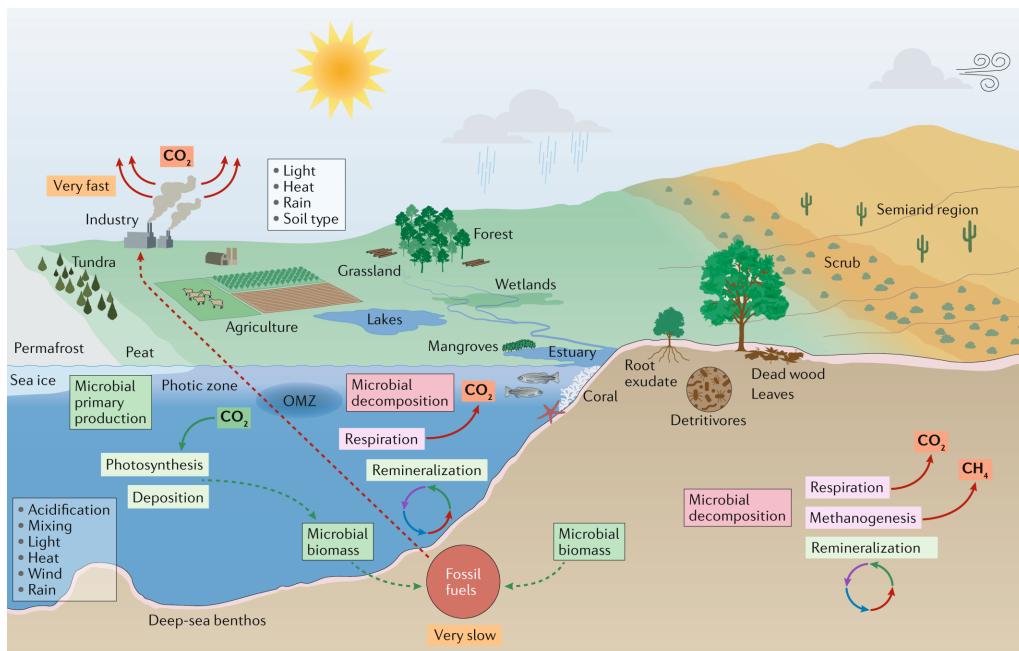


FIGURE 1.1: Marine microbial communities contribute to CO<sub>2</sub> sequestration, nutrients recycle and thus to the release of CO<sub>2</sub> to the atmosphere. Soil microbial communities decomposers organic matter and release nutrients in the soil from [1] doi: [10.1038/s41579-019-0222-5](https://doi.org/10.1038/s41579-019-0222-5), under Creative Commons Attribution 4.0 International License

## 1. INTRODUCTION

---

### 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]

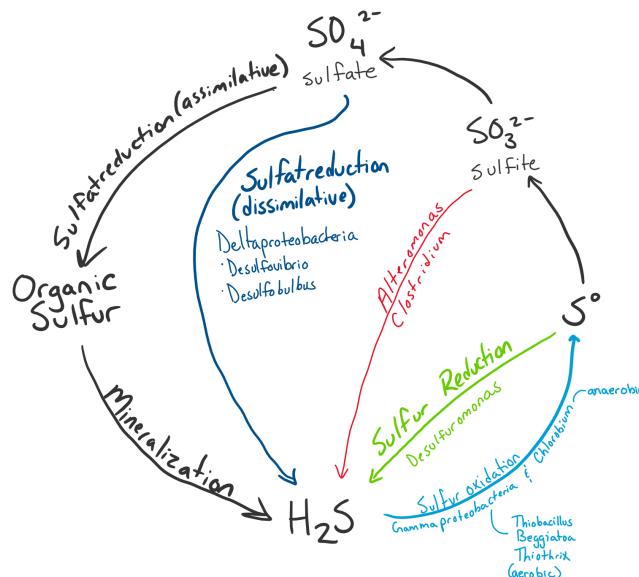
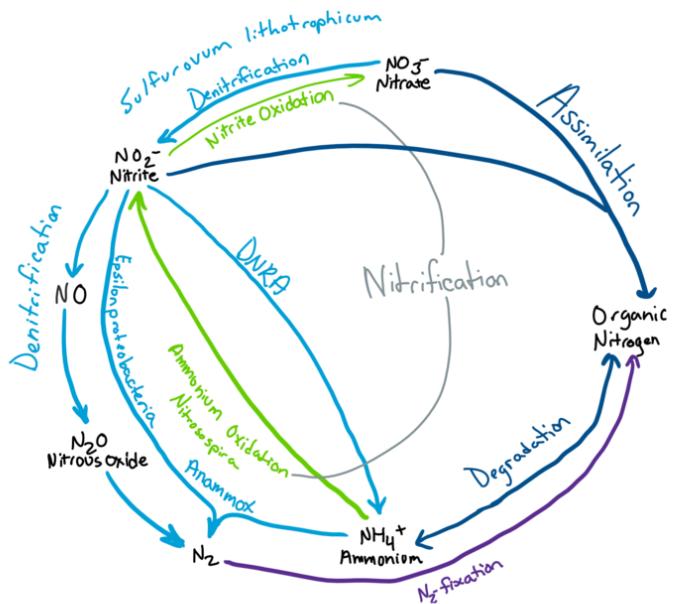


FIGURE 1.2: Sulfur cycle. Figure taken from [2]

### 1.1.3 Microbial interactions: unravelling the microbiome



**DNRA** = dissimilatory nitrate reduction to ammonium

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 [11, 12]. Environmental DNA (eDNA) metabarcoding, targeting DNA directly isolated from environmental samples (e.g., water, soil or sediment, [13]), is considered a holistic approach [14] in terms of biodiversity assessment, providing high detection capacity. At the same time, it allows wide scale rapid bio-assessment [14] at a relatively low cost as compared to traditional biodiversity survey methods [15]. 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 [16].

### 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" [17] 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).

### 1.3. Data integration & data mining in the era of omics

---

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 modern 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 ([18]); the term "strategy" is used as per [19]. 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, combining 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

## 1.7. Aims and objectives

---

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 inferred. 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<sup>1</sup>

### 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](https://doi.org/10.1093/gigascience/giaa022).

### 2.1.1 Introduction

Environmental DNA (eDNA) metabarcoding inaugurates a new era in bio- and eco-monitoring [20]. eDNA refers to genetic material obtained directly from environmental samples (soil, sediment, water, etc.) without any obvious signs of biological source material [21]. 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 [15]. eDNA metabarcoding attempts to turn the page on the way biodiversity is perceived and monitored [15]. This combination is considered to be a potential holistic

---

<sup>1</sup>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.

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

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 [22], QIIME 2 [23], LotuS [24]). 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](#), ScreenForBio [25], and PIPITS [26]. However, there is still 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](#)) to serve the broad biological community via commonly accessible infrastructures.

### 2.1.2 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 [27] 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 [28] and Singularity [29], 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 [30]. 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 [30].

## 2.1. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

---

As high-throughput sequencing (HTS) data become more and more accurate, ASVs, i.e., marker gene amplified sequence reads that differ in  $\geq 1$  nucleotide from each other, become easier to resolve [15] [31]. The use of ASVs instead of OTUs has been suggested [31]; however, the choice of which approach to use should be based on each study's objective(s) [32].

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 [33] and CROP [34], are used for the clustering of reads in (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 [35] allows ASV inference in all cases.

Taxonomic assignment is performed in an alignment-based approach, making use of the CREST LCAClassifier [36] and the Silva database [37] for the case of 16S and 18S rRNA marker genes; the Unite database [38] is used for the ITS gene. In the 16S marker gene case, phylogeny-based assignment is also supported, based on RAxML-ng [39], EPA-ng [40], and Silva [37]. For the COI marker gene, the RDPClassifier [41] and the MIDORI database [42] are used for the taxonomic assignment. In addition, ecological and phylogenetic analysis are facilitated via the phyloseq R package [43].

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](#).

### 2.1.3 Methods & Implementation

PEMA's architecture comprises 4 main parts taking place in tandem (Fig. 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 [44] 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 [45] 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 [46] 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 [47], an algorithm of the SPAdes assembly toolkit [48], revises incorrectly called bases. PANDAseq [49] assembles the overlapping paired-end reads, and then the obiuniq program of OBITools [50] groups all the identical sequences in every sample, keeping track of their abundances. The VSEARCH package [33] 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).

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

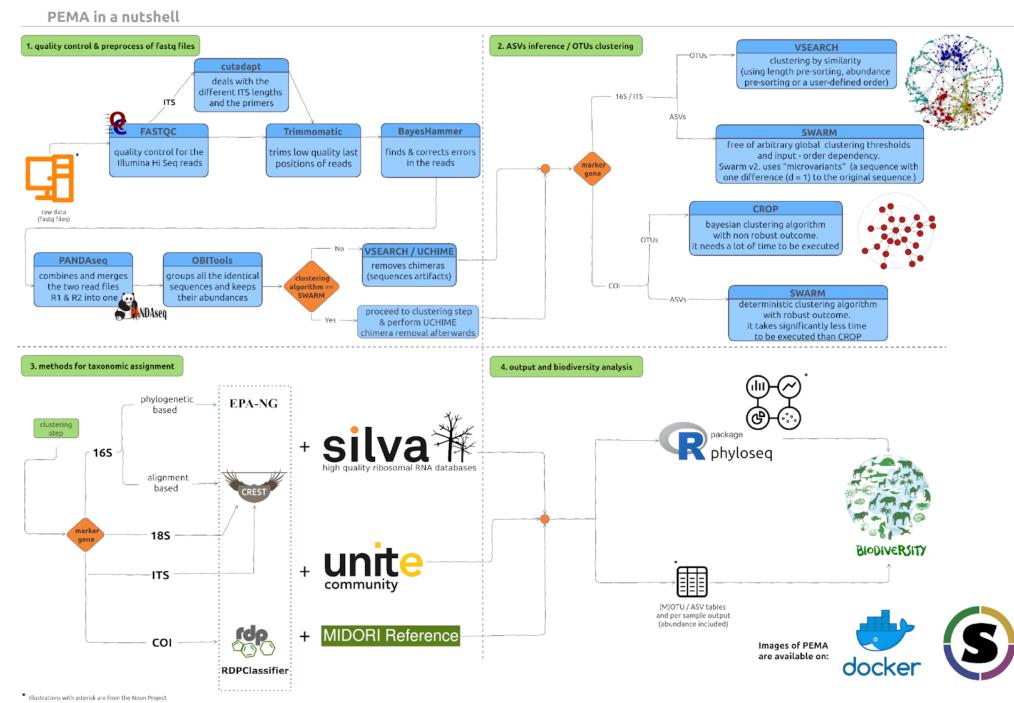


FIGURE 2.1: PEMA comprises 4 parts. The first step (top left) is the quality control and pre-processing of the Illumina sequencing reads. This step is common for both 16S rRNA and COI marker genes. The second step (top right) is the clustering of reads to (M)OTUs or their inferring to ASVs. The third 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

### 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 [33] is used for OTU clustering, while ASVs can be identified by the Swarm v2 algorithm [35]. VSEARCH is an accurate and fast tool that can handle large datasets; at the same time it is a great alternative for USEARCH [51] because it is distributed under an open source license.

For the ITS and COI marker genes, CROP [34], 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 [34]. ASV inference is conducted by Swarm v2 [35]

## 2.1. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

---

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 [37] and the Unite [52] 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 [37] versions new PEMA versions will be available.

For the COI marker gene, PEMA uses the RDPClassifier [41] and the MIDORI reference database [42] to assign taxonomy of the MOTUs. The MIDORI database contains quality-controlled metazoan mitochondrial gene sequences from GenBank [53].

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 [39] and gappa [phat algorithm] [54], Fig. 2A). PaPaRa [55] and EPA-ng [40] 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 [56] (Fig. 2B).

### 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 [43], 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 [43] output, a multiple sequence alignment (MSA) and a phylogenetic tree of the OTU/ASVs retrieved can be returned; for the MSA, the MAFFT [57, 58] aligner is invoked while the latter is built by RAxML-ng [39].

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

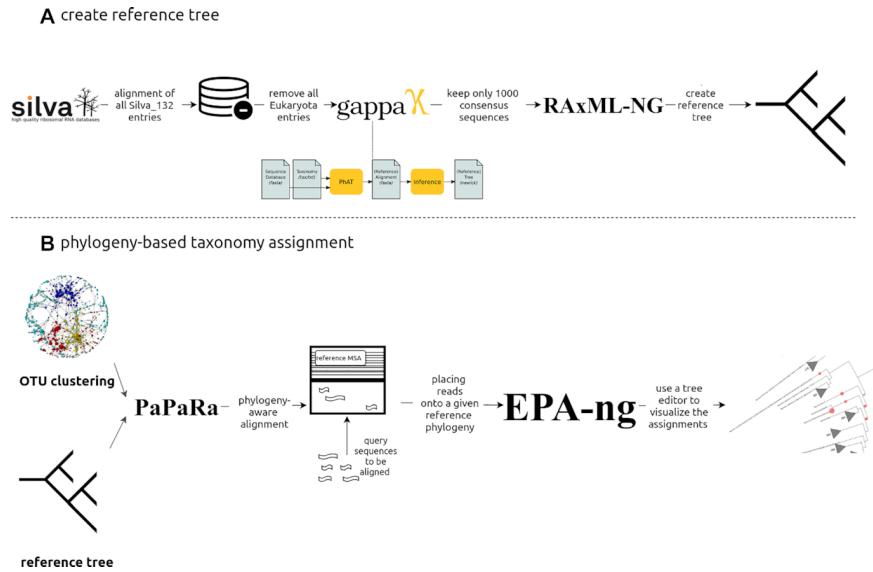


FIGURE 2.2: Phylogeny - based taxonomy assignment. A: Building a reference tree for the phylogeny-based taxonomy assignment to 16S rRNA marker gene OTUs: from the latest edition of Silva SSU, all entries referring to Bacteria and Archaea were used and using the “art” algorithm, 10,000 consensus taxa 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.

### PEMA container-based installation

An easy way of installing PEMA is via its containers. A Dockerized PEMA version is available. Singularity users can *pull* the PEMA image from as described in [PEMA GitHub repository](#). Between the 2 containers, the Singularity-based one is recommended for HPC environments owing to Singularity’s improved security and file accessing properties, see <https://dev.to/grokcode/singularity-a-docker-for-hpc-environments-i6p>. PEMA can also be found in the bio.tools (id: PEMA) and SciCrunch (PEMA, [RRID:SCR\\_017676](#)) databases. For detailed documentation, see [here](#).

### 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. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

---

### PEMA modules added after publication

<sup>2</sup>

#### 2.1.4 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. [59] with 20 different bacterial species was studied. Correspondingly, in the case of the 18S rRNA marker gene, a mock community of Bradley et al. [60] with 12 algal species was used; for the ITS, one of Bakker [61] including 19 different fungal taxa; and for the case of the COI marker gene, a mock community of Bista et al. [62] 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. [4] 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. [6] 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* [63] 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

---

<sup>2</sup>REMEMBER TO WRITE THIS PART

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

<b>Marker gene</b>	<b>Precision</b>	<b>Recall</b>	<b>F1</b>
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)

is the precision and recall harmonic mean and is calculated by means of the following formula:  $F1 = 2 \times (precision \times recall) / (precision + recall)$  [64].

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 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. [59], 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 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).

## 2.1. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

---

### 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 [61], PEMA identified the genus Chytromyces 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 [61], 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 [62] 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. [62] 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

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

<b>Feature</b>	<b>LotuS</b>	<b>QIIME 2</b>	<b>mothur</b>	<b>Barque</b>	<b>PEMA</b>
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

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 [22], QIIME 2 [23], LotuS [24] and Barque. Table 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.

## 2.1. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

QIIME 2						
Parameter	LotuS	mothur	Deblur	DADA2	PEMA	Pavloudi et al. [4]
No. of OTUs	9,849	142,669	517	1,023	6,028	7,050
Execution time (h)	~9	~67 <sup>3</sup>	2.5	~5	~1.5	~26

TABLE 2.3: ~56

### 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. [4] dataset with mothur [22], QIIME 2 [23], LotuS [24] and PEMA was conducted.

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 [27], 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 3. LotuS and PEMA resulted in a final number of OTUs comparable to that of Pavloudi et. al [4]. 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 Fig. 3. The phyla that were found in the samples are similar to the ones that were found in the original study [4]. 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 [4], the general trend

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

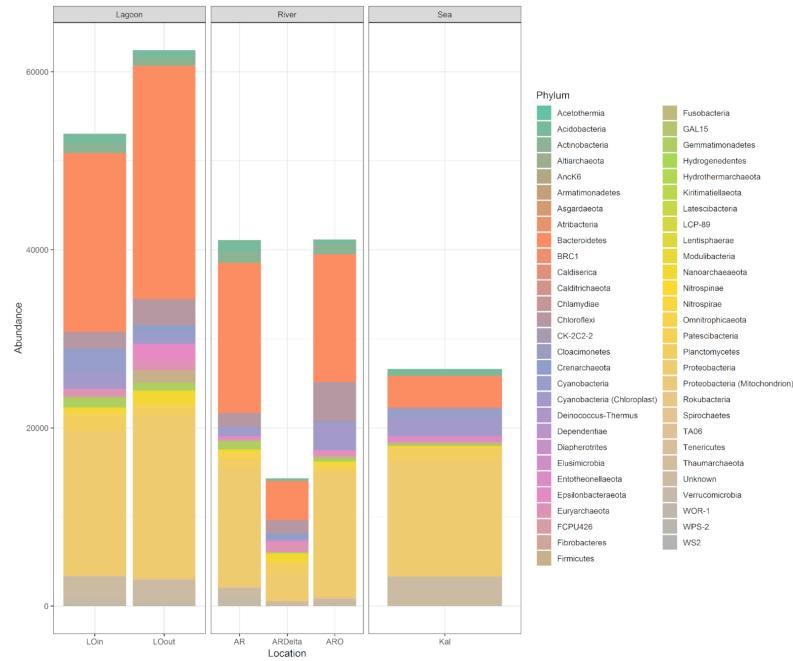


FIGURE 2.3: OTU bar plot at the phylum level. Bar plot depicting the taxonomy of the retrieved OTUs from PEMA for the dataset of Pavloudi et al. [4], at the phylum level for the case of the 16S marker gene. AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; Loin: Logarou station inside the lagoon; LOout: Logarou station in the channel connecting the lagoon to the gulf; Kal: Kalamitsi.

of a decreasing number of OTUs with increasing salinity was observed as in the original study (Supplementary Fig. 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 Figs S2–S4), with an extreme case of missing the order of Betaproteobacterales (Supplementary Figs 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. [4]. However, this was not the case with Deblur (PERMANOVA:  $P > 0.05$ ). Overall, PEMA's output is in accordance with the original study [4], and seen through this perspective PEMA performed equally well with the other tested pipelines, along with having the shortest execution time.

### COI marker gene analysis evaluation

Bista et al. [6] 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

## 2.1. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

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

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. [6] dataset with Barque; 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. [6] study VSEARCH [33] was used with a clustering at 97% similarity threshold. Afterwards, the BLAST+ (megablast) algorithm [65] 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 [6]. 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 4); according to Mahé et al. [35], this is the most important parameter because it affects the number of MOTUs that are being created. The resulting MOTUs were classified against the MIDORI reference database [42] using RDPClassifier [41]. The results of the processing of the sequences are reported in Supplementary Table S3. For the case of Barque, the BOLD Database was used [66].

As shown in Table 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 [6] 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 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 [47]), 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 5).

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

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

TABLE 2.5: Comparison of the taxonomy of retrieved MOTUs among PEMA, Barque, and the positive controls of Bista et al. [6]; \* Taxonomies identical to the published study (species level), \*\* Taxonomies identical to the published study (genus level).

### 2.1.5 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. [31], "ASV methods infer the biological sequences in the sample prior to the introduction of amplification and sequencing 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 [31]. Swarm v2 produces all possible *microvariants* of an amplicon to implement an exact-string comparison [35]. 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. [67], which variant-based approaches manage to address. However

## 2.1. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

---

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](#) 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 [68].

### 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](#), [Life-Watch ERIC](#), [EMBRC ERIC](#)). 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

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

microbial pathology and gut microbiome, in line with modern research needs, from low volume to big data.

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

### 2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data<sup>4</sup>

#### 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](https://doi.org/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 [69]. However, it is essential that comprehensive public databases containing well curated, up-to-date sequences from voucher specimens are available [70]. 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-barcode, is only as good and accurate as the reference databases [71]. Nevertheless, there is not a truly “universal” genetic marker that is capable of being amplified for all species across different taxa [72]. Different markers have been used for different taxonomic groups [11]. 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 [30]. 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 [73]. There are cases where COI has been the standard marker for metabarcoding, such as in the assessment of freshwater macroinvertebrates [74] even though not all taxonomic groups can be differentiated to the species level using this locus [11]; for example, in case of fish other loci are widely used such as 12S rRNA gene (hereafter referred to as 12S rRNA) [75].

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 [76, 77].

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

<sup>4</sup>For author contributions, please refer to the relevant section. Modified version of the published review.

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

studies, leading to overestimates of the number of taxa present in a sample [78]. Numts are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms [79].

Thus, as Mioduchowska et al. (2018) [80] highlight, when universal primers are used targeting the COI locus, it is possible to co-amplify both non-target numts and prokaryotes [81]. 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 [80].

Even though there are various known issues [16], COI is indeed considered as the “gold standard” for community DNA metabarcoding of bulk metazoan samples [82]; bulk is an environmental sample containing mainly organisms from the taxonomic group under study providing high quality and quantity of DNA [83]. However, as highlighted in the same study, this is not the case for eDNA samples. As Stat et al. (2017) [14] 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 [84, 85, 86]. Cold-adapted marine gammaproteobacteria are an indicative example for this case as shown by Siddall et al. (2009) [81].

### 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 [87]) 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 [81], 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., [88, 89]). 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) [90]. 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) [40] (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. [91].

### 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

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

(version: GB239) [42]. 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 [66]. 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 [54]; 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 [57, 58].

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 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 [92].

A total of 1,109 consensus sequences (70% of total consensus sequences) were built

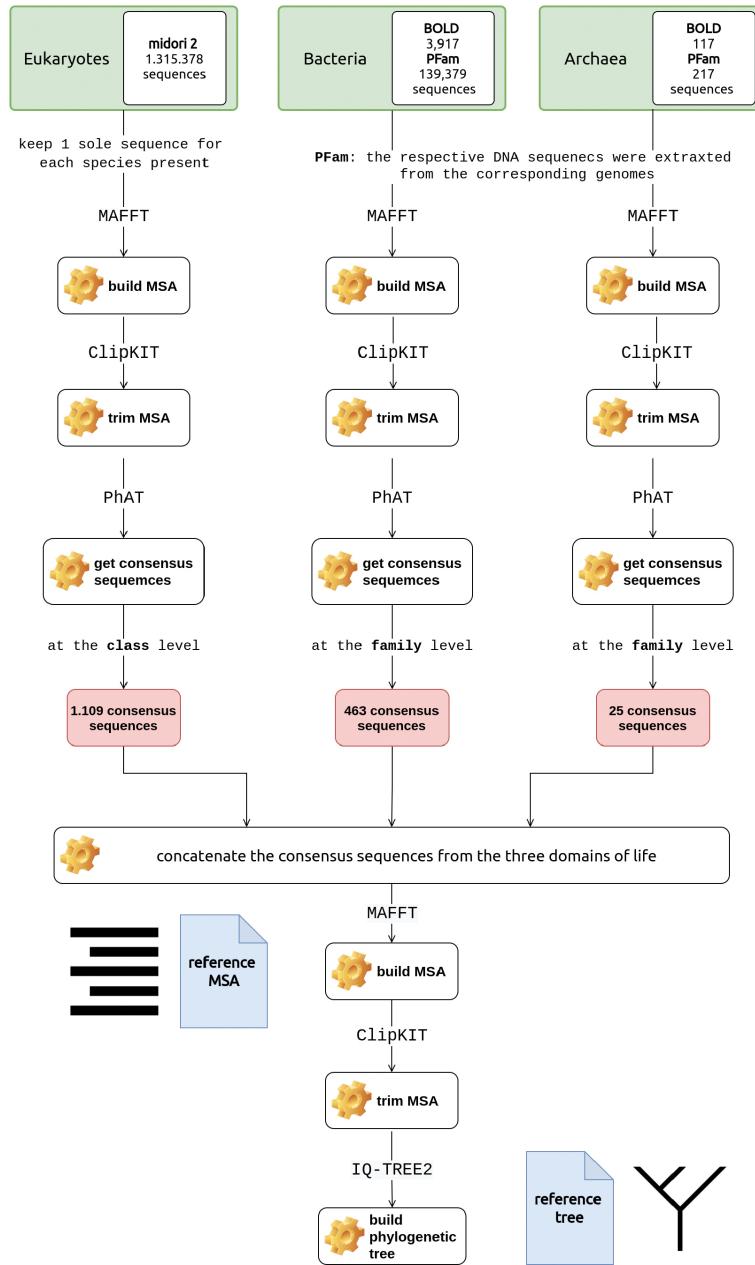


FIGURE 2.4: Overview of the approach followed to build the COI reference tree of life. Sequences were retrieved from Midori 2 (eukaryotes) and BOLD (bacteria and archaea) repositories. Consensus sequences at the family level were built for each domain specific dataset. MAFFT and consensus sequences at 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.

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

---

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 [93] 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 [94, 95]. 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 [96], 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 [33] 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 [55]. The query sequences are then split to build a discrete query MSA. Finally, the Evolutionary Placement Algorithm EPA-ng [40] 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

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

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 [97]. 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).

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

---

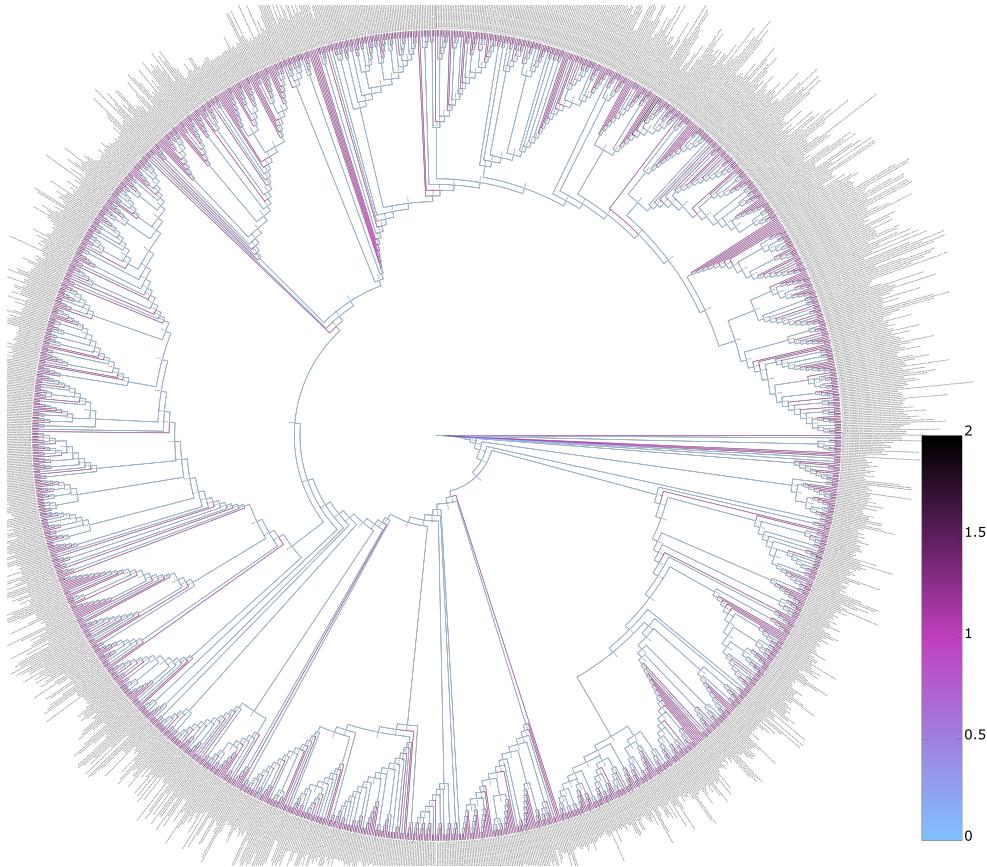


FIGURE 2.5: Phylogenetic tree of the consensus sequences retrieved; the tree that DARN makes use of. Light blue: bacterial branches. Dark green: archaeal branches. White: eukaryotic branches.

### 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](#)). 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 [23] and DADA2 [98] and, b. PEMA [99]. Similarly, 18 mangrove and 18 reef marine eDNA samples from Honduras, were analyzed using a. [JAMP v0.74](#) and DnoisE [100] 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) [101] 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 [35] that PEMA invokes was set equal to 2 and 10 to cover a great range of different cases [102]. 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	Env type	Sample type	Primer set	Amplicon length (bp)	Bioinfo pipeline(s)	# of ASVs	~% of sequence assignments per domain (if PEMMA, using d = 10)			
							Eukaryotes	Bacteria	Archaea   distant	
ERS6449795– ERS6449829	marine	eDNA	jgHCO2198 - jgLCO1490 and LoboF1 - LoboR1	658	QIIME2 - Dada2 PEMA (d = 10) JAMP dada2 PEAR vsearch DnoisE	13,376	11	88.0	0.02	0.003
ERS6463899– ERS6463901	marine reef	eDNA	mlCOLintF - jgHCO2198	313	PEMA (d = 10)	11,545	46	50.0	1	3
ERS6463906– ERS6463911 ERS6463913– ERS6463918 ERS6463920– ERS6463922					JAMP dada2 PEAR vsearch DnoisE PEMA (d = 10)	663	40	60.0	-	0.6
ERS6463744– ERS6463761	marine mangrove					5,879	49	47.0	1.0	2.0
ERR3460466 ERR3460467 ERR3460470 ERS6488992 ERS6488993 ERS6488994 ERS6488995	marine marine marine lentic lentic lentic lotic	bulk bulk eDNA eDNA ePTDr2 eDNA	mlCOLintF - jgHCO2198 fwhF2 - EPTDr2 BF3 - BR2	313 142 458	PEMA (d = 2) PEMA (d = 10) PEMA (d = 10)	193 74 416 315 823 1,940	99 97 71 85 90 64	1 0.0 28.0 7 0.4 4	- - 0 3 0.4 2	- - 1 5 - 4

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.

## 2. SOFTWARE DEVELOPMENT TO ESTABLISH QUALITY HTS-ORIENTED BIOINFORMATICS METHODS FOR MICROBIAL DIVERSITY ASSESSMENT

---

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. [84, 87] 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 [84]. Furthermore, dark matter seems to be particularly common in eDNA as compared to bulk samples [82]. However, it should be mentioned that the high number of prokaryotic sequences in COI metabarcoding data is also reflecting known issues with contamination [103, 104, 105], incorrectly labeled reference sequences [106] and holobionts [107, 108] 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 [109, 75] 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 [110].

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

### **2.3. A workflow for marine Genomic Observatories data analysis**

---

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**

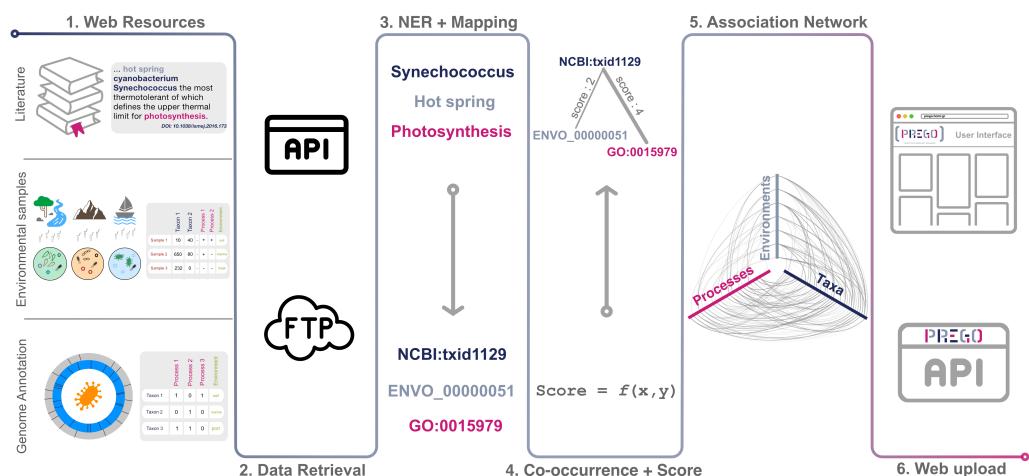


FIGURE 3.1: PREGO methodology: figure in the publication under submission

# **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: [111]

#### **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 5335, 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!

#### 4. SOFTWARE DEVELOPMENT TO ESTABLISH METABOLIC FLUX SAMPLING APPROACHES AT THE COMMUNITY LEVEL

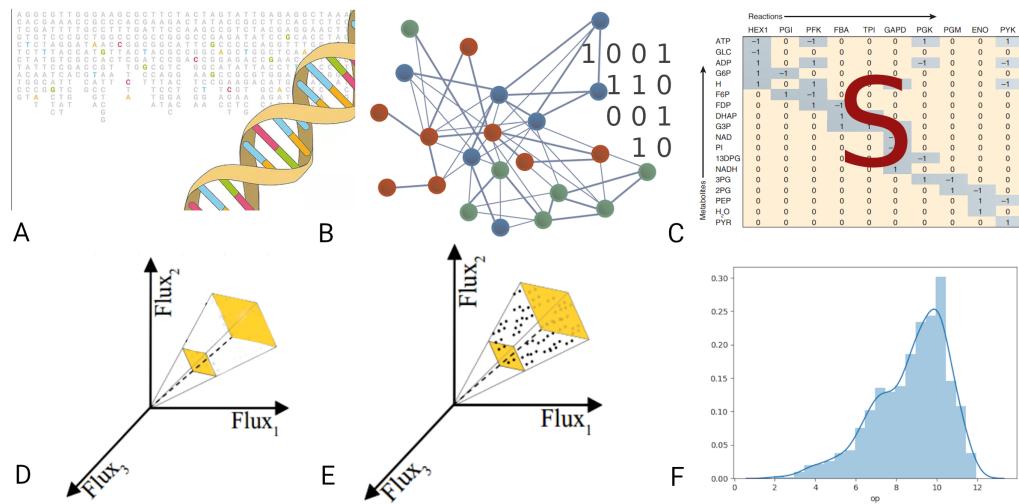


FIGURE 4.1: From DNA sequences to distributions of metabolic fluxes. (A) The genes of an organism provide us with the enzymes that it can potentially produce. Enzymes are like a blueprint for the reactions they can catalyze. (B) Using the enzymes we identify the reactions in the organism. (C) We construct 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 [112] 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<sup>1</sup> 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

<sup>1</sup>[https://github.com/GeomScale/volume\\_approximation/tree/v1.1.0-2](https://github.com/GeomScale/volume_approximation/tree/v1.1.0-2)

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

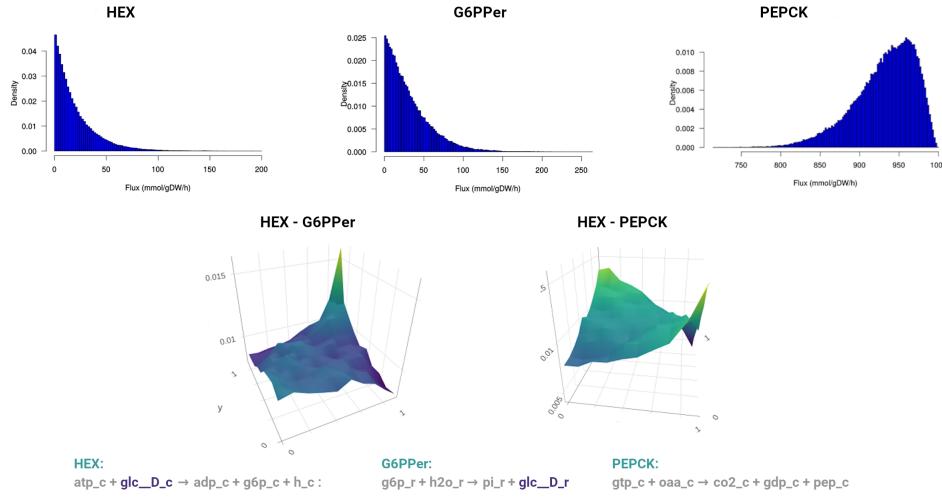


FIGURE 4.2: Flux distributions in the most recent human metabolic network Recon3D [5]. We estimate the flux distributions of the reactions catalyzed by the enzymes Hexokinase (D-Glucose:ATP) (HEX), Glucose-6-Phosphate Phosphatase, Edoplasmic Reticular (G6PPer) and Phosphoenolpyruvate carboxykinase (GTP) (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) [113]. 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

#### 4. SOFTWARE DEVELOPMENT TO ESTABLISH METABOLIC FLUX SAMPLING APPROACHES AT THE COMMUNITY LEVEL

---

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 [111]. 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( $P, p, \rho, \tau, W$ )

**Require:** polytope  $P$ ; point  $p \in P$ ; upper bound on the number of reflections  $\rho$ ; parameter  $\tau$  to adjust the length of the trajectory; walk length  $W$ .

**Ensure:** a point in  $P$  (uniformly distributed in  $P$ ).

```

for  $j = 1, \dots, W$  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
         $\ell \leftarrow \{p_i + tu_i, 0 \leq t \leq L\}$  {this is a segment}
        if  $\partial P \cap \ell = \emptyset$  then
             $p_{i+1} \leftarrow p_i + Lu_i$  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.  $\|s\| = 1$ ,  $L \leftarrow L - |\ell \cap \partial P|$ ,  $u_{i+1} \leftarrow u_i - 2(u_i^T s)s$  {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  $P$ ,  $\partial P$ , 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

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

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

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

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

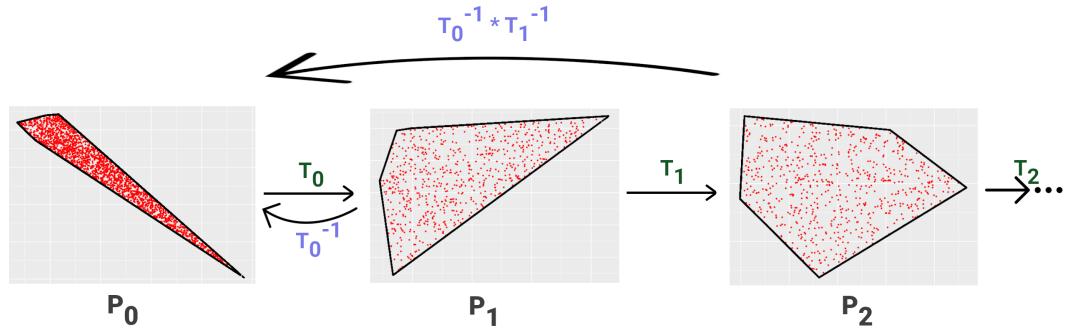


FIGURE 4.3: An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer  $n$  and starts at phase  $i = 0$  sampling from  $P_0$ . In each phase it samples a maximum number of points  $\lambda$ . If the sum of Effective Sample Size in each phase becomes larger than  $n$  before the total number of samples in  $P_i$  reaches  $\lambda$  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: [96]

#### 6.1.1 Introduction

#### 6.1.2 Contribution

#### 6.1.3 Methods

#### 6.1.4 Results

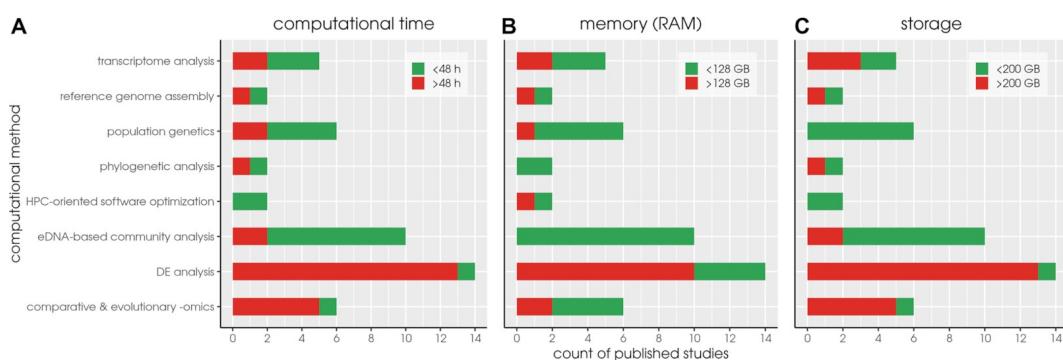


FIGURE 6.1: Computing requirements of the published studies performed on the IMBBC HPC facility over the last decade. Figure from publication.

6. AN OVERVIEW OF THE COMPUTATIONAL REQUIREMENTS & SOLUTIONS IN MICROBIAL ECOLOGY

---

**6.1.5 Discussion**

# **Chapter 7**

## **Conclusions**

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



# **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] C. Pavloudi, J. B. Kristoffersen, A. Oulas, M. De Troch, and C. Arvanitidis, “Sediment microbial taxonomic and functional diversity in a natural salinity gradient challenge remane’s “species minimum” concept,” *PeerJ*, vol. 5, p. e3687, 2017.
- [5] 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.
- [6] I. Bista, G. Carvalho, K. Walsh, M. Seymour, M. Hajibabaei, D. Lallias, M. Christmas, and S. Creer, “Annual time-series analysis of aqueous dna reveals ecologically relevant dynamics of lake ecosystem biodiversity. nat. commun. 8, 14087,” 2017.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.

## BIBLIOGRAPHY

---

- [11] 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.
- [12] 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.
- [13] P. Taberlet, E. Coissac, M. Hajibabaei, and L. H. Rieseberg, “Environmental dna,” 2012.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] R. Strohman, “Maneuvering in the complex path from genotype to phenotype,” *Science*, vol. 296, no. 5568, pp. 701–703, 2002.
- [19] 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.
- [20] A. Pavan-Kumar, P. Gireesh-Babu, and W. Lakra, “Dna metabarcoding: a new approach for rapid biodiversity assessment,” *J Cell Sci Mol Biol*, vol. 2, no. 1, p. 111, 2015.
- [21] P. F. Thomsen and E. Willerslev, “Environmental dna—an emerging tool in conservation for monitoring past and present biodiversity,” *Biological conservation*, vol. 183, pp. 4–18, 2015.
- [22] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, “Introducing mothur: open-source, platform-independent, community-supported software for

- describing and comparing microbial communities,” *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.
- [23] 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.
  - [24] F. Hildebrand, R. Tadeo, A. Y. Voigt, P. Bork, and J. Raes, “Lotus: an efficient and user-friendly otu processing pipeline,” *Microbiome*, vol. 2, no. 1, pp. 1–7, 2014.
  - [25] J. Axtner, A. Crampton-Platt, L. A. Hörig, A. Mohamed, C. C. Xu, D. W. Yu, and A. Wilting, “An efficient and robust laboratory workflow and tetrapod database for larger scale environmental dna studies,” *GigaScience*, vol. 8, no. 4, p. giz029, 2019.
  - [26] H. S. Gweon, A. Oliver, J. Taylor, T. Booth, M. Gibbs, D. S. Read, R. I. Griffiths, and K. Schonrogge, “Pipits: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the illumina sequencing platform,” *Methods in ecology and evolution*, vol. 6, no. 8, pp. 973–980, 2015.
  - [27] P. Cingolani, R. Sladek, and M. Blanchette, “Bigdatascript: a scripting language for data pipelines,” *Bioinformatics*, vol. 31, no. 1, pp. 10–16, 2015.
  - [28] B. B. Rad, H. J. Bhatti, and M. Ahmadi, “An introduction to docker and analysis of its performance,” *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 17, no. 3, p. 228, 2017.
  - [29] G. M. Kurtzer, V. Sochat, and M. W. Bauer, “Singularity: Scientific containers for mobility of compute,” *PloS one*, vol. 12, no. 5, p. e0177459, 2017.
  - [30] 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.
  - [31] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis,” *The ISME journal*, vol. 11, no. 12, pp. 2639–2643, 2017.
  - [32] C. Pauvert, M. Buée, V. Laval, V. Edel-Hermann, L. Fauchery, A. Gautier, I. Lesur, J. Vallance, and C. Vacher, “Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline,” *Fungal Ecology*, vol. 41, pp. 23–33, 2019.
  - [33] 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.
  - [34] X. Hao, R. Jiang, and T. Chen, “Clustering 16s rrna for otu prediction: a method of unsupervised bayesian clustering,” *Bioinformatics*, vol. 27, no. 5, pp. 611–618, 2011.

## BIBLIOGRAPHY

---

- [35] 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.
- [36] A. Lanzén, S. L. Jørgensen, D. H. Huson, M. Gorfer, S. H. Grindhaug, I. Jonassen, L. Øvreås, and T. Urich, “Crest–classification resources for environmental sequence tags,” *PloS one*, vol. 7, no. 11, p. e49334, 2012.
- [37] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, “The silva ribosomal rna gene database project: improved data processing and web-based tools,” *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.
- [38] M. C. Rillig, M. Ryo, A. Lehmann, C. A. Aguilar-Trigueros, S. Buchert, A. Wulf, A. Iwasaki, J. Roy, and G. Yang, “The role of multiple global change factors in driving soil functions and microbial biodiversity,” *Science*, vol. 366, no. 6467, pp. 886–890, 2019.
- [39] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, “Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference,” *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, 2019.
- [40] 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.
- [41] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, “Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy,” *Applied and environmental microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.
- [42] 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.
- [43] P. J. McMurdie and S. Holmes, “phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data,” *PloS one*, vol. 8, no. 4, p. e61217, 2013.
- [44] “Fastqc,” Jun 2015.
- [45] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.
- [46] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet. journal*, vol. 17, no. 1, pp. 10–12, 2011.
- [47] S. I. Nikolenko, A. I. Korobeynikov, and M. A. Alekseyev, “Bayeshammer: Bayesian clustering for error correction in single-cell sequencing,” in *BMC genomics*, vol. 14, pp. 1–11, Springer, 2013.

- [48] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, *et al.*, “Spades: a new genome assembly algorithm and its applications to single-cell sequencing,” *Journal of computational biology*, vol. 19, no. 5, pp. 455–477, 2012.
- [49] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld, “Pandaseq: paired-end assembler for illumina sequences,” *BMC bioinformatics*, vol. 13, no. 1, pp. 1–7, 2012.
- [50] F. Boyer, C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac, “obitoools: A unix-inspired software package for dna metabarcoding,” *Molecular ecology resources*, vol. 16, no. 1, pp. 176–182, 2016.
- [51] R. C. Edgar, “Search and clustering orders of magnitude faster than blast,” *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.
- [52] R. H. Nilsson, K.-H. Larsson, A. F. S. Taylor, J. Bengtsson-Palme, T. S. Jeppesen, D. Schigel, P. Kennedy, K. Picard, F. O. Glöckner, L. Tedersoo, *et al.*, “The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications,” *Nucleic acids research*, vol. 47, no. D1, pp. D259–D264, 2019.
- [53] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, and E. W. Sayers, “Genbank.,” *Nucleic acids research*, vol. 46, no. D1, pp. D41–D47, 2018.
- [54] 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.
- [55] S. A. Berger and A. Stamatakis, “Papara 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension,” *Heidelberg Institute for Theoretical Studies*, 2012.
- [56] I. Letunic and P. Bork, “Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation,” *Nucleic acids research*, vol. 49, no. W1, pp. W293–W296, 2021.
- [57] 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.
- [58] 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.
- [59] D. M. Gohl, P. Vangay, J. Garbe, A. MacLean, A. Hauge, A. Becker, T. J. Gould, J. B. Clayton, T. J. Johnson, R. Hunter, *et al.*, “Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies,” *Nature biotechnology*, vol. 34, no. 9, pp. 942–949, 2016.

## BIBLIOGRAPHY

---

- [60] I. M. Bradley, A. J. Pinto, J. 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*, vol. 82, no. 19, pp. 5878–5891, 2016.
- [61] M. G. Bakker, “A fungal mock community control for amplicon sequencing experiments,” *Molecular ecology resources*, vol. 18, no. 3, pp. 541–556, 2018.
- [62] I. Bista, G. R. Carvalho, M. Tang, K. Walsh, X. Zhou, M. Hajibabaei, S. Shokralla, M. Seymour, D. Bradley, S. Liu, *et al.*, “Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples,” *Molecular Ecology Resources*, vol. 18, no. 5, pp. 1020–1034, 2018.
- [63] P. W. Harrison, B. Alako, C. Amid, A. Cerdeño-Tárraga, I. Cleland, S. Holt, A. Hussein, S. Jayathilaka, S. Kay, T. Keane, *et al.*, “The european nucleotide archive in 2018,” *Nucleic acids research*, vol. 47, no. D1, pp. D84–D88, 2019.
- [64] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.
- [65] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, “Blast+: architecture and applications,” *BMC bioinformatics*, vol. 10, no. 1, pp. 1–9, 2009.
- [66] 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.
- [67] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn, “Swarm: robust and fast clustering method for amplicon-based studies,” *PeerJ*, vol. 2, p. e593, 2014.
- [68] S. I. Glassman and J. B. Martiny, “Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units,” *MSphere*, vol. 3, no. 4, pp. e00148–18, 2018.
- [69] 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.
- [70] 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.
- [71] 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.

- [72] 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.
- [73] P. D. Hebert, S. Ratnasingham, and J. R. De Waard, “Barcode 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.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] 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.
- [80] 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.
- [81] M. E. Siddall, F. M. Fontanella, S. C. Watson, S. Kvist, and C. Erséus, “Barcode bamboozled by bacteria: convergence to metazoan mitochondrial primer targets by marine microbes,” *Systematic Biology*, vol. 58, no. 4, pp. 445–451, 2009.
- [82] 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.
- [83] P. Taberlet, A. Bonin, L. Zinger, and E. Coissac, “Analysis of bulk samples,” in *Environmental DNA*, pp. 140–143, Oxford University Press.

## BIBLIOGRAPHY

---

- [84] 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.
- [85] 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.
- [86] 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.
- [87] 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.
- [88] F. Sinniger, J. Pawłowski, 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.
- [89] 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.
- [90] G. Bernard, J. S. Pathmanathan, R. Lannes, P. Lopez, and E. Baptiste, “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.
- [91] 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.
- [92] 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.
- [93] 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.
- [94] 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.

- [95] 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.
- [96] 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.
- [97] 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.
- [98] 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.
- [99] 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.
- [100] 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 metaphylogeny,” *BMC bioinformatics*, vol. 22, no. 1, pp. 1–24, 2021.
- [101] 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.
- [102] S. Kamenova, “A flexible pipeline combining clustering and correction tools for prokaryotic and eukaryotic metabarcoding. peer community in ecology 1: 100043,” 2020.
- [103] 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.
- [104] 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.
- [105] 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.

## BIBLIOGRAPHY

---

- [106] 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.
- [107] 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.
- [108] E. Salvucci, “Microbiome, holobiont and the net of life,” *Critical reviews in microbiology*, vol. 42, no. 3, pp. 485–494, 2016.
- [109] H. Weigand, A. J. Beermann, F. Ćiampor, 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.
- [110] 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.
- [111] 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.
- [112] 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.
- [113] 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.