ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

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

Haris Zafeiropoulos

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

Promotors:

Prof. Emmanouil Ladoukakis
Dr Evangelos Pafilis
Dr Christoforos Nikolaou

Academic year 2021 – 2022

Members of the examination committee
&
reading committee

Prof. Emmanouil Ladoukakis
Univeristy of Crete
Biology Department

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

Dr Christoforos Nikolaou
Biomedical Sciences Research Center "Alexander Fleming"
Institute of Bioinnovation

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

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

Preface

Haris Zafeiropoulos

Contents

Preface **i**

Contents **ii**

Abstract **iv**

Περίληψη **v**

List of Figures and Tables **vi**

List of Abbreviations and Symbols **vii**

1 Introduction **1**

1.1 Microbial communities: structure & function 1

1.1.1 The role of microbial communities in biogeochemical cycles 1

1.1.2 Microbial interactions: unravelling the microbiome 1

1.2 Bioinformatics challenges in HTS approaches 1

1.3 Data integration & data mining in the era of omics 1

1.3.1 Metadata: a key issue for the microbiome community 1

1.3.2 Ontologies & databases: the corner stone of mordern biology . . . 3

1.4 Sampling the flux space of a metabolic model: challenges & potential . . . 3

1.5 The hypersaline Tristomo swamp: a case study of an extreme environment 3

1.6 Systems biology from a computational resources point-of-view 3

1.7 Aims and objectives 3

2 Software development to establish quality HTS-oriented bioinformatics methods for microbial diversity assessment **5**

2.1 PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes 5

2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data 6

2.3 A workflow for marine Genomic Observatories data analysis 6

3 Software development to build a knowledge-base at the systems biology level **9**

3.1 PREGO: a literature- and data-mining resource to associate microorganisms, biological processes, and environment types 9

4 Software development to establish metabolic flux sampling approaches at the community level **11**

4.1 Genome-scale metabolic model analysis 11

4.2 A New MCMC Algorithm for Sampling the Flux Space of Metabolic Networks 11
4.3 Flux sampling at the community level 11

5 Microbial interactions inference in communities of a hypersaline swamp elucidate mechanisms governing taxonomic & functional profiles 13
5.1 Amplicon & shotgun metagenomic analysis 13
5.2 Inferring microbial interactions 13

6 0s and 1s in marine molecular research 15
6.1 Computing resources: a prerequisite & a limitation in modern microbial ecology 15
6.2 High Performance Computing and Cloudification: scaling up bioinformatics analysis 15

7 Conclusions 17

Bibliography 21

Abstract

Περίληψη

Και στα ελληνικά

List of Figures and Tables

List of Figures

2.1 The PEMA workflow: figure from publication 6
2.2 DARN methodology: figure in the publication 7

3.1 PREGO methodology: figure in the publication under submission 10

4.1 Our MMCS algorithm and its first phases. Figure published on SoCG21 11

6.1 Computing requirements of the published studies performed on the IMBBC HPC facility over the last decade. Figure from publication. 15

List of Tables

List of Abbreviations and Symbols

Abbreviations

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

Chapter 1

Introduction

1.1 Microbial communities: structure & function

Microbes, i.e. Bacteria, Archaea and small Eukaryotes such as protoza, are omnipresent and impact global ecosystem functions [1] through their abundance [2], versatility [3] and interactions [4].

1.1.1 The role of microbial communities in biogeochemical cycles

1.1.2 Microbial interactions: unravelling the microbiome

1.2 Bioinformatics challenges in HTS approaches

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" [5] 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

1.4 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

Publication relative to this chapter: [6].

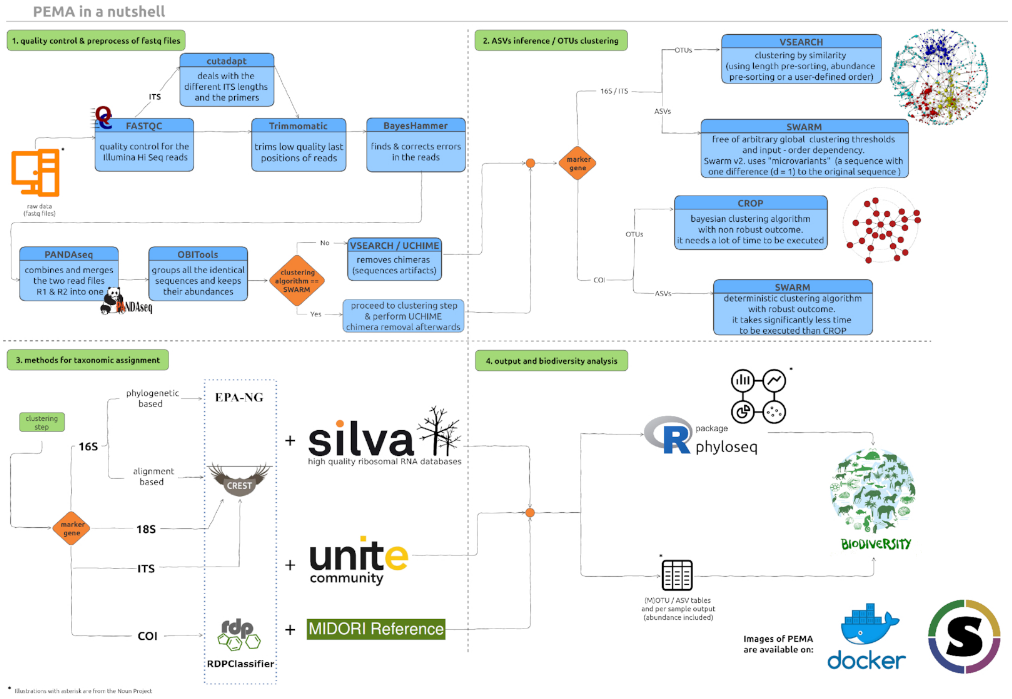

PEMA in a nutshell
1. quality control & preprocess of fastq files
raw data
(fastq files)
FASTQC
quality control for the Illumina Hi Seq reads
ITS
cutadapt
deals with the different ITS lengths and the primers
Trimmomatic
trims low quality last positions of reads
BayesHammer
finds & corrects errors in the reads
PANDAseq
combines and merges the two read files R1 & R2 into one
PANDAseq
OBITools
groups all the identical sequences and keeps their abundances
clustering algorithm == SWARM
No
VSEARCH / UCHIME
removes chimeras (sequences artifacts)
Yes
proceed to clustering step & perform UCHIME chimera removal afterwards
2. ASVs inference / OTUs clustering
marker gene
16S / ITS
OTUs
VSEARCH
clustering by similarity (using length pre-sorting, abundance pre-sorting or a user-defined order)
ASVs
SWARM
free of arbitrary global clustering thresholds and input - order dependency. Swarm v2. uses "microvariants" (a sequence with one difference (d = 1) to the original sequence)
COI
OTUs
CROP
bayesian clustering algorithm with non robust outcome. it needs a lot of time to be executed
ASVs
SWARM
deterministic clustering algorithm with robust outcome. it takes significantly less time to be executed than CROP
3. methods for taxonomic assignment
clustering step
marker gene
16S
phylogenetic based
alignment based
18S
ITS
COI
EPA-NG
CREST
RDPClassifier
+ silva
high quality ribosomal RNA databases
+ unite community
+ MIDORI Reference
4. output and biodiversity analysis
package
phyloseq
BIODIVERSITY
(M)OTU / ASV tables and per sample output (abundance included)
Images of PEMA are available on:
docker
* Illustrations with asterisk are from the Noun Project

FIGURE 2.1: The PEMA workflow: figure from publication

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

Publication relative to this chapter: [7]

2.3 A workflow for marine Genomic Observatories data analysis

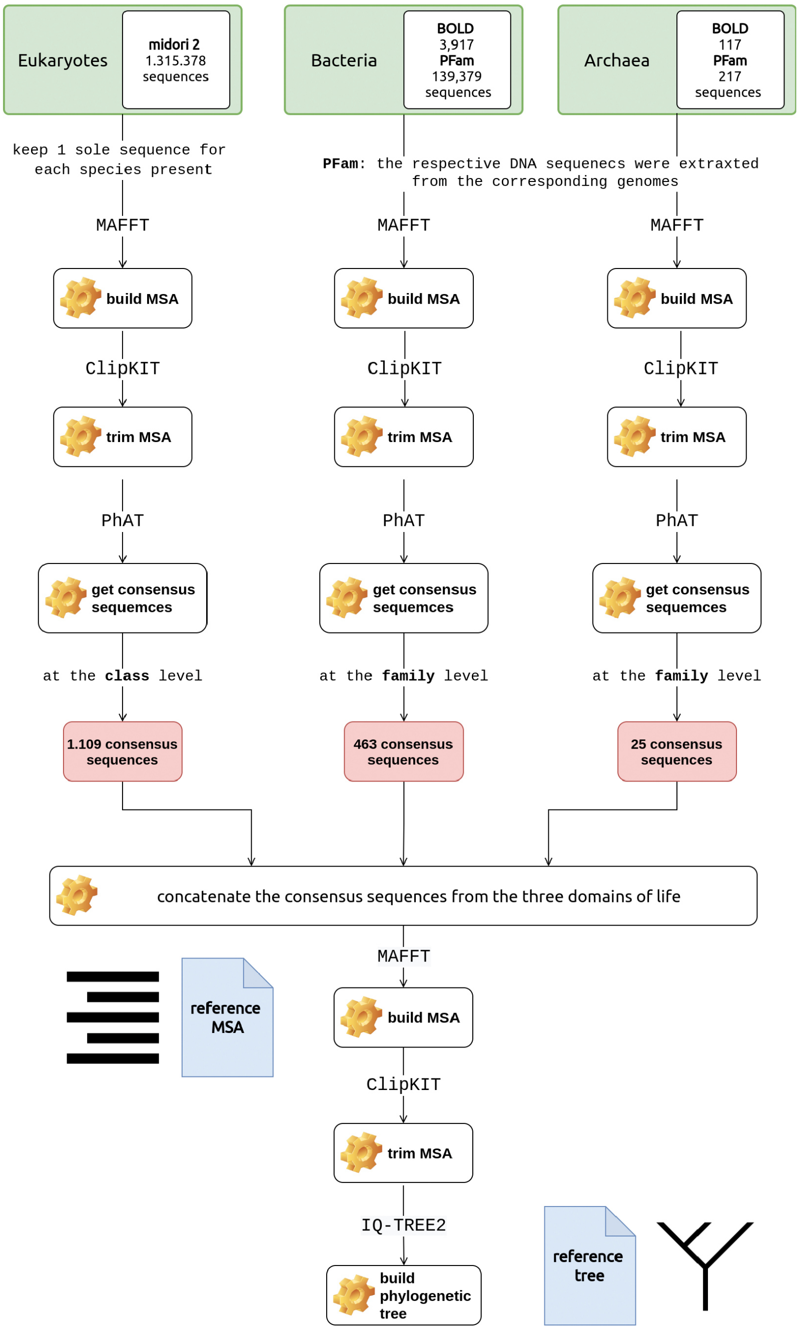


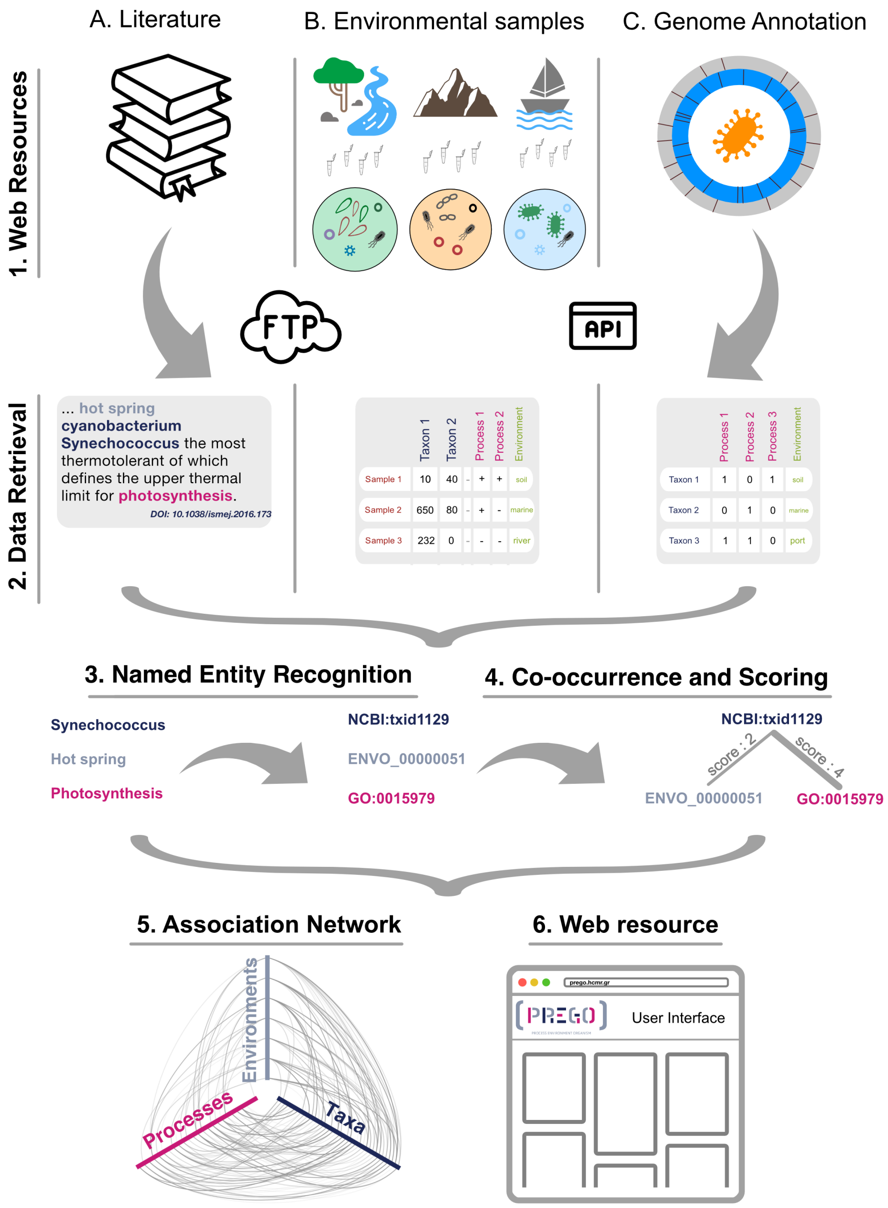
FIGURE 2.2: DARN methodology: figure in the publication

Chapter 3

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

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

Publication relative to this chapter: under submission



	Taxon 1	Taxon 2		Process 1	Process 2	Environment
Sample 1	10	40	-	+	+	soil
Sample 2	650	80	-	+	-	marine
Sample 3	232	0	-	-	-	river

	Process 1	Process 2	Process 3	Environment
Taxon 1	1	0	1	soil
Taxon 2	0	1	0	marine
Taxon 3	1	1	0	port

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

Chapter 4

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

4.1 Genome-scale metabolic model analysis

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

Publication relative to this chapter: [8]

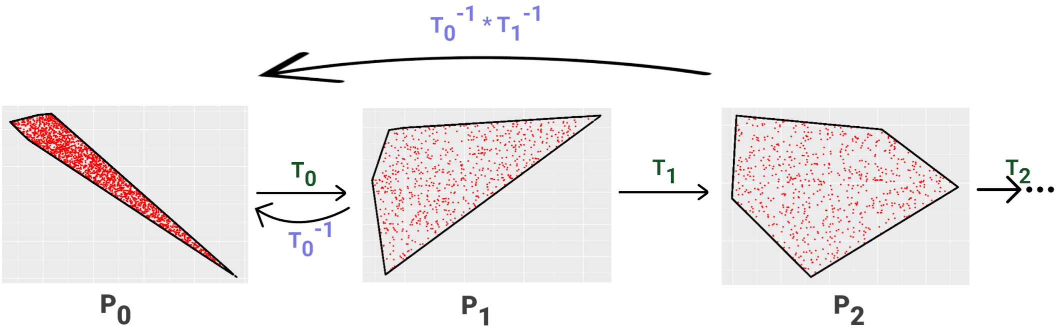


FIGURE 4.1: Our MMCS algorithm and its first phases. Figure published on SoCG21

4.3 Flux sampling at the community level

Chapter 5

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

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

5.1 Amplicon & shotgun metagenomic analysis

darn and PEMA will be used at this point, among other software

5.2 Inferring microbial interactions

PREGO and dingo will be used to this end

Chapter 6

0s and 1s in marine molecular research

Publication relative to this chapter: [9]

6.1 Computing resources: a prerequisite & a limitation in modern microbial ecology

6.2 High Performance Computing and Cloudification: scaling up bioinformatics analysis

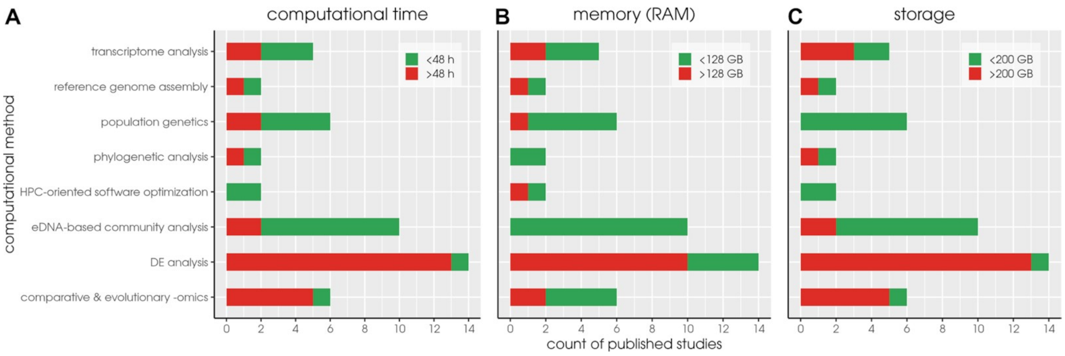


FIGURE 6.1: Computing requirements of the published studies performed on the IMBBC HPC facility over the last decade. Figure from publication.

Chapter 7

Conclusions

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

Appendices

Bibliography

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] H. Zafeiropoulos, L. Gargan, S. Hintikka, C. Pavloudi, and J. Carlsson, “The dark matter investigator (darn) tool: getting to know the known unknowns in coi amplicon data,” *Metabarcoding and Metagenomics*, vol. 5, p. e69657, 2021.

[8] 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.

[9] 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.