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

Here I ll talk about people.

Haris Zafeiropoulos

Contents

Preface i

Contents ii

Abstract v

Περίληψη vi

List of Figures and Tables vii

List of Abbreviations and Symbols xi

1 Introduction 1

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 mordern 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

2 Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment 9

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 Abstract . . . 9

2.1.2 Introduction . . . 10

2.1.3 Contribution . . . 11

2.1.4 Methods & Implementation . . . 12

2.1.5 Results & Validation . . . 15

2.1.6 Discussion . . . 23
2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data . . . 25
2.2.1 Abstract . . . 25
2.2.2 Introduction . . . 25
2.2.3 Contribution . . . 27
2.2.4 Methods & Implementation . . . 27
2.2.5 Results & Validation . . . 31
2.2.6 Discussion . . . 35
2.3 A workflow for marine Genomic Observatories data analysis . . . 35

3 Software development to build a knowledge-base at the systems biology level 37
3.1 PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types . . . 37
3.1.1 Abstract . . . 37
3.1.2 Introduction . . . 37
3.1.3 Contribution . . . 37
3.1.4 Methods & Implementation . . . 37
3.1.5 Results & Validation . . . 37
3.1.6 Discussion . . . 37

4 Software development to establish metabolic flux sampling approaches at the community level 43
4.1 A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks 43
4.1.1 Introduction . . . 43
4.1.2 Contribution . . . 44
4.1.3 Methods & Implementation . . . 46
4.1.4 Results . . . 47
4.1.5 Discussion . . . 47

5 Studying the microbiome as a whole: the way forward 49
5.1 Microbial interactions inference in communities of a hypersaline swamp elucidate mechanisms governing taxonomic & functional profiles . . . 49
5.1.1 Introduction . . . 49
5.1.2 Contribution . . . 49
5.1.3 Methods . . . 49
5.1.4 Results . . . 49
5.1.5 Discussion . . . 49

6 An overview of the computational requirements & solutions in microbial ecology 51
6.1 0s and 1s in marine molecular research: a regional HPC perspective . . . 51
6.1.1 Abstract . . . 51
6.1.2 Introduction . . . 52
6.1.3 Contribution . . . 53
6.1.4 Methods . . . 54
6.1.5 Results . . . 56

6.1.6 Discussion 59
6.1.7 Conclusions 66

7 Conclusions **67**

Bibliography **71**

Short CV **86**

Abstract

Short description on this dissertation.

Περίληψη

Και στα ελληνικά

List of Figures and Tables

List of Figures

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

1.2 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. 21

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. 29

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. 32

3.1 PREGO analysis methodology. PREGO periodically retrieves three distinct types of data from open access resources. Scientific text, environmental sample data and genomic annotations are handled with respective methodologies in order to standardize their entities. Named Entity Recognition and Co-occurrence analysis is the common framework in order to build a weighted association network with nodes being the entity identifiers. Lastly, all these associations are available through a Web interface and an API. All these steps have been implemented in an autonomous way with regular cycles of updates (see Appendix B). 38

3.2 PREGO web user interface. (a) There are two search fields, plain text and taxa sequences. (b-d) three associations tabs each one presenting associations of the querred entity with the respective entities, Environments (b), Biological Process (c) and Molecular Function (d). Three channels of information are distinguishing the associations based on the original data. (e) Documents tab presents the scientific articles that mention the queried entity highlighted with color. (f) Downloads tab provides the associations of each channel (when available) to be downloaded in JSON and TSV format. 39

3.3 Each association is supported by original data. (a) Literature channel has a pop-up functionality that displays the scientific articles that each specific association occurs with highlighted color. (b) Environmental Samples channel redirects to the samples that support the specific association (currently only is supported MGnify). (c) Annotated Genomes channel similarly redirects to the isolates ids that each association is based on (both Struo and JGI IMG are supported). 40

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. 44

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. 45

4.3 An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer n and starts at phase $i = 0$ sampling from P_0. In each phase it samples a maximum number of points λ. If the sum of Effective Sample Size in each phase becomes larger than n before the total number of samples in P_i reaches λ then the algorithm terminates. Otherwise, we proceed to a new phase. We map back to P_0 all the generated samples of each phase. . 47

6.1 Block diagram of the *Zorba* architecture. This is the IMBBC HPC facility architecture in its current setup, after 12 years of development. There are 2 login nodes and 1 intermediate where users may develop their analyses. Computational nodes are split into 4 partitions with different specs and policy terms: bigmem supporting processes requiring up to 640 GB RAM, batch handling mostly (but not exclusively) parallel-driven jobs (either in a single node or across several nodes), minibatch aiming to serve parallel jobs with reduced resource requirements, and fast partition for non-intensive jobs. All servers, except file systems, run Debian 9 (kernel 4.9.0-8-amd64). CCBY icons from the Noun Project: "nfs file document icon" by IYIKON, PK; "Earth" By mungang kim, KR; "database": By Vectorstall, PK; "switch" by Bonegolem, IT . 55

6.2 Bar chart with the number of publications that have used IMBBC HPC facility resources, grouped by scientific field. The different methods for data acquisition are also presented. WGS, whole-genome sequencing; WTS, whole-transcriptome sequencing. 57

6.3 Red bars denote published research with high resource requirements of the various computational methods employed at the IMBBC HPC facility due to (a) long computational times (> 48 h), (b) high memory requirements (> 128 GB), or (c) high storage requirements (> 200 GB). For instance, no eDNA-based community analyses performed at *Zorba* thus far have required a large amounts of memory. 58

List of Tables

2.1 Summary benchmark of PEMA marker - gene – specific mock community recovery (precision) . . . 17
2.2 Comparison of the basic features of the different pipelines . . . 19
2.3 OTU predictions and execution time for the different pipelines. ~ 56 if the reference database is already built . . . 20
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. . . . 22
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". . . . 28
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. . . . 34

3.1 Source databases that are integrated in PREGO and the number of items retrieved. The Open Access subset of PubMed Central has Creative Commons licence available for commercial and noncommercial use. JGI has its own licence, the same applies for BioProject, MEDLINE and PubMed as well. . . . 38
3.2 The entities of PREGO after the NER and mapping of every source. Counts of distinct entities of Taxa, Environments (ENVO terms), Biological Processes (Gene Ontology Biological process) and Molecular Function (Gene Ontology Molecular Function). . . . 41

List of Abbreviations and Symbols

Abbreviations

NGS	Next Generation Sequencing
HPC	High Performance Computing
MCMC	Markov Chain Monte Carlo
MMCS	Multiphase Monte Carlo Sampling
PREGO	PRocess Environment OrGanism
PEMA	Pipeline for Environmental DNA Metabarcoding Analysis
DARN	Dark mAtteR iNvestigator

Chapter 1

Introduction

1.1 Microbial ecology

1.1.1 Microbial communities: structure & function

Microbes, i.e. Bacteria, Archaea and small Eukaryotes such as protoza, are omnipresent and impact global ecosystem functions [7] through their abundance [8], versatility [9] and interactions [10].

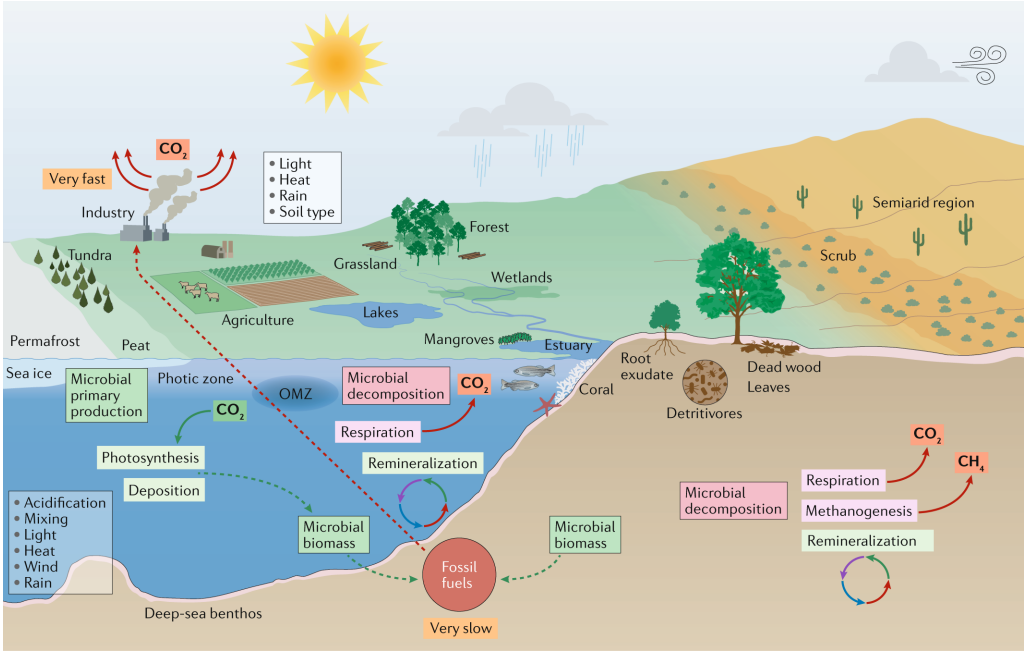

FIGURE 1.1: Marine icrobial communities contribute to CO_2 sequestration, nutrients recycle and thus to the release of CO_2 to the atmosphere. Soil microbial communities decomposers organic matter and release nutrients in the soil from [1] doi: 10.1038/s41579-019-0222-5, under Creative Commons Attribution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

1.1.2 The role of microbial communities in biogeochemical cycles

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

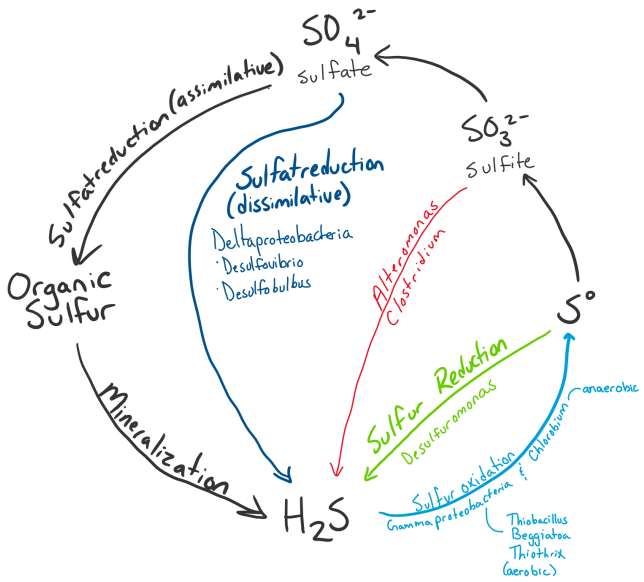

FIGURE 1.2: Sulfur cycle. Figure taken from [2]

1.1.3 Microbial interactions: unravelling the microbiome

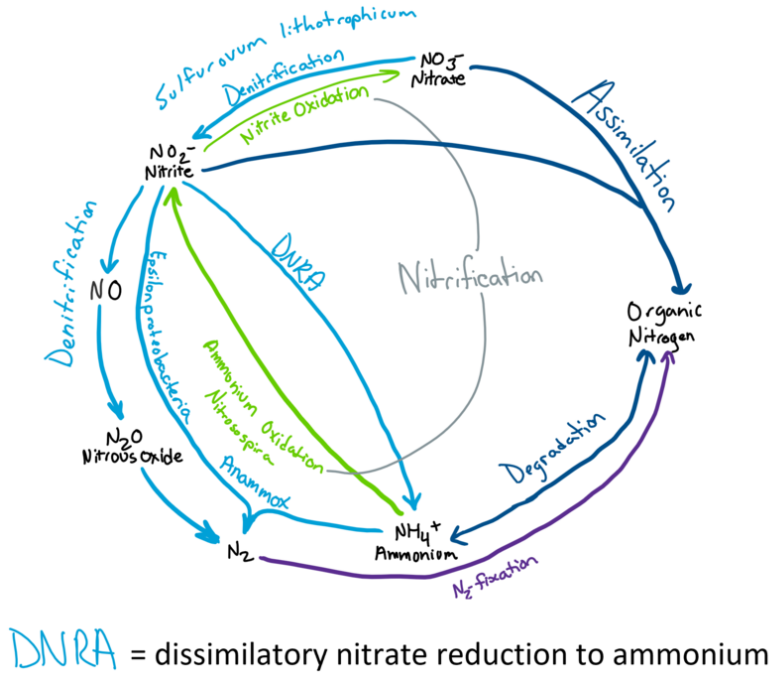

FIGURE 1.3: Nitrogen cycle. Figure taken from [3]

1.2 The era of omics

1.2.1 High Throughput Sequencing approaches

DNA metabarcoding is a rapidly evolving method that is being more frequently employed in a range of fields, such as biodiversity, biomonitoring, molecular ecology and others [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).

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

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

1.3.2 Ontologies & databases: the corner stone of mordern biology

Databases

- GenBank, ENA
- repositories such as MGnify
- PubMed

Ontologies:

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

1.4 Metabolic modeling at the omics era

1.4.1 Genome-scale metabolic model analysis

The relationship between genotype and phenotype is fundamental to biology. Many levels of control are introduced when moving from one to the other. Systems biology aims at deciphering "the strategy" both at the cell and at higher levels of organization, in case of multicell species, that enables organisms to produce orderly adaptive behavior in the face of widely varying genetic and environmental conditions ([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, combinining several new technologies that allow large scale analysis of hundreds of samples. In addition, a software tool called darn, was built to investigate the unassigned sequences in amplicon data of the COI marker gene.

In **Chapter 3**, data integration, data mining and text-mining methods were exploited to build a knowledge-base, called prego, including millions of associations between:

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

2. microbial taxa and biological processes they occur

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

In **Chapter 4**, the challenges of flux sampling in metabolic models of high dimensions was presented along with a Multiphase Monte Carlo Sampling (MMCS) algorithm we developed.

In **Chapter 5**, sediment samples from a hypersaline swamp in Tristomo, Karpathos Greece were analysed using both amplicon and shotgun metagenomics. The taxonomic and the functional profiles of the microbial communities present there were investigated. Key microbial interactions for the assemblages were infered. All the methods developed and presented in the previous chapters were used to enhance the analysis of this microbiome.

In **Chapter 6**, the history of the IMBBC-HCMR HPC facility was presented indicating the vast needs of computing resources in modern analyses in general and in microbial studies more specifically.

Finally, in the **Conclusions** chapter, general discussion and conclusions that have derived from this research were presented.

Chapter 2

Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment

2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes[1]

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

2.1.1 Abstract

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

[1] For author contributions, please refer to the relevant section. Modified version of the published review; extra features have been added and discussed on this thesis.

across operating systems; thus, they strongly facilitate pipeline development and usage. Likewise programming languages specialized for big data pipelines incorporate features like roll-back checkpoints and on-demand partial pipeline execution.

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

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

2.1.2 Introduction

Environmental DNA (eDNA) metabarcoding inaugurates a new era in bio- and eco-monitoring [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 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.

2.1.3 Contribution

PEMA (Pipeline for Environmental DNA Metabarcoding Analysis) is an open source pipeline that bundles state-of-the-art bioinformatic tools for all necessary steps of amplicon analysis and aims to address the aforementioned issues. It is designed for paired-end sequencing studies and is implemented in the BDS [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].

As high-throughput sequencing (HTS) data become more and more accurate, ASVs, i.e., marker gene amplified sequence reads that differ in ≥ 1 nucleotide from each other, become easier to resolve [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 .

2.1.4 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.

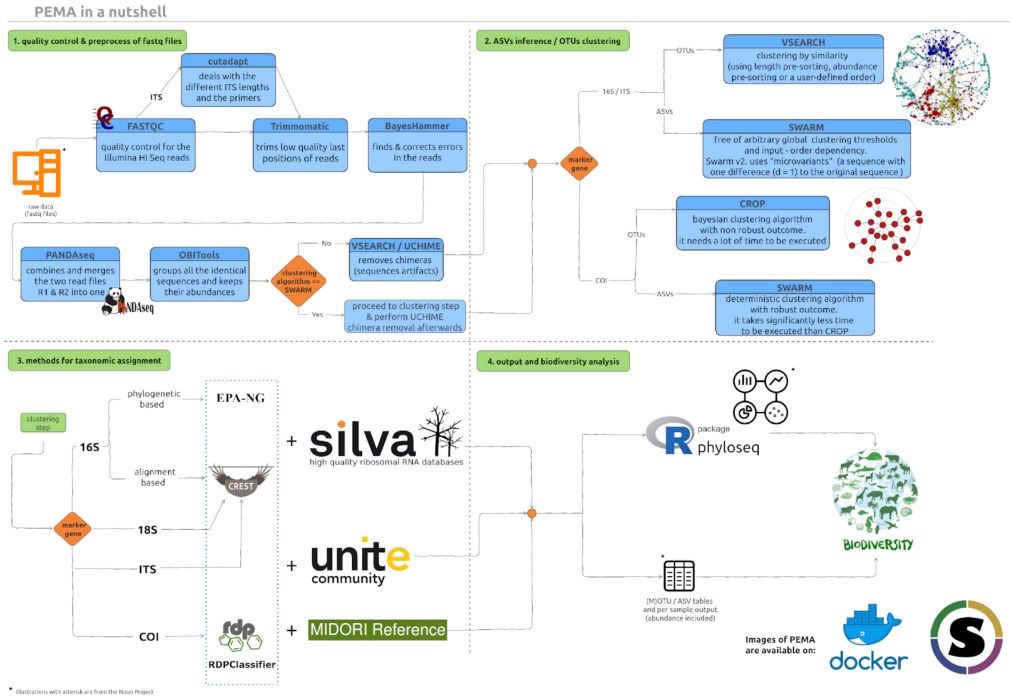

FIGURE 2.1: PEMA comprises 4 parts. The first step (top left) is the quality control and pre-processing of the Illumina sequencing reads. This step is common for both 16S rRNA and COI marker genes. The second step (top right) is the clustering of reads to (M)OTUs or their inferring to ASVs. The third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

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] 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).

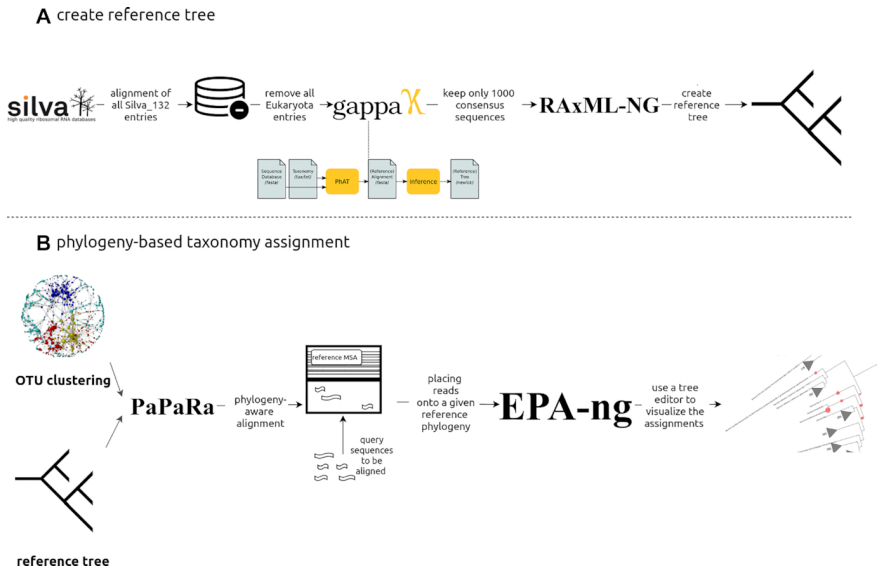

FIGURE 2.2: Phylogeny - based taxonomy assignment. A: Building a reference tree for the phylogeny-based taxonomy assignment to 16S rRNA marker gene OTUs: from the latest edition of Silva SSU, all entries referring to Bacteria and Archaea were used and using the "art" algorithm, 10,000 consensus taxa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

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

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

Via the phyloseq R package [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].

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 SciCruch (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.

PEMA modules added after publication

2

2.1.5 Results & Validation

Evaluation

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

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

[2] REMEMBER TO WRITE THIS PART

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 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.

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

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

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).

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 Chytriomyces in all 3 samples, as well as the Ustilaginaceae family. Only 1 FP assignment was recorded. When the CROP algorithm was used, PEMA's output was less accurate; the *Fusarium* species contained in the mock community were not identified further than their family (Nectriaceae). As mentioned by Bakker [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 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.

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

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

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

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

		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[3]	2.5	~5	~1.5	~26

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

(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 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 Betaproteobacteriales (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 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

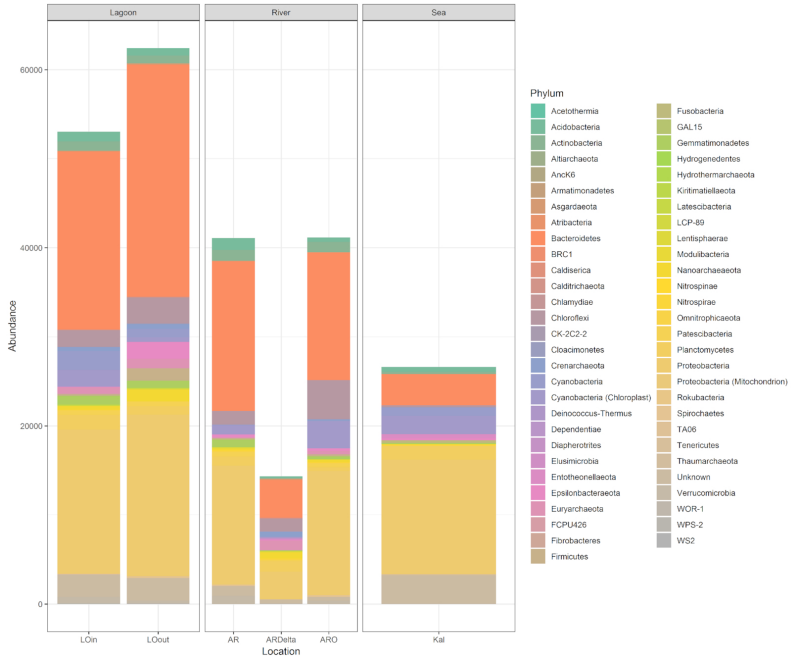

FIGURE 2.3: OTU bar plot at the phylum level. Bar plot depicting the taxonomy of the retrieved OTUs from PEMA for the dataset of Pavloudi et al. [4], at the phylum level for the case of the 16S marker gene. AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; LOin: Logarou station inside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

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

Parameter	***d = 1***	***d = 2***	***d = 3***	***d = 10***	***d = 13***
MOTUs after pre-process and clustering steps	83,791	59,833	33,227	7,384	4,829
MOTUs after chimera removal	80,347	57,863	32,539	7,339	4,796
Non-singleton MOTUs	6,381	4,947	2,658	1,914	1,634
Assigned species	62	83	86	86	84
Execution time (h)	2:01:35	2:09:49	1:51:44	2:17:26	2:31:15

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

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

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

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).

2.1.6 Discussion

OTU clustering vs ASV inference

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

However, the OTU concept is not compulsorily related to the clustering approach; it is widely used to describe results based on its biological meaning but it does not imply clustering. In addition, according to Callahan et al. [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 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, LifeWatch 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 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[4]

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

2.2.1 Abstract

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

2.2.2 Introduction

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

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

reference databases, assessing sample OTU composition. The taxonomy assignment step of the eDNA metabarcoding method and thus, the identification via DNA-barcoding, is only as good and accurate as the reference databases [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 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].

2.2.3 Contribution

The co-amplification of prokaryotes explained above, is a major reason for why many Operational Taxonomic Units (OTUs) and/or Amplicon Sequence Variants (ASVs) in eDNA metabarcoding studies cannot get taxonomy assignments when metazoan reference databases are used (c.f. Aylagas et al. 2016 [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].

2.2.4 Methods & Implementation

Building the COI tree of life

Sequences for the COI region from all the three domains of life were retrieved from curated databases. Eukaryotic sequences were retrieved from the Midori reference 2 database (version: GB239) [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 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

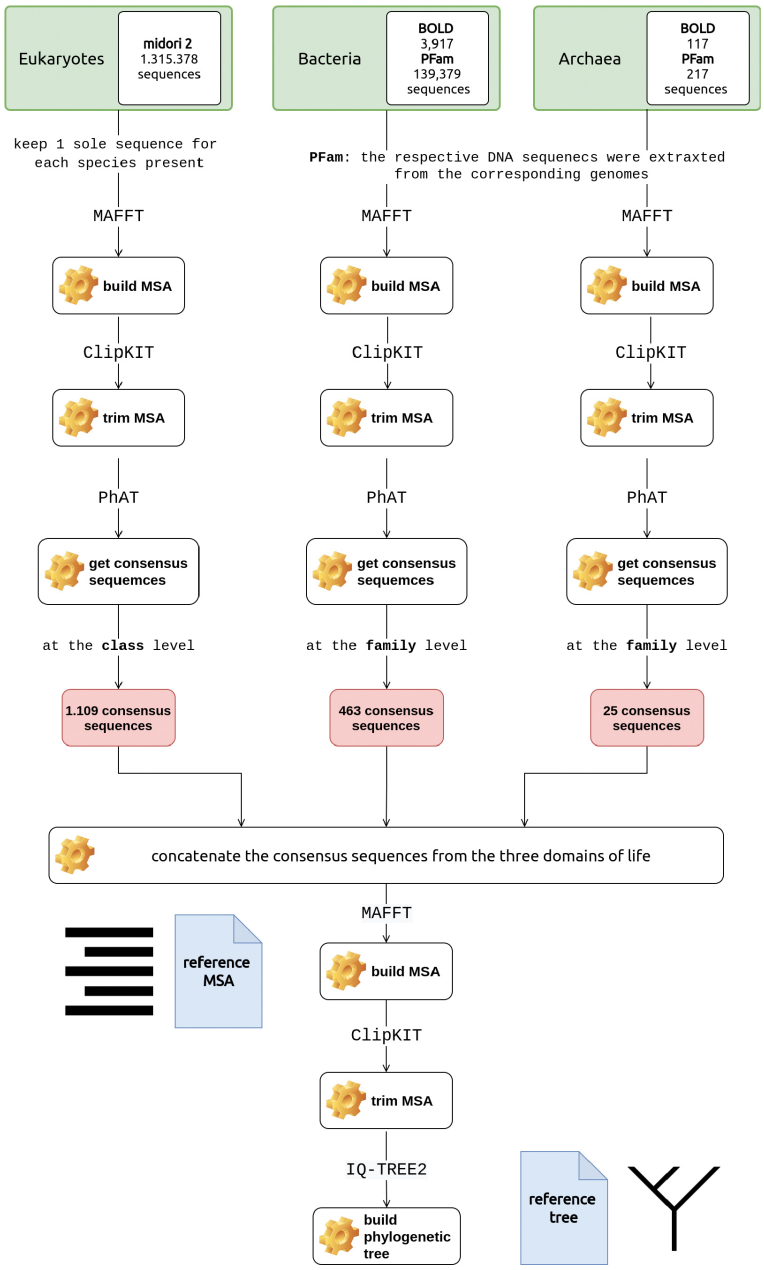

FIGURE 2.4: Overview of the approach followed to build the COI reference tree of life. Sequences were retrieved from Midori 2 (eukaryotes) and BOLD (bacteria and archaea) repositories. Consensus sequences at the family level were built for each domain specific dataset. MAFFT and consensus sequences at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

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 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.

2.2.5 Results & Validation

Evaluation of the phylogenetic tree

The inferred phylogenetic tree is shown in Figure 2, with the bacterial (light blue) and archaeal (dark green) branches highlighted; in Suppl. material 3: Fig. S1 the distribution of the eukaryotic phyla on the tree is presented. As shown, bacteria and archaea can be distinguished from eukaryotes. Scattered bacterial branches that are present among eukaryotic ones represent the diversity of the COI locus. To evaluate the phylogenetic tree, the set of consensus sequences were placed on it using the EPA-ng algorithm. The placements (see .jplace through a phylogenetic tree viewer, e.g. iTOL) verified that the phylogenetic tree built is valid, as the consensus sequences have been placed in their corresponding taxonomic branches (Suppl. material 4: Fig. S2; the figure was built using the heat-tree module of the gappa examine tool).

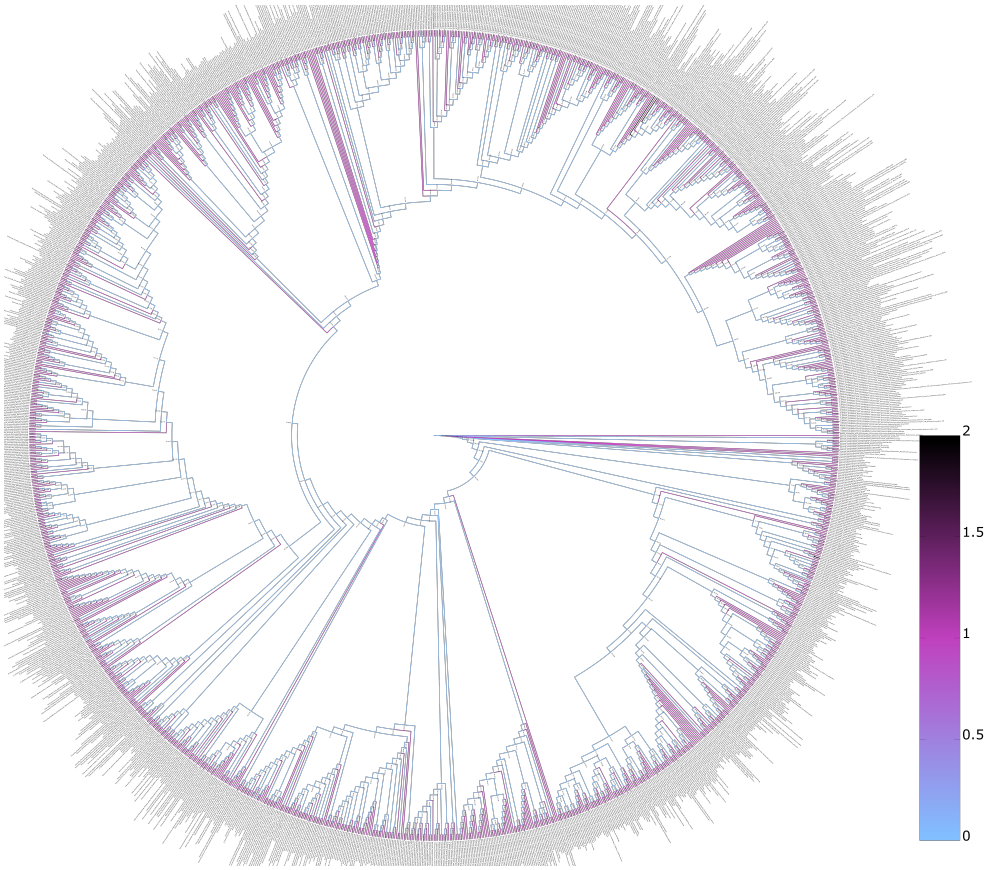

FIGURE 2.5: Phylogenetic tree of the consensus sequences retrieved; the tree that DARN makes use of. Light blue: bacterial branches. Dark green: archaeal branches. White: eukaryotic branches.

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 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.

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

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

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

2.2.6 Discussion

By making use of a COI - oriented reference phylogenetic tree built from 1,593 consensus sequences, to phylogenetically place sequences from COI metabarcoding samples onto it, the surmise for including bacteria, algae, fungi etc. [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 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[1]

3.1.1 Abstract

3.1.2 Introduction

3.1.3 Contribution

3.1.4 Methods & Implementation

3.1.5 Results & Validation

kljfdfg 3.1.5

PREGO contents

3.1.6 Discussion

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

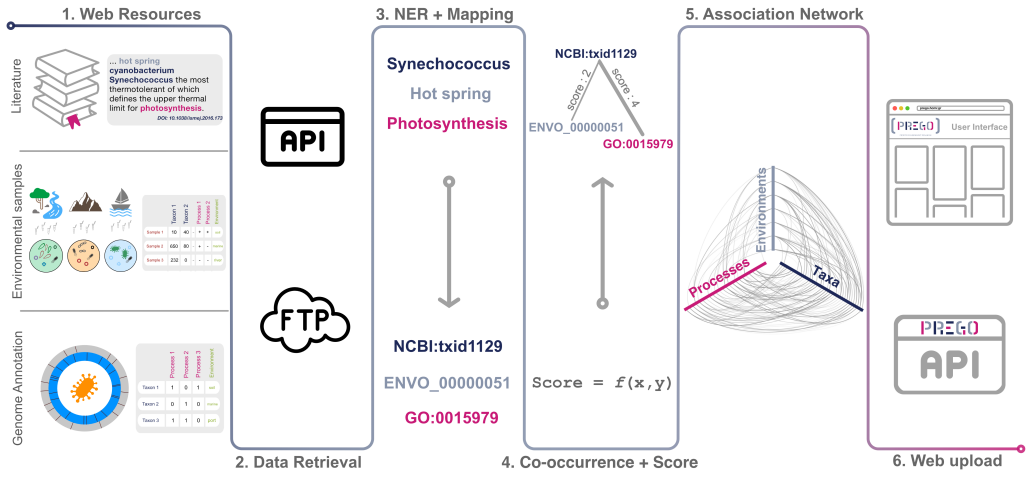

FIGURE 3.1: PREGO analysis methodology. PREGO periodically retrieves three distinct types of data from open access resources. Scientific text, environmental sample data and genomic annotations are handled with respective methodologies in order to standardize their entities. Named Entity Recognition and Co-occurrence analysis is the common framework in order to build a weighted association network with nodes being the entity identifiers. Lastly, all these associations are available through a Web interface and an API. All these steps have been implemented in an autonomous way with regular cycles of updates (see Appendix B).

Source	# items	Data type	Metadata	License
MEDLINE and PubMed	33 million	abstracts (text)	no	NLM Copyright
PubMed Central OA Subset	2.7 million	full article (text)	no	CC for Commercial, non-commercial
JGI IMG	9,644	Isolates Annotated genomes	yes	JGI Data Policy
Struo	21,276	Annotated genomes	no	MIT, CC BY-SA 4.0
BioProject	18,752	Annotated genomes with abstracts (text)	yes	INSDC policy
MG-RAST	16,096	markergene samples	yes	CC0
	7,965	metagenomic samples	yes	CC0
MGnify	10,500	markergene samples	yes	CC-BY, CC0

TABLE 3.1: Source databases that are integrated in PREGO and the number of items retrieved. The Open Access subset of PubMed Central has Creative Commons licence available for commercial and noncommercial use. JGI has its own licence, the same applies for BioProject, MEDLINE and PubMed as well.

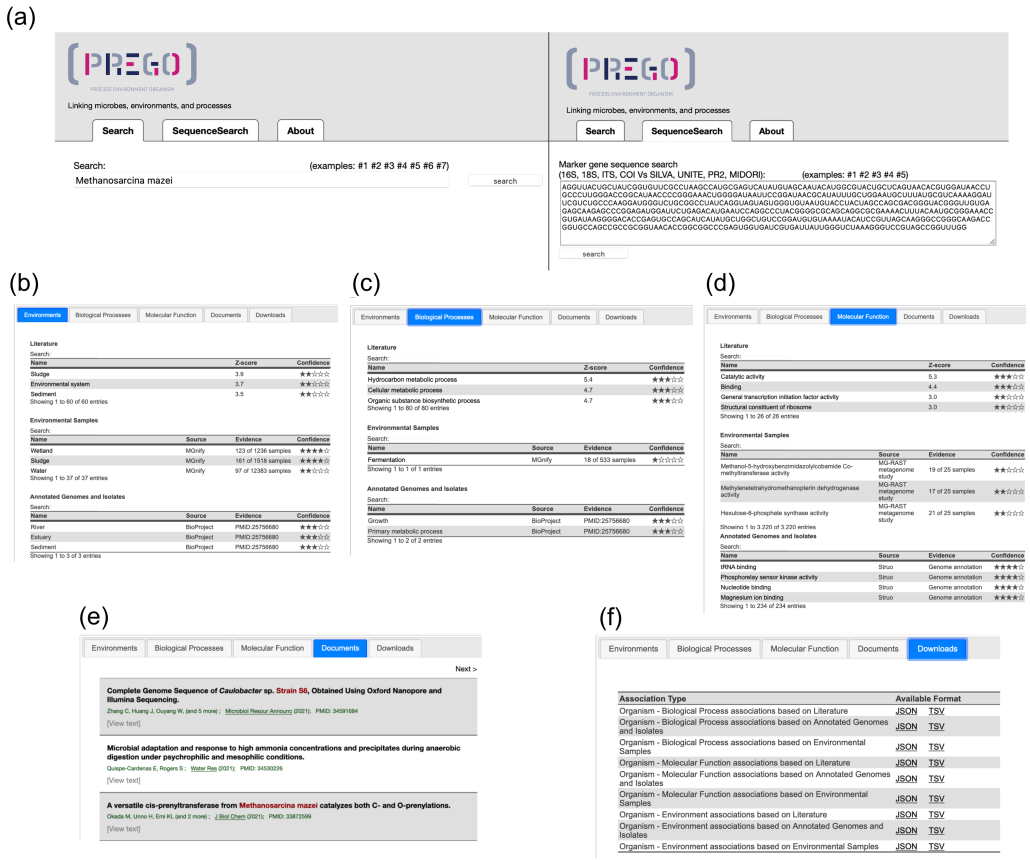

FIGURE 3.2: PREGO web user interface. (a) There are two search fields, plain text and taxa sequences. (b-d) three associations tabs each one presenting associations of the querred entity with the respective entities, Environments (b), Biological Process (c) and Molecular Function (d). Three channels of information are distinguishing the associations based on the original data. (e) Documents tab presents the scientific articles that mention the queried entity highlighted with color. (f) Downloads tab provides the associations of each channel (when available) to be downloaded in JSON and TSV format.

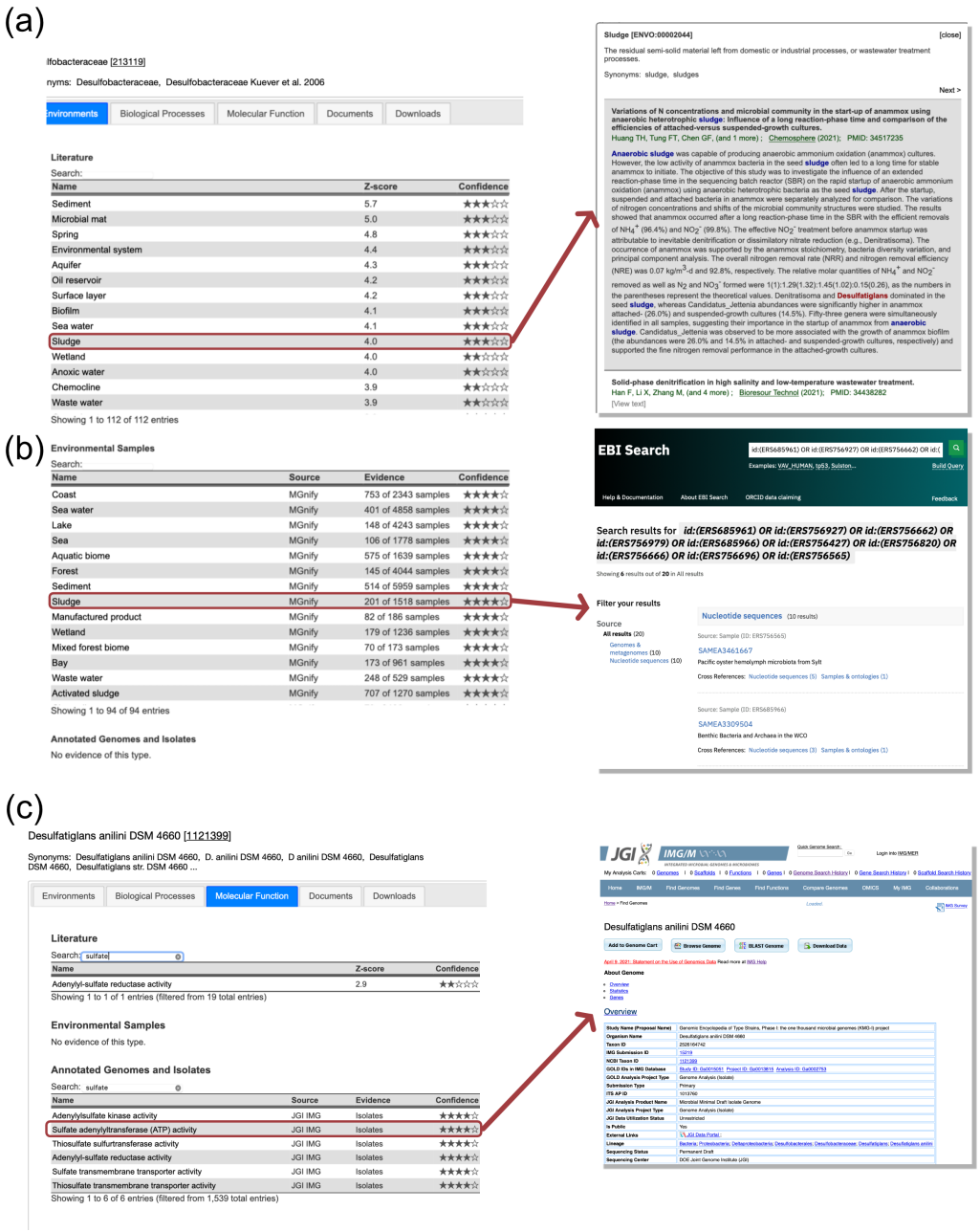

FIGURE 3.3: Each association is supported by original data. (a) Literature channel has a pop-up functionality that displays the scientific articles that each specific association occurs with highlighted color. (b) Environmental Samples channel redirects to the samples that support the specific association (currently only is supported MGnify). (c) Annotated Genomes channel similarly redirects to the isolates ids that each association is based on (both Struo and JGI IMG are supported).

Channel	Source	Taxonomy		Environments	Biological Processes	Molecular Functions
Literature	MEDLINE PubMed - PMC OA	Strains	8,929	1,077	15,079	7,318
		Species	240,377			
		Total	342,506			
Environmental samples	MG-RAST amplicon	Strains	1,392	162	-	-
		Species	4,324			
		Total	5,859			
	MG-RAST metagenome	Strains	2,522	258	-	3,839
		Species	4,406			
		Total	7,157			
	MGnify amplicon	Strains	2	216	11	-
		Species	1,471			
		Total	2,955			
Annotated Genomes & Isolates	JGI IMGisolates	Strains	2,398	241	-	3,670
		Species	11,203			
		Total	13,849			
	STRUO	Strains	6	-	-	2,789
		Species	19,289			
		Total	19,325			
	BioProject	Strains	5,754	309	626	-
		Species	3,373			
		Total	9,393			
Total	All	Strains	12,840	1,090	15,091	7,971

TABLE 3.2: The entities of PREGO after the NER and mapping of every source. Counts of distinct entities of Taxa, Environments (ENVO terms), Biological Processes (Gene Ontology Biological process) and Molecular Function (Gene Ontology Molecular Function).

Channel	Source	Environments - Processes	Environments - Functions	Taxonomy	Taxa - Environments	Taxa - Processes	Taxa - Function
Literature	MEDLINE PubMed - PMC OA	883,997	422,579	Strains	69,968	590,630	384,079
				Species	778,877	3,501,635	1,961,920
				Total	1,669,608	7,969,310	4,613,827
Environmental samples	MG-RAST amplicon	-	-	Strains	13,645		
				Species	39,007	-	-
				Total	53,439		
	MG-RAST metagenome	-	620,846	Strains	262,106		8,626,328
				Species	103,913	-	10,715,548
				Total	372,301		19,950,096
	MGnify amplicon	-	-	Strains	18	-	
				Species	30,122	351	-
				Total	111,976	2,097	
	JGI IMG isolates	-	-	Strains	8,229		3,461,693
				Species	42,141	-	13,216,559
				Total	50,888		16,821,850
Annotated Genomes and Isolates	STRUO	-	-	Strains			1,803
				Species	-	-	4,070,195
				Total			4,079,312
	BioProject	-	-	Strains	3,263	7,473	
				Species	4,187	4,294	
				Total	7,641	12,169	
Total	All	883,997	1,043,425	Strains	357,229	598,103	12,473,903
				Species	998,247	3,506,280	29,964,222
				Total	2,265,853	7,983,576	45,465,085

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 5 335, took less than 30 hours. To the best of our knowledge, that is out of reach for existing software.

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

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

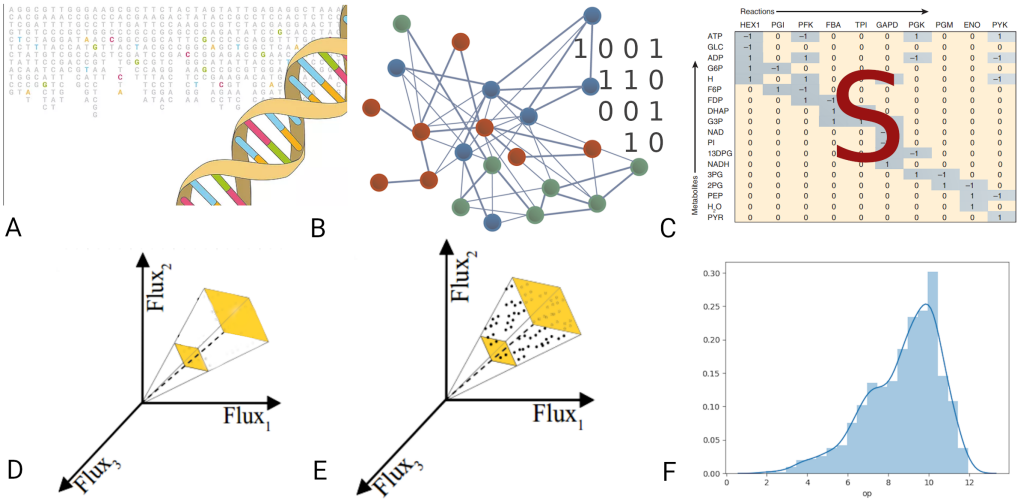

FIGURE 4.1: From DNA sequences to distributions of metabolic fluxes. (A) The genes of an organism provide us with the enzymes that it can potentially produce. Enzymes are like a blueprint for the reactions they can catalyze. (B) Using the enzymes we identify the reactions in the organism. (C) We construct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

4.1.2 Contribution

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

This improves the efficiency of the random walk in the next phase, see For sampling, we propose an improved variant of Billiard Walk that enjoys faster arithmetic complexity per step. We also handle efficiently the potential arithmetic inaccuracies near to the boundary, see [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[1] provides a complete software framework to sample efficiently in metabolic networks. We demonstrate the efficiency of our tools by performing experiments on almost all the metabolic networks that are publicly available and by comparing with the state-of-the-art software packages, like cobra

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

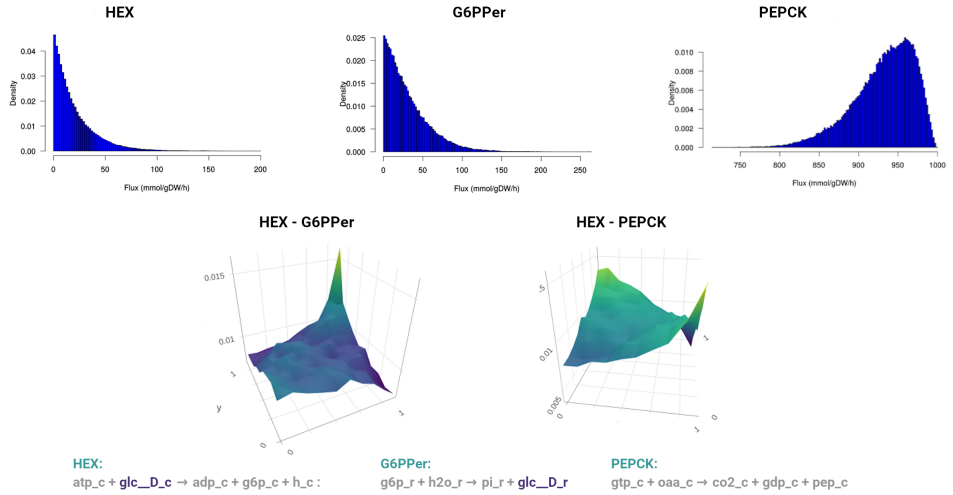

FIGURE 4.2: Flux distributions in the most recent human metabolic network Recon3D [5]. We estimate the flux distributions of the reactions catalyzed by the enzymes Hexokinase (D-Glucose:ATP) (HEX), Glucose-6-Phosphate Phosphatase, Edoplasmic Reticular (G6PPer) and Phosphoenolpyruvate carboxykinase (GTP)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

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

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, ρ, τ, W)

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

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

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

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

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

Multiphase Monte Carlo Sampling algorithm

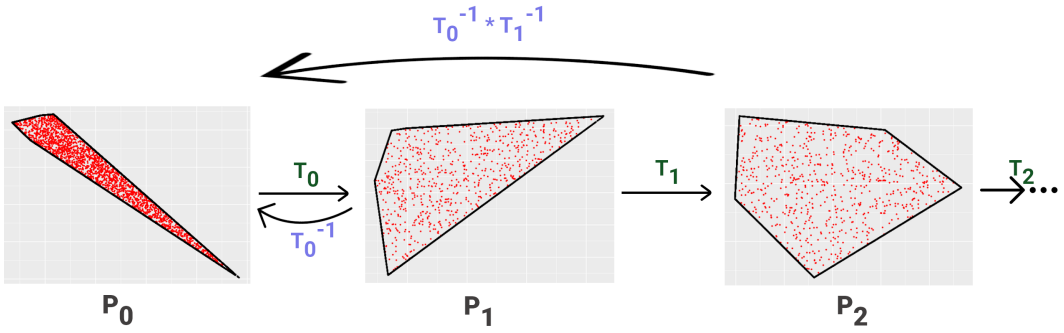

FIGURE 4.3: An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer n and starts at phase $i = 0$ sampling from P_0. In each phase it samples a maximum number of points λ. If the sum of Effective Sample Size in each phase becomes larger than n before the total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

4.1.4 Results

4.1.5 Discussion

Flux sampling at the community level!

Chapter 5

Studying the microbiome as a whole: the way forward

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

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

5.1.1 Introduction

5.1.2 Contribution

5.1.3 Methods

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

5.1.4 Results

5.1.5 Discussion

Chapter 6

An overview of the computational requirements & solutions in microbial ecology

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

Citation: Zafeiropoulos, H., Gioti, A., Ninidakis, S., Potirakis, A., Paragkamian, S., ... & Pafilis, E. (2021). 0s and 1s in marine molecular research: a regional HPC perspective. GigaScience, 10(8), giab053, doi: 10.1093/gigascience/giab053

6.1.1 Abstract

High-performance computing (HPC) systems have become indispensable for modern marine research, providing support to an increasing number and diversity of users. Pairing with the impetus offered by high-throughput methods to key areas such as non-model organism studies, their operation continuously evolves to meet the corresponding computational challenges.

Here, we present a Tier 2 (regional) HPC facility, operating for over a decade at the Institute of Marine Biology, Biotechnology, and Aquaculture of the Hellenic Centre for Marine Research in Greece. Strategic choices made in design and upgrades aimed to strike a balance between depth (the need for a few high-memory nodes) and breadth (a number of slimmer nodes), as dictated by the idiosyncrasy of the supported research. Qualitative computational requirement analysis of the latter revealed the diversity of marine fields, methods, and approaches adopted to translate data into knowledge. In addition, hardware and software architectures, usage statistics, policy, and user management aspects of the facility are presented. Drawing upon the last decade's experience from the different levels of operation of the Institute of Marine Biology, Biotechnology, and Aquaculture HPC facility, a number of lessons are presented; these have contributed to the facility's future

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

directions in light of emerging distribution technologies (e.g., containers) and Research Infrastructure evolution. In combination with detailed knowledge of the facility usage and its upcoming upgrade, future collaborations in marine research and beyond are envisioned.

6.1.2 Introduction

The ubiquitous marine environments (more than 70% of the global surface [114]) mold Earth's conditions to a great extent. The interconnected abiotic [7] and biotic factors (from bacteria [7] to megafauna [115]), shape biogeochemical cycles [116] and climates [117, 118] from local to global scales. In addition, marine systems have high socio-economic value [119] as an essential source of food and by supporting renewable energy and transport, among other services [120]. The study of marine environments involves a series of disciplines (scientific fields): from Biodiversity [121] and Oceanography to (eco)systems biology [122] and from Biotechnology [123] to Aquaculture [124].

To shed light on the evolutionary history of (commercially important) marine species [125], as well as on how invasive species respond and adapt to novel environments [126], the analysis of their genetic stock structure is fundamental [127]. Similarly, biodiversity assessment is essential to elucidate ecosystem functioning [128] and to identify taxa with potential for bioprospecting applications [129]. Furthermore, systems biology approaches provide both theoretical and technical backgrounds in which integrative analyses flourish [130]. However, conventional methods do not offer the information needed to explore the aforementioned scientific topics.

High-throughput sequencing (HTS) and sister methods have launched a new era in many biological disciplines [131, 132]. These technologies allowed access to the genetic, transcript, protein, and metabolite repertoire [133] of studied taxa or populations, and facilitated the analysis of organism-environment interactions in communities and ecosystems [134]. Whole-genome sequencing and whole-transcriptome sequencing approaches provide valuable information for the study of non-model taxa [135]. This information can be further enriched by genotyping-by-sequencing approaches, such as restriction site-associated DNA sequencing [136], or by investigating gene expression dynamics through Differential Expression (DE) analyses [137]. Moving from single species to assemblages, molecular-based identification and functional profiling of communities has become available through marker (metabarcoding), genome (metagenomics), or transcriptome (metatranscriptomics) sequencing from environmental samples [138]. To a great extent, these methods address the problem of how to produce and get access to the information on different biological systems and molecules.

These 0s and 1s of information (i.e., the data) come along with challenges regarding their management, analysis, and integration [139]. The computational requirements for these tasks exceed the capacity of a standard laptop/desktop by far, owing to the sheer volume of the data and to the computational complexity of the bioinformatic algorithms employed for their analysis. For example, building the de novo genome assembly of a non-model Eukaryote may require algorithms of nondeterministic polynomial time complexity. This analysis can reach up to several hundreds or thousands of GB of memory

(RAM) [140]. Hence, the challenges of how to exploit all these data and how to transform data into knowledge set the present framework in biological research [141, 142].

To address these computational challenges, the use of high-performance computing (HPC) systems has become essential in life sciences and systems biology [143]. HPC is the scientific field that aims at the optimal incorporation of technology, methodology, and the application thereof to achieve “the greatest computing capability possible at any point in time and technology" [144]. Such systems range from a small number to several thousands of interconnected computers (compute nodes). According to the Partnership for Advanced Computing in Europe, the European HPC facilities are categorized as: (i) European Centres (Tier 0), (ii) national centers (Tier 1), and (iii) regional centers (Tier 2) [33] [145]. As the Partnership for Advanced Computing in Europe highlights, "computing drives science and science drives computing" in a great range of scientific fields, from the endeavor to maintain a sustainable Earth to efforts to expand the frontiers in our understanding of the universe [146]. On top of the heavy computational requirements, biological analyses come with a series of other practical issues that often affect the bioinformatics-oriented HPC systems.

Researchers with purely biological backgrounds often lack the coding skills or even the familiarity required for working with Command Line Interfaces [146]. Virtual Research Environments are web-based e-service platforms that are particularly useful for researchers lacking expertise and/or computing resources [147]. Another common issue is that most analyses include a great number of steps, with the software used in each of these having equally numerous dependencies. A lack of continuous support for tools with different dependencies, as well as frequent and non-periodical versioning of the latter, often results in broken links and further compromises the reproducibility of analyses [36]. Widely used containerization technologies—e.g., Docker [28] and Singularity [29]—ensure reproducibility of software and replication of the analysis, thus partially addressing these challenges. By encapsulating software code along with all its corresponding dependencies in such containers, software packages become reproducible in any operating system in an easy-to-download-and-install fashion, on any infrastructure.

6.1.3 Contribution

The Institute of Marine Biology Biotechnology and Aqua-culture (IMBBC) has been developing a computing hub that, in conjunction with national and European Research Infrastructures (RIs), can support state-of-the-art marine research. The regional IMBBC HPC facility allows processing of data that derive from the Institute’s sequencing platforms and expeditions and from multiple external sources in the context of interdisciplinary studies. Here, we present insights from a thorough analysis of the research supported by the facility and some of its latest usage statistics in terms of resource requirements, computational methods, and data types; the above have contributed in shaping the facility along its lifespan.

6.1.4 Methods

The IMBBC HPC Facility: From a Single Server to a Tier 2 System

The IMBBC HPC facility was launched in 2009 to support computational needs over a range of scientific fields in marine biology, with a focus on non-model taxa [39]. The facility was initiated as an infrastructure of the Institute of Marine Biology and Genetics of the Hellenic Centre for Marine Research. Its development has followed the development of national RIs (Fig. 1; also see Section A1 in Zafeiropoulos et al. [148]). The first nodes were used to support the analysis of data sets generated from methods such as eDNA metabarcoding and multiple omics. Since 2015, the facility also supports Virtual Research Environments, including e-services and virtual laboratories. The current configuration of the facility presented herein is named *Zorba* (Fig. 1, Box 4) and will be upgraded within 2021 (see Section Future Directions). Hereafter, *Zorba* refers to the specific system setup from 2015 and onwards, while the facility throughout its lifespan will be referred to as "IMBBC HPC".

Zorba currently consists of 328 CPU cores, 2.3 TB total memory, and 105 TB storage. Job submission takes place on the 4 available computing partitions, or queues, as explained in Fig. 2. *Zorba* at its current state achieves a peak performance of 8.3 trillion double-precision floating-point operations per second, or 8.3 Tflops, as estimated by LinPack benchmarking [149]. On top of these, a total 7.5 TB is distributed to all servers for the storage of environment and system files. Interconnection of both the compute and login nodes takes place via an infiniband interface with a capacity of 40 Gbps, which features very high throughput and very low latency. Infiniband is also used for a switched interconnection between the servers and the 4 available file systems. A thorough technical description of *Zorba* is available in Section A2 of Zafeiropoulos et al. [148].

More than 200 software packages are currently installed and available to users at *Zorba*, covering the most common analysis types. These tools allow assembly, HTS data preprocessing, phylogenetic tree construction, ortholog finding, and population structure modeling, to name a few. Access to these packages is provided through Environment Modules, a broadly used means of accessing software in HPC systems [150].

During the last 2 years, *Zorba* has been moving from system-dependent pipelines previously developed at IMBBC (e.g., ParaMetabarCoding) towards containerization of available and new pipelines/tools. A complete metabarcoding analysis tool for various marker genes (PEMA) [99], the chained and automated use of STACKS, software for population genetics analysis from short-length sequences [151] (latest version), a set of statistical functions in R for the computation of biodiversity indices, and analyses in cases of high computational demands [152], as well as a programming workflow for the automation of biodiversity historical data curation (DECO) are among the in-house developed containers. The standard container/image format used on *Zorba* is Singularity. Singularity images can be served by any *Zorba* partition; Docker images can run instantly as Singularity images. A thorough description of the software containers developed in *Zorba* can be found in Section D of Zafeiropoulos et al. [148].

Zorba's daily functioning is ensured by a core team of 4 full-time, experienced staff:

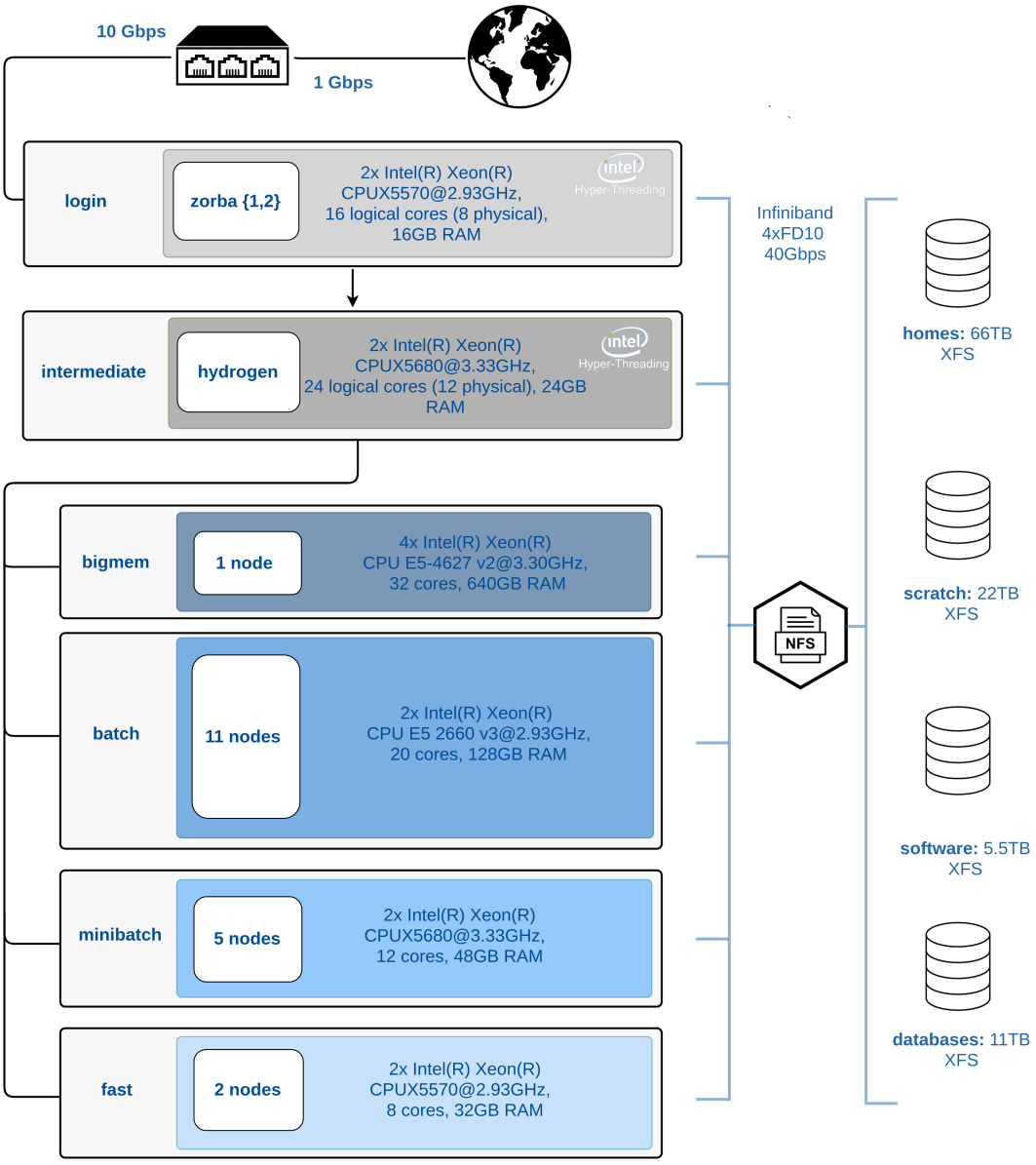

FIGURE 6.1: Block diagram of the *Zorba* architecture. This is the IMBBC HPC facility architecture in its current setup, after 12 years of development. There are 2 login nodes and 1 intermediate where users may develop their analyses. Computational nodes are split into 4 partitions with different specs and policy terms: bigmem supporting processes requiring up to 640 GB RAM, batch handling mostly (but not exclusively) parallel-driven jobs (either in a single node or across several nodes), minibatch aiming to serve parallel jobs with reduced resource requirements, and fast partition for non-intensive jobs. All servers, except file systems, run Debian 9 (kernel 4.9.0-8-amd64). CCBY icons from the Noun Project: "nfs file document icon" by IYIKON, PK; "Earth" By mungang kim, KR; "database": By Vectorstall, PK; "switch" by Bonegolem, IT

a hardware officer, 2 system administrators, and a permanent researcher in biodiversity informatics and data science.

More than 70 users (internal and external scientists), investigators, postdoctoral researchers, technicians, and doctoral/postgraduate students have gained access to the HPC infrastructure thus far. Support is provided officially through a help desk ticketing system. An average of 31 requests/month have been received (since June 2019), with the most demanded categories being troubleshooting (38.2%) and software installation (23.8%). Since October 2017, monthly meetings among HPC users have been established to regularly discuss such issues.

Proper scheduling of the submitted jobs and fair resource sharing is a major task that needs to be confronted day to day. To address this, a specific usage policy for each of the various partitions and a scheduling software tool set have been adopted in *Zorba*. Policy terms are dynamically adapted to the HPC hardware architecture and to the usage statistics, with revisions being discussed between the HPC core team and users. The Simple Linux Utility for Resource Management (SLURM) open-source cluster management system orchestrates the job scheduling and allocates resources, and a booking system helps users to organize their projects and administrators to monitor the resource reservations on a mid- to long-term basis. A SLURM Database Daemon (slurmdbd) has also been installed to allow logging and recording of job usage statistics into a separate SQL database (see Section C1 in Zafeiropoulos et al. [148]). An extended description of user and job administrations and orchestration can be found in Section C1 of Zafeiropoulos et al. [148]).

Training has been an integral component of the HPC facility mindset since its launch and enables knowledge sharing across MSc and PhD students and researchers within and outside the Institute. Introductory courses are organized on a regular basis, aimed at familiarizing new users with Unix environments, programming, and HPC usage policy and resource allocation (e.g., job submission in SLURM). Furthermore, the IMBBC HPC facility has served, since 2011, as an international training platform for specific types of bioinformatic analyses (see Section C2 in Zafeiropoulos et al. [148]). For instance, the facility has provided computational resources for workshops on Microbial Diversity, Genomics and Metagenomics, Genomics in Biodiversity, Next-Generation Sequencing technologies and informatics tools for studying marine biodiversity and adaptation in the long term, or Ecological Data Analysis using R. The plan is to enhance and diversify the educational component of the HPC facility by providing courses on a more permanent basis and targeting a larger audience. An extensive listing of training activities is given in Section C2 of Zafeiropoulos et al. [148].

6.1.5 Results

Computational Breakdown of the IMBBC HPC-Supported Research

Systematic labelling of IMBBC HPC-supported published studies ($n = 47$) was performed to highlight their resource requirements. Each study was manually labelled with the relevant scientific field, the data acquisition method, the computational methods, and its resource requirements; all the annotations were validated by the corresponding authors

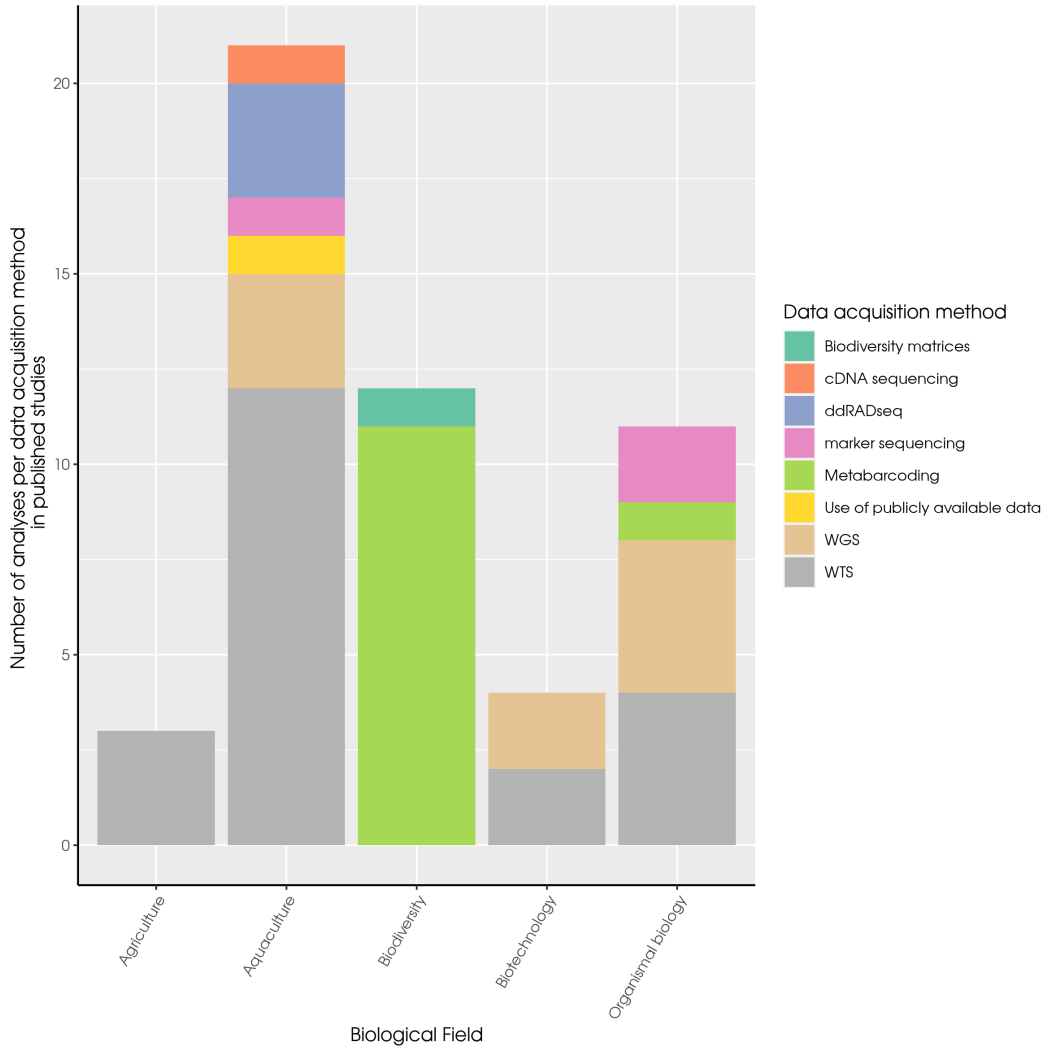

FIGURE 6.2: Bar chart with the number of publications that have used IMBBC HPC facility resources, grouped by scientific field. The different methods for data acquisition are also presented. WGS, whole-genome sequencing; WTS, whole-transcriptome sequencing.

(see Section D2 in Zafeiropoulos et al. [148]). It should be stated that the conclusions of this overview are specific to the studies conducted at IMBBC.

The scientific fields of Aquaculture (40% of studies), Biodiversity (26% of studies), and Organismal biology (19% of studies) account for the majority of the research publications supported by the IMBBC HPC facility (Fig. 3; Supplementary File imbbc_hpc_labelling_data.xlsx in Zafeiropoulos et al. [148]).

In comparison, studies in the Biotechnology and Agriculture fields indicate contemporary and beyond-marine orientations of research at IMBBC, respectively (see Section B2 in Zafeiropoulos et al. [40]). In addition, 8 methods of data acquisition (experimental or *in silico*) have been defined (Fig. 3). Among these methods, whole-genome sequencing and

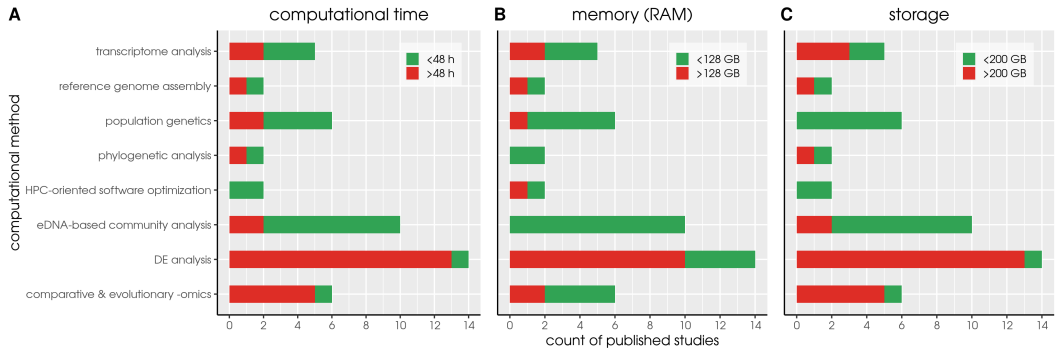

FIGURE 6.3: Red bars denote published research with high resource requirements of the various computational methods employed at the IMBBC HPC facility due to (a) long computational times (> 48 h), (b) high memory requirements (> 128 GB), or (c) high storage requirements (> 200 GB). For instance, no eDNA-based community analyses performed at *Zorba* thus far have required a large amounts of memory.

whole-transcriptome sequencing have been widely used in multiple fields (Biotechnology, Organismal Biology, Aquaculture). Conversely, Double digest restriction-site associated sequencing (ddRADseq) has been solely employed for population genetic studies in the context of Aquaculture.

The 47 published studies employed different computational methods (sets of tasks executed on the HPC facility). These studies served different purposes, from a range of bioinformatics analyses to HPC-oriented software optimization. The computational methods were categorized in 8 classes (Fig. 4). The resource requirements of each computational method were evaluated in terms of memory usage, computational time, and storage. Reflecting the current *Zorba* capacity, studies which, in any part of their analysis, exceeded 128 GB of memory or/and 48 hours of running time or/and 200 GB physical space were classified as studies with high demands (see Supplementary file imbbc_hpc_labelling_data.xlsx in [148]).

As shown in Fig. 4, the 2 most commonly used computational methods have rather different resource requirements. While DE analysis shows a notable trend for both long computational times (Fig. 4a) and high memory (Fig. 4b), eDNA-based community analysis does not have high resource requirements either in computation time or memory. High memory was commonly associated with computational methods, including de novo assembly; all relevant research concerned non-model taxa and involved short-read sequencing or combinations of short- and long-read sequencing. By contrast, phylogenetic analysis studies did not involve intensive RAM use; this is largely due to the fact that software used by IMBBC users adopts parallel solutions for tree construction. Long computational times (Fig. 4a) were most often observed at the functional annotation step in transcriptome analysis, DE analysis, and comparative and evolutionary omics, when this step involved BLAST queries of thousands of predicted genes against large databases, such as nr (NCBI). Finally, a common challenge emerging from all bioinformatic approaches is significant storage limitations (Fig. 4c); this challenge was associated with the use of HTS

technologies that produce large amounts of raw data, the analysis of which involves the creation of numerous intermediate files.

Overall, published studies using the IMBBC HPC facility show a degree of variance with respect to the types of tools used (depending on the user, their bioinformatic literacy, and other factors), each of which is more or less optimized with respect to HPC use. Moreover, the variance in computational needs observed within each type of computational method reflects the diversity of the studied taxonomic groups. For instance, transcriptome analysis (involving de novo assembly and functional annotation steps) was employed for the study of taxa as diverse as bacteria, sponges, fungi, fish, and goose barnacles. The complexity of each of these organisms' transcriptomes can, to a large extent, explain the differences observed in computational time, memory, and storage.

Furthermore, *Zorba* CPU and RAM statistics collected since 2019 displayed some overall patterns, including an average computation load per month of less than or close to 50% of its max capacity (50% of 236 kilocorehours/month) for most (20) of the 24 months of the logging period. Memory requirements were also heterogeneous: most (90%) of the 44,000 jobs performed in the same 24-month period required less than 10 GB of RAM, and 0.30% of the jobs required more than 128 GB of RAM (i.e., exceeding the memory capacity of the main compute nodes [batch partition]). The detailed usage statistics of *Zorba* are described in Section B1 and Supplementary file zorba_usage_statistics.xlsx of Zafeiropoulos et al. [148].

6.1.6 Discussion

Scientific Impact Stories

Below, some examples of research results that were made possible with the IMBBC HPC facility are described. This list of use cases is by no means exhaustive, but rather an attempt to highlight different fields of research supported by the facility, along with their distinct computational features.

Invasive species range expansion detected with eDNA data from Autonomous Reef Monitoring Structures

The Mediterranean biodiversity and ecosystems are experiencing profound transformations owing to Lessepsian migration, international shipping, and aquaculture, which lead to the migration of nearly 1,000 alien species [46]. The first step towards addressing the effects of these invasions is monitoring of the introduced taxa. A powerful tool in this direction has been eDNA metabarcoding, which has enchanced detection of invasive species [153], often preceding macroscopic detection. One such example is the first record of the nudibranch *Anteaeolidiella lurana* (Ev. Marcus & Er. Marcus, 1967) in Greek waters in 2020 [154]. An eDNA metabarcoding analysis allowed for detection of the species with high confidence on fouling communities developed on Autonomous Reef Monitoring Structures (ARMS). This finding, confirmed with image analysis of photographic records on a later deployment period, is an example of work conducted within the framework of the European ASSEMBLE plus programme ARMS-MBON (Marine Biodiversity Observa-

tion Network). PEMA software [99] was used in this study, as well as in the 30-month pilot phase of ARMS-MBON [155].

Providing omics resources for large genome-size, non-model taxa

Zorba has been used for building and annotating numerous de novo genome and transcriptome assemblies of marine species, such as the gilthead sea bream *Sparus aurata* [156] or the greater amberjack *Seriola dumerili* [157]. Both genome and transcriptome assemblies of species with large genomes often exceed the maximum available memory limit, eventually affecting the strategic choices for *Zorba* future upgrades (see Section Future Directions). For instance, building the draft genome assembly of the seagrass Halophila stipulacea (estimated genome size 3.5 GB) using Illumina short reads has been challenging even for seemingly simple tasks, such as a kmer analysis [158]. Taking advantage of short- and long-read sequencing technologies to construct high-quality reference genomes, the near-chromosome level genome assembly of Lagocephalus sceleratus (Gmelin, 1789) was recently completed as a case study of high ecological interest due to the species' successful invasion throughout the Eastern Mediterranean [159]. In the context of this study, an automated containerized pipeline allowing high-quality genome assemblies from Oxford Nanopore and Illumina data was developed (SnakeCube [160]). The availability of standardized pipelines offers great perspective for in-depth studies of numerous marine species of interest in aquaculture and conservation biology, including rigorous phylogenomic analyses to position each species in the tree of life (e.g., Natsidis et al. [161]).

DE analysis of aquaculture fish species sheds light on critical phenotypes

Distinct, observable properties, such as morphology, development, and behavior, characterize living taxa. The corresponding phenotypes may be controlled by the interplay between specific genotypes and the environment. To capture an individual's genotype at a specific time point, molecular tools for transcript quantification have followed the fast development of technologies, with Expressed Sequence Tags as the first approach to be historically used, especially suited for non-model taxa [56]. Nowadays, the physiological state of aquaculture species is retrieved through investigation of stage-specific and immune- and stress response–specific transcriptomic profiles using RNAseq. The corresponding computational workflows involve installing various tools at *Zorba* and implementing a series of steps that often take days to compute. These analyses, besides detecting transcripts at a specific physiological state, have successfully identified regulatory elements, such as microRNAs. Through the construction of a regulatory network with putative target genes, microRNAs have been linked to the transcriptome expression patterns. The most recent example is the identification of microRNAs and their putative target genes involved in ovary maturation [162].

Large-scale ecological statistics: are all taxa equal?

The nomenclature of living organisms, as well as their descriptions and their classifications under a specific nomenclature code, have been studied for more than 2 centuries.

Up to now, all the species present in an ecosystem have been considered equal in terms of their contributions to diversity. However, this axiom has been tested only once before, on the United Kingdom's marine animal phyla, showing the inconsistency of the traditional Linnaean classification between different major groups [163]. In Arvanitidis et al. [164], the average taxonomic distinctness index ($\Delta+$) and its variation ($Lambda+$) were calculated on a matrix deriving from the complete World Register of Marine Species [165], containing more than 250,000 described species of marine animals. It is the R-vLab web application, along with its HPC high RAM back-end components (on bigmem, see Section The IMBBC HPC Facility: From a Single Server to a Tier 2 System) that made such a calculation possible. This is the first time such a hypothesis has been tested on a global scale. Preliminary results show that the 2 biodiversity indices exhibit complementary patterns and that there is a highly significant yet non-linear relationship between the number of species within a phylum and the average distance through the taxonomic hierarchy.

Discovery of novel enzymes for bioremediation

Polychlorinated biphenyls are complex, recalcitrant pollutants that pose a serious threat to wildlife and human health. The identification of novel enzymes that can degrade such organic pollutants is being intensively studied in the emerging field of bioremediation. In the context of the Horizon 2020 Tools And Strategies to access original bioactive compounds by Cultivating MARine invertebrates and associated symbionts (TASCMAR) project, global ocean sampling provided a large biobank of fungal invertebrate symbionts and, through large-scale screening and bioreactor culturing, a marine-derived fungus able to remove a polychlorinated biphenyl compound was identified for the first time. *Zorba* resources and domain expertise in fungal genomics were used as a Centre for the Study and Sustainable Exploitation of Marine Biological Resources (CMBR) service for the analysis of multi-omic data for this symbiont. Following genome assembly of *Cladosporium sp.* TM-S3 [166], transcriptome assembly and a phylogenetic analysis revealed the full diversity of the symbiont's multicopper oxidases, enzymes commonly involved in oxidative degradation [167]. Among these, 2 laccase-like proteins shown to remove up to 71% of the polychlorinated biphenyl compound are now being expressed to optimize their use as novel biocatalysts. This step would not have been possible without the annotation of the Cladosporium genome with transcriptome data; mapping of the purified enzymes' LC-MS (Liquid chromatography–mass spectrometry) spectra against the set of predicted proteins allowed for identification of their corresponding sequences.

Lessons Learned

Depth and breadth are both required for a bioinformatics-oriented HPC

In our experience, the vast majority of the analyses run at the IMBBC HPC infrastructure are CPU-intensive. RAM-intensive jobs (> 128 GB RAM, see Section Computational Breakdown of the IMBBC HPC-Supported Research) represent only 0.3% of the total jobs executed over the last 2 years (see Section B1 in [148]). Despite the difference in the frequency of executed jobs with distinct requirements, serving both types of jobs and

ensuring their successful completion is equally important for addressing fundamental marine research questions (as shown in Section Computational Breakdown of the IMBBC HPC-Supported Research). The need for both HPC depth (a few high-memory nodes) and breadth (a number of slimmer nodes) has been previously reported [143]. This need reflects the idiosyncrasy of different bioinformatics analysis steps, often even within the same workflow. High-memory nodes are necessary for tasks such as de novo assembly of large genomes, while the availability of as many less powerful nodes as possible can speed up the execution of less demanding tasks and free resources for other users. Future research directions and the available budget further dictate tailoring of the HPC depth and breadth. Cloud-based services—e.g., for containerized workflows—may also facilitate this process once these become more affordable.

Quota . . . overloaded

We observed that independently of the type of analysis, storage was an issue for all *Zorba* users (Fig. 4). A high percentage of these issues relate to the raw data from HTS projects. These data are permanently stored in the home directories, occupying significant space. This, in conjunction with the fact that users delete their data with great reluctance, makes storage a major issue of daily use in *Zorba*. In specific cases where users' quota was exceeded uncontrollably, the *Zorba* team has been applying compression of raw and output data in contact with the user, but this is by no means a stable strategy. More generally, with the performance of the existing storage configuration in *Zorba* close to reaching its limits due to the increase in users and its concurrent use, several solutions have been adopted to resolve the issue. The most long-lasting solution has been the adoption of a per user quota system to allow storage sustainability and fairness in our allocation policy. This quota system nevertheless constitutes a limiting factor in pipeline execution, since lots of software tools produce unpredictably too many intermediate files, which not only increase storage but also cause job failures due to space restrictions. We managed the above issue by adding a scratch file system as an intermediate storage area for the runtime capacity needs. Following completion of their analysis, a user retains only the useful files and the rest are permanently removed. A storage upgrade scheduled within 2021 (see Section Future Directions) is expected to alleviate current storage challenges in *Zorba*. However, given the ever-increasing data production (e.g., as the result of decreasing sequencing costs and/or of rising imaging technologies), the responsible storage use approaches described here remain only partial solutions to anticipated future storage needs. Centralized (Tier 1 or higher) storage solutions represent a longer-term solution, which is in line with current views on how to handle big data generated by international research consortia in a long-lasting manner.

Continuous intercommunication among different disciplines matters

Smooth functioning of an HPC system and exploitation of its full potential for research requires stable employment of a core team of computer scientists and engineers, in close collaboration with an extended team of researchers. At least 4 disciplines are involved in *Zorba*-related issues: computer scientists, engineers, biologists (in the broad sense,

including ecologists, genomicists, etc.), and bioinformaticians with varying degrees of literacy in biology and informatics and various domain specializations (comparative genomics, biodiversity informatics, bacterial metagenomics, etc). The continuous communication among representatives of these 4 disciplines has substantially contributed to research supported by *Zorba* and to the evolution of the HPC system itself over time. In our experience, an HPC system cannot function effectively and for long without full-time system administrators, nor with bioinformaticians alone. Although it has not been the case since the system's onset, investment in monthly meetings, seminars, and training events (in biology, containers, domain-specific applications, and computer science; see Section The IMBBC HPC Facility: From a Single Server to a Tier 2 System) is the only way to establish stable intercommunication among different players of an HPC system. Such proximity translates into timely and adequate systems and bioinformatics analysis support, an element that in its turn translates into successful research (see Section Computational Breakdown of the IMBBC HPC-Supported Research). It should be noted that the overall good experience in connectivity among different HPC players derives from *Zorba* being a Tier 2 system, with a number of active permanent users in double digits. The establishment of such inter-communication was relatively straightforward to implement with periodic meetings and the assistance of ticketing and other management solutions (see Section C1 in [148]).

The way forward: develop locally and share and deploy centrally

The various approaches regarding the function of an HPC system are strongly related to the different viewpoints of the academic communities towards the relatively new disciplines of bioinformatics and big data. These approaches are strongly affected by national and international decisions that affect the ability to fund supercomputer systems. There are advantages in deploying bioinformatics-oriented HPC systems in centralized (Tier 0 and Tier 1) facilities: better prices at hardware purchases, easier access to HPC-tailored facilities (for instance, in terms of the cooling system and physical space), or experienced technical personnel (see also [143]). However, synergies between regional (Tier 2) and centralized HPC systems are fundamental for moving forward in supporting the diverse and demanding needs of bioinformatics. An example of such synergies concerns technical solutions (e.g., containerization) that address long-standing software sharing issues. In our experience, a workflow/pipeline can be developed by experts within the context of a specific project in a regional HPC facility. Once a production version of the pipeline is packaged, it can be distributed to centralized systems to cover a broader user audience (see Section The IMBBC HPC Facility From a Single Server to a Tier 2 System). Singularity containers have been developed to utterly suit HPC environments, mostly because they permit root access of the system in all cases. In addition, Singularity is compatible with all Docker images and can be used with Graphics Processing Units (GPUs) and Message Passing Interface (MPI) applications. This is why we chose to run containers in a Singularity format at *Zorba*. However, as Docker containers are widely used, especially in cloud computing (see more about cloud computing in Section Cloud Computing), workflows and services produced at IMBBC are offered in both container formats. Containers are an already established technology, used by the biggest cloud providers worldwide and

increasingly by non-profit research institutes. Despite indirect costs (e.g., costs to containerize legacy software), we believe that these technologies will become the norm in the future, especially in the context of reproducibility and interoperability of bioinformatics analysis.

Software optimizations for parallel execution

The most common ways of achieving implicit or explicit parallelization in modern multicore systems for bioinformatics, computational biology, and systems biology software tools are the software threads—provided by programming languages—and/or the OpenMP API [168]. These types of multiprocessing make good use of the available cores on a multicore system (single node), but they are not capable of combining the available CPU cores from more than 1 node. Some other software tools use MPI to spawn processing chunks to many servers and/or cores or (even better) combine MPI with OpenMP/Threads to maximize the parallelization in hybrid models of concurrency. Such designs are now used to a great extent in some cases, such as phylogeny inference software that makes use of Monte Carlo Markov Chain samplers. However, these cases are but a small number compared to the majority of bioinformatics tasks, while their usage in other analyses is low. At the hardware level, simultaneous multithreading is not enabled in the compute nodes of the IMBBC HPC infrastructure. Since the majority of analyses running on the cluster demand dedicated cores, hardware multithreading does not perform well. In our experience, the existence of more (logical) cores in compute nodes misleads the least experienced users into using more threads than the physically available ones, which slows down their executions. In comparison, assisting servers (filesystems, login nodes, web servers) make use of hardware multithreading, since they serve numerous small tasks from different users/sources that commonly contain Input/Output (I/O) operations. GPUs provide an alternative way for parallel execution, but they are supported by a limited number of bioinformatics software tools. Nevertheless, GPUs can optimize the execution process in specific, widely used bioinformatic analyses, such as sequence alignment [169, 170], image processing in microtomography (e.g., microCT), or basecalling of Nanopore raw data.

Cloud Computing

A recent alternative to traditional HPC systems, such as that described in this review, is cloud computing. Cloud computing is the way of organizing computing resources so they are provided over the Internet ("the cloud"). This paradigm of computing requires the minimum management effort possible [171]. Cloud computing providers exist in both commercial vendors and academic/publicly funded institutions and infrastructures (for more on cloud computing for bioinformatics, see Langmead and Nellore [172]). Computing resources can be reserved from individuals, institutions, organizations, or even scientific communities. The most widely-known commercial cloud providers are the "big 3" of cloud computing—namely, Amazon Web Services, Google Cloud Platform, and Microsoft Azure — while other cloud vendors are constantly emerging. Academic/publicly funded providers are also available: e.g., the EMBL–EBI Embassy Cloud.

Cloud computing services are being increasingly adopted in research, mainly because they offer simplicity and high availability to users with reduced or even no experience in HPC systems, through web interfaces. For this type of user, the time needed for data manipulation, software installation, and user-system interaction is significantly reduced compared to using a local HPC facility.

Container technologies, especially Docker, along with container-management systems such as Kubernetes combined with OpenStack, have been widely used in a number cloud computing systems, in particular in the research domain. It should be noted, however, that tool experimentation and benchmarking is more limited in cloud computing compared to local facilities and is costly, since it demands additional core hours of segmented computation. In-house HPC infrastructures can be fully configured to suit specific research area needs (storage available, fast interconnection for MPI jobs, number of CPUs versus available RAM, assisting services, etc.). Moreover, in cases where InfiniBand interconnection, a computer networking communications standard, is adopted in HPC, the performance in jobs and software that take advantage of it is substantial. Given the features and advantages of each approach (mentioned above) one could foresee the scenario of combining them to address the research community needs.

Future Directions

An upgrade of the existing hardware design of Zorba has been scheduled in 2021, funded by the CMBR research infrastructure (Fig. 1). More specifically:

3 nodes of 40 CPU physical cores will be added through new partitions (120 cores in total); the total RAM will be increased by 3.5 TB; 100 TB of cold storage will be installed and is expected to alleviate the archiving problem at the existing homes/scratch file systems; and the total usable existing storage capacity for users in home and scratch partitions will be increased by approximately 100 TB.

With this upgrade, it is expected that the total computational power of *Zorba* will be increased by approximately 6 TFlops, while the infrastructure will be capable of serving memory-intensive jobs requiring up to 1.5 TB of RAM, hosted on a single node. Eventually, more users will be able to concurrently load and analyze big data sets on the file systems. Over the coming 2 years, *Zorba* is also expected to have 2 major additions:

- the acquisition of a number of GPU nodes to build a new partition, especially for serving software that has been ported to run on GPUs; and
- the design of a parallel file system (Ceph or Lustre) to optimize concurrent I/O operations to speed up CPU-intensive jobs.

The expectation is that the upcoming upgrade of *Zorba* will further enhance collaborations with external users, since the types of bioinformatic tasks supported by the infrastructure are common to other disciplines beyond marine science, such as environmental omics research in the broad term. A nationwide survey targeting the community of researchers studying the environment and adopting the same approaches (HTS, biodiversity monitoring) has revealed that their computational and training needs are on the rise (A. Gioti et al., unpublished observations). Usage peaks and valleys were observed

in *Zorba* (see Section B1 in [148]), similarly to other HTS-oriented HPC systems [143]. It is therefore feasible to share *Zorba*'s idling time with other scientific communities. Besides, the *Zorba* upgrade is very timely in coming during a period where additional computational infrastructures emerge: the Cloud infrastructure Hypatia, funded by the Greek node of ELIXIR, is entering its production phase. It will constitute a national Tier 1 HPC facility, designed to host 50 computational nodes of different capabilities (regular servers, GPU-enabled servers, Solid-State Drive-enabled servers, etc.) and provide users the option to either create custom virtual machines for their computational services or to upload and execute workflows of containerized scientific software packages. In this context, a strategic combination of *Zorba* and Hypatia is expected to contribute to a strong computational basis in Greece. It is also expected that *Zorba* functionality will be augmented also through its connection with the Super Computing Installations of LifeWatch ERIC (European Research Infrastructure Consortium) (e.g., Picasso facility in Malaga, Spain). Building upon the lessons learned in the last 12 years, a foreseeable challenge for the facility is the enhancement of its usage monitoring to the example of international HPC systems [173], in order to allow even more efficient use of computational resources.

6.1.7 Conclusions

Zorba is an established Tier 2 HPC regional facility operating in Crete, Greece. It serves as an interdisciplinary computing hub in the eastern Mediterranean, where studies in marine conservation, invasive species, extreme environments, and aquaculture are of great scientific and socio-economic interest. The facility has supported, since its launch over a decade ago, a number of different fields of marine research, covering all kingdoms of life; it can also share part of its resources to support research beyond the marine sciences.

The operational structure of *Zorba* enables continuous communication between users and administrators for more effective user support, troubleshooting, and job scheduling. More specifically, training, regular meetings, and containerization of in-house pipelines have proven constructive for all teams, students, and collaborators of IMBBC. This operational structure has evolved over the years based on the needs of the facility's users and the available resources. The practical solutions adopted—from hardware (e.g., depth/breadth balanced structure, user quotas, and temporary storage) to software (e.g., modularized bioinformatics application maintenance and containerization) and human resource management (e.g., frequent intercommunication, continuous cross-discipline training)—reflect IMBBC research to a large extent. However, and by incrementing previous reviews [143], other Institutes and HPC facilities can be informed on the lessons learned (see Section Lessons Learned), and reflect on the computational requirement analysis of the methods presented (see Section Computational Breakdown of the IMBBC HPC-Supported Research) through the spectrum of their own research so as to plan ahead.

HPC facilities could reach a benefit greater than the sum of their capacities once they interconnect. The IMBBC HPC facility lies at the crossroad of 3 RIs, CMBR (Greek node of EMBRC-ERIC), LifeWatchGreece (Greek node of LifeWatch ERIC), and ELIXIR Greece, and via these will pursue further collaboration at larger Tier 0 and Tier 1 levels.

Chapter 7

Conclusions

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

Appendices

Bibliography

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

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

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

[4] 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 edna 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.

[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 i llumina 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.

[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, "obitools: 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.

[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, "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. suppl_1, pp. S96–S99, 2003.

[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, "Barcoding 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.

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

[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. Bapteste, "Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery," *Genome biology and evolution*, vol. 10, no. 3, pp. 707–715, 2018.

[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 metaphylogeography," *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.

[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. Čiampor, F. O. Costa, Z. Csabai, S. Duarte, M. F. Geiger, M. Grabowski, F. Rimet, B. Rulik, *et al.*, "Dna barcode reference libraries for the monitoring of aquatic biota in europe: Gap-analysis and recommendations for future work," *Science of the Total Environment*, vol. 678, pp. 499–524, 2019.

[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.

[114] NOAA, "How much water is in the ocean?," 2021. [Online; accessed 07-January-2022].

[115] J. A. Estes, M. Heithaus, D. J. McCauley, D. B. Rasher, and B. Worm, "Megafaunal impacts on structure and function of ocean ecosystems," *Annual Review of Environment and Resources*, vol. 41, pp. 83–116, 2016.

[116] K. R. Arrigo, "Marine microorganisms and global nutrient cycles," *Nature*, vol. 437, no. 7057, pp. 349–355, 2005.

[117] F. Boero and E. Bonsdorff, "A conceptual framework for marine biodiversity and ecosystem functioning," *Marine Ecology*, vol. 28, pp. 134–145, 2007.

[118] L. M. Beal, W. P. De Ruijter, A. Biastoch, and R. Zahn, "On the role of the agulhas system in ocean circulation and climate," *Nature*, vol. 472, no. 7344, pp. 429–436, 2011.

[119] K. Remoundou, P. Koundouri, A. Kontogianni, P. A. Nunes, and M. Skourtos, "Valuation of natural marine ecosystems: an economic perspective," *environmental science & policy*, vol. 12, no. 7, pp. 1040–1051, 2009.

[120] H.-O. Pörtner, D. C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, and N. Weyer, "The ocean and cryosphere in a changing climate," 2019.

[121] E. Sala and N. Knowlton, "Global marine biodiversity trends," *Annu. Rev. Environ. Resour.*, vol. 31, pp. 93–122, 2006.

[122] T. Tonon and D. Eveillard, "Marine systems biology," *Frontiers in genetics*, vol. 6, p. 181, 2015.

[123] H. M. Dionisi, M. Lozada, and N. L. Olivera, "Bioprospection of marine microorganisms: biotechnological applications and methods," *Revista argentina de microbiología*, vol. 44, no. 1, pp. 49–60, 2012.

[124] J. H. Tidwell and G. L. Allan, "Fish as food: aquaculture's contribution," *EMBO reports*, vol. 2, no. 11, pp. 958–963, 2001.

[125] G. Carvalho and L. Hauser, "Molecular genetics and the stock concept in fisheries," in *Molecular genetics in fisheries*, pp. 55–79, Springer, 1995.

[126] A. K. Sakai, F. W. Allendorf, J. S. Holt, D. M. Lodge, J. Molofsky, K. A. With, S. Baughman, R. J. Cabin, J. E. Cohen, N. C. Ellstrand, *et al.*, "The population biology of invasive species," *Annual review of ecology and systematics*, vol. 32, no. 1, pp. 305–332, 2001.

[127] G. A. Begg and J. R. Waldman, "An holistic approach to fish stock identification," *Fisheries research*, vol. 43, no. 1-3, pp. 35–44, 1999.

[128] M. Loreau, "Biodiversity and ecosystem functioning: recent theoretical advances," *Oikos*, vol. 91, no. 1, pp. 3–17, 2000.

[129] M. C. Leal, J. Puga, J. Serodio, N. C. Gomes, and R. Calado, "Trends in the discovery of new marine natural products from invertebrates over the last two decades–where and what are we bioprospecting?," *PLoS One*, vol. 7, no. 1, p. e30580, 2012.

[130] J. Norberg, D. P. Swaney, J. Dushoff, J. Lin, R. Casagrandi, and S. A. Levin, "Phenotypic diversity and ecosystem functioning in changing environments: a theoretical framework," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11376–11381, 2001.

[131] E. R. Mardis, "Next-generation dna sequencing methods," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 387–402, 2008.

[132] J. Kulski, *Next generation sequencing: advances, applications and challenges*. BoD–Books on Demand, 2016.

[133] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[134] J. G. Bundy, M. P. Davey, and M. R. Viant, "Environmental metabolomics: a critical review and future perspectives," *Metabolomics*, vol. 5, no. 1, pp. 3–21, 2009.

[135] V. Cahais, P. Gayral, G. Tsagkogeorga, J. Melo-Ferreira, M. Ballenghien, L. Weinert, Y. Chiari, K. Belkhir, V. Ranwez, and N. Galtier, "Reference-free transcriptome assembly in non-model animals from next-generation sequencing data," *Molecular ecology resources*, vol. 12, no. 5, pp. 834–845, 2012.

[136] N. A. Baird, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson, "Rapid snp discovery and genetic mapping using sequenced rad markers," *PloS one*, vol. 3, no. 10, p. e3376, 2008.

[137] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in rna-seq: a matter of depth," *Genome research*, vol. 21, no. 12, pp. 2213–2223, 2011.

[138] J. E. Goldford, N. Lu, D. Bajić, S. Estrela, M. Tikhonov, A. Sanchez-Gorostiaga, D. Segrè, P. Mehta, and A. Sanchez, "Emergent simplicity in microbial community assembly," *Science*, vol. 361, no. 6401, pp. 469–474, 2018.

[139] I. Merelli, H. Pérez-Sánchez, S. Gesing, and D. D'Agostino, "Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives," *BioMed research international*, vol. 2014, 2014.

[140] J.-i. Sohn and J.-W. Nam, "The present and future of de novo whole-genome assembly," *Briefings in bioinformatics*, vol. 19, no. 1, pp. 23–40, 2018.

[141] C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng, "Big data bioinformatics," *Journal of cellular physiology*, vol. 229, no. 12, pp. 1896–1900, 2014.

[142] S. Pal, S. Mondal, G. Das, S. Khatua, and Z. Ghosh, "Big data in biology: The hope and present-day challenges in it," *Gene Reports*, p. 100869, 2020.

[143] S. Lampa, M. Dahlö, P. I. Olason, J. Hagberg, and O. Spjuth, "Lessons learned from implementing a national infrastructure in sweden for storage and analysis of next-generation sequencing data," *Gigascience*, vol. 2, no. 1, pp. 2047–217X, 2013.

[144] T. Sterling, M. Brodowicz, and M. Anderson, *High performance computing: modern systems and practices*. Morgan Kaufmann, 2017.

[145] Wikipedia contributors, "Supercomputing in europe — Wikipedia, the free encyclopedia," 2021. [Online; accessed 7-January-2022].

[146] E. Lindahl, "The scientific case for computing in europe 2018-2026," 2018.

[147] L. Candela, D. Castelli, and P. Pagano, "Virtual research environments: an overview and a research agenda. data sci. j," *J*, vol. 12, pp. 75–81, 2013.

[148] H. Zafeiropoulos, A. Gioti, S. Ninidakis, A. Potirakis, S. Paragkamian, N. Angelova, A. Antoniou, T. Danis, E. Kaitetzidou, P. Kasapidis, J. B. Kristoffersen, V. Papadogiannis, C. Pavloudi, Q. V. Ha, J. Lagnel, N. Pattakos, G. Perantinos, D. Sidirokastritis, P. Vavilis, G. Kotoulas, T. Manousaki, E. Sarropoulou, C. S. Tsigenopoulos, C. Arvanitidis, A. Magoulas, and E. Pafilis, "The IMBBC HPC facility: history, configuration, usage statistics and related activities." HZ and NG contributed equally at this work. Correspondence to: E. Pafilis; pafilis@hcmr.gr, Apr. 2021.

[149] J. J. Dongarra, P. Luszczek, and A. Petitet, "The linpack benchmark: past, present and future," *Concurrency and Computation: practice and experience*, vol. 15, no. 9, pp. 803–820, 2003.

[150] T. Castrignanò, S. Gioiosa, T. Flati, M. Cestari, E. Picardi, M. Chiara, M. Fratelli, S. Amente, M. Cirilli, M. A. Tangaro, *et al.*, "Elixir-it hpc@ cineca: high performance computing resources for the bioinformatics community," *BMC bioinformatics*, vol. 21, no. 10, pp. 1–17, 2020.

[151] J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, "Stacks: an analysis tool set for population genomics," *Molecular ecology*, vol. 22, no. 11, pp. 3124–3140, 2013.

[152] C. Varsos, T. Patkos, A. Oulas, C. Pavloudi, A. Gougousis, U. Z. Ijaz, I. Filiopoulou, N. Pattakos, E. V. Berghe, A. Fernández-Guerra, *et al.*, "Optimized r functions for analysis of ecological community data using the r virtual laboratory (rvlab)," *Biodiversity data journal*, no. 4, 2016.

[153] K. E. Klymus, N. T. Marshall, and C. A. Stepien, "Environmental dna (edna) metabarcoding assays to detect invasive invertebrate species in the great lakes," *PloS one*, vol. 12, no. 5, p. e0177643, 2017.

[154] M. Bariche, S. A. Al-Mabruk, M. A. Ateş, A. Büyük, F. Crocetta, M. Dritsas, D. Edde, A. Fortič, E. Gavriil, V. Gerovasileiou, *et al.*, "New alien mediterranean biodiversity records (march 2020)," *Mediterranean Marine Science*, vol. 21, no. 1, pp. 129–145, 2020.

[155] S. Katsanevakis, M. Coll, C. Piroddi, J. Steenbeek, F. Ben Rais Lasram, A. Zenetos, and A. C. Cardoso, "Invading the mediterranean sea: biodiversity patterns shaped by human activities," *Frontiers in Marine Science*, vol. 1, p. 32, 2014.

[156] M. Pauletto, T. Manousaki, S. Ferraresso, M. Babbucci, A. Tsakogiannis, B. Louro, N. Vitulo, V. H. Quoc, R. Carraro, D. Bertotto, *et al.*, "Genomic analysis of sparus aurata reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish," *Communications biology*, vol. 1, no. 1, pp. 1–13, 2018.

[157] E. Sarropoulou, A. Y. Sundaram, E. Kaitetzidou, G. Kotoulas, G. D. Gilfillan, N. Papandroulakis, C. C. Mylonas, and A. Magoulas, "Full genome survey and dynamics of gene expression in the greater amberjack seriola dumerili," *GigaScience*, vol. 6, no. 12, p. gix108, 2017.

[158] A. Tsakogiannis, T. Manousaki, V. Anagnostopoulou, M. Stavroulaki, and E. T. Apostolaki, "The importance of genomics for deciphering the invasion success of the seagrass halophila stipulacea in the changing mediterranean sea," *Diversity*, vol. 12, no. 7, p. 263, 2020.

[159] T. Danis, A. Tsakogiannis, J. B. Kristoffersen, D. Golani, D. Tsaparis, P. Kasapidis, G. Kotoulas, A. Magoulas, C. S. Tsigenopoulos, and T. Manousaki, "Building a high-quality reference genome assembly for the the eastern mediterranean sea invasive sprinter lagocephalus sceleratus (tetraodontiformes, tetraodontidae)," *Biorxiv*, 2020.

[160] N. Angelova, T. Danis, L. Jacques, C. Tsigenopoulos, and T. Manousaki, "SnakeCube: containerized and automated next- generation sequencing (NGS) pipelines for genome analyses in HPC environments," Apr. 2021.

[161] P. Natsidis, A. Tsakogiannis, P. Pavlidis, C. S. Tsigenopoulos, and T. Manousaki, "Phylogenomics investigation of sparids (teleostei: Spariformes) using high-quality proteomes highlights the importance of taxon sampling," *Communications biology*, vol. 2, no. 1, pp. 1–10, 2019.

[162] M. Papadaki, E. Kaitetzidou, C. C. Mylonas, and E. Sarropoulou, "Non-coding rna expression patterns of two different teleost gonad maturation stages," *Marine Biotechnology*, vol. 22, no. 5, pp. 683–695, 2020.

[163] R. Warwick and P. Somerfield, "All animals are equal, but some animals are more equal than others," *Journal of Experimental Marine Biology and Ecology*, vol. 366, no. 1-2, pp. 184–186, 2008.

[164] C. D. Arvanitidis, R. M. Warwick, P. J. Somerfield, C. Pavloudi, E. Pafilis, A. Oulas, G. Chatzigeorgiou, V. Gerovasileiou, T. Patkos, N. Bailly, *et al.*, "Research infrastructures offer capacity to address scientific questions never attempted before: Are all taxa equal?," tech. rep., PeerJ Preprints, 2018.

[165] L. Vandepitte, B. Vanhoorne, W. Decock, S. Vranken, T. Lanssens, S. Dekeyzer, K. Verfaille, T. Horton, A. Kroh, F. Hernandez, *et al.*, "A decade of the world register of marine species–general insights and experiences from the data management team: Where are we, what have we learned and how can we continue?," *PLoS One*, vol. 13, no. 4, p. e0194599, 2018.

[166] A. Gioti, R. Siaperas, E. Nikolaivits, G. Le Goff, J. Ouazzani, G. Kotoulas, and E. Topakas, "Draft genome sequence of a cladosporium species isolated from the mesophotic ascidian didemnum maculosum," *Microbiology resource announcements*, vol. 9, no. 18, pp. e00311–20, 2020.

[167] E. Nikolaivits, R. Siaperas, A. Agrafiotis, J. Ouazzani, A. Magoulas, A. Gioti, and E. Topakas, "Functional and transcriptomic investigation of laccase activity in the presence of pcb29 identifies two novel enzymes and the multicopper oxidase repertoire of a marine-derived fungus," *Science of The Total Environment*, vol. 775, p. 145818, 2021.

[168] L. Dagum and R. Menon, "Openmp: an industry standard api for shared-memory programming," *IEEE computational science and engineering*, vol. 5, no. 1, pp. 46–55, 1998.

[169] P. D. Vouzis and N. V. Sahinidis, "Gpu-blast: using graphics processors to accelerate protein sequence alignment," *Bioinformatics*, vol. 27, no. 2, pp. 182–188, 2011.

[170] M. S. Nobile, P. Cazzaniga, A. Tangherloni, and D. Besozzi, "Graphics processing units in bioinformatics, computational biology and systems biology," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 870–885, 2017.

[171] P. Mell, T. Grance, *et al.*, "The nist definition of cloud computing," 2011.

[172] B. Langmead and A. Nellore, "Cloud computing for genomic data analysis and collaboration," *Nature Reviews Genetics*, vol. 19, no. 4, pp. 208–219, 2018.

[173] M. Dahlö, D. G. Scofield, W. Schaal, and O. Spjuth, "Tracking the ngs revolution: managing life science research on shared high-performance computing clusters," *GigaScience*, vol. 7, no. 5, p. giy028, 2018.

Short CV

Education

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

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

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

Research projects - working Experience

- **A workflow for marine Genomic Observatories data analysis** (2020 - ongoing)
  Role: technical support & principal investigator
  This EOSC-Life funded project aims at developing a workflow for the analysis of EMBRC's Genomic Observatories (GOs) data, allowing researchers to deal better with this increasing amount of the data and make them more easily interpretable.

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

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

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

Awards

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

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

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

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

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

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

Publications

- **PREGO:**
  Zafeiropoulos, H.,

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

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

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

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

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