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

July, 2022

Members of the examination committee
&
reading committee

Prof. Emmanouil Ladoukakis
Univeristy of Crete
Biology Department

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

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

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

Prof. Panagiotis Sarris
University of Crete
Department of Biology

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

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

Preface

This thesis is the product of a series of chance-and-necessity events. My home village, my father and a book [1] made me appreciate nature and think highly of the relationship between humans and their environment. The first two, they also made me, among other things, rather stubborn. This stubbornness led me to first join the Technical University of Patras, even if Biology was my passion since secondary school, but then, somehow back to biology and the University of Athens. Yet, it was only chance that got me in the city that shaped me the most. A few months before defending my bachelor's thesis, Prof. Pedro Jordano, who has contributed the most to the study of ecological networks and interactions, pointed out to me that I need first to get some expertise in Bioinformatics if I am interested in his work. This was the first time I ever considered of Bioinformatics. It was the same year that a relative MSc in the University of Crete ran for the first time. A few years later and mostly thanks to Dr. Christina Pavloudi it is hard to tell *what if* I was not studying microbial assemblages.

Therefore, it is all the good and bad circumstances and choices as well as the people related to each of those that brought me here today, and it is them that I would like to acknowledge first.

I would like to thank my promotors; Prof. Manolis Ladoukakis, a university teacher with whom we met when I invaded his office crying for help and he just said *"breathe; we will figure this out"*. Since then, I enjoy his guidance for which I am grateful. Dr. Evangelos Pafilis with whom we started working together back in 2017 in the framework of my MSc thesis and 6 years later, here we are. Dr. Christoforos Nikolaou was also among our MSc teachers and has been an influence to me since I first came in Crete. Also, Dr. Christina Pavloudi with whom we met during my MSc and since then, she has not stopped triggering my curiosity and has (almost) never complained with my endless questions to her. Special thanks to Dr. Apostolos Chalkis, probably the greatest example of a necessity-and-chance case in my research story so-far, with whom we met on the streets of Athens and a few years later he called me to ask me if I had ever heard of metabolic networks; this was the first time I ever heard of them. Prof. Elias Tsigaridas and Dr. Vissarion Fisikopoulos thank you for your patience and spirit; I learnt more than a lot working with the GeomScale group. I would also like to say a great thank you to the Area52 lab. To Prof. Jens Carlsson who convinvced me to look for "aliens" and proved me there are people that have no idea about how a good olive oil tastes like! To Dr. Sanni Hintikka and to Dr. Laura Gargan that would always make some time for me, even from the back seat of a car during lunch time.

[1] *My life as an Indian*, by J. W. Schultz (1907)

Last but not least, Dr. George Kotoulas who has been an inspiration to me from so many and different aspects.

I would also like thank and fairwell all members of the Biodiversity lab and colleagues in HCMR. Χλαπάτσα, Niki, Eva, Emma, Eirini, George, Bill, Antoni, Stelio, Natassa, Noche Sangre, Adrianouko, Eirini, Despoina, thank you for making the every-day-life so companionable; I will miss you all. The most special thanks to Savvas Paragkamian; over the last years, we shared a desk, a great number of issues and bugs, our ideas and our temper.

Last, I would like to declare my respect and gratitude to all those that set the example for me at every turn. At a collective level, the Communist Liberation youth; it might gets tough from time to time, but struggling to interpret this world and fighting to change it, is most nobel fight one might fight. And of course, to my corner; Tsocha, Nef, Angelique, Annoula, Leo, καλύτερα της πιάτσας, in the most various ways, you have been my *lee side*. I love you all and you are in my heart. Σκουπίδια this applies for all of you too! My parents that have tought me that there is nothing common in *common people*. I am proud of you and there is no way I can thank you enough. My sister, my super hero. Last, στο δίπλα μαξιλάρι αριστερά/ κοιμάται ο άνθρωπος ο φίλος και η λύση μου. Αφούλη thank you for being my friend, my inspiration, my boxing bag, my colleague, my courage, my caress, my coffee and my tobacco.

Που πας· άκουσες ποτέ την κραυγή της θάλασσας·
...
Και μέσα στα ανοιχτά με σκέπασε ένα κίνημα
σκόρπια διαδήλωση και δύσκολο ξεκίνημα
και σου ζητάω συγγνώμη αν λίγο τη χάλασα
μα η διαδήλωση μια μέρα έγινε θάλασσα

To a little seed, meant to join us any time now
Στον νεότερο από τους σποράκους που όπου να'ναι φτάνει

Haris Zafeiropoulos

Contents

Abstract

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

Aim of this PhD was to enhance microbiome data analyses by developing software addressing on-going computational challenges on the study of microbial communities. On top of that, to exploit such state-of-the-art methods to study microbial assemblages in extreme environments. To this end, the Tristomo marsh in Karpathos island (Greece), was chosen as a study case.

Environmental DNA and metabarcoding have been widely used to estimate the biodiversity (the *who*) and the structure of communities. Vast amount of sequencing data targeting certain marker genes depending the taxonomic group of interest become available thanks to High Throughput Sequencing technologies. However, the bioinformatics analysis of such data require multiple steps and parameter settings as well as increase computing resources. Workflows along with computing infrastructures ease this need to a great extent; in this nontion, a Pipeline for environmental DNA Metabarcoding Analysis (PEMA) was developed (Chapter 2.1). However, eDNA metabarcoding has limitations too. Cytochrome c oxidase subunit I (COI) marker gene is a commonly used marker gene, especially in studies targeting eukaryotic taxa. It is well known that in COI studies a great number of the derived Operational Taxonomic Unitss (OTUs) get no taxonomic hits. The presence of pseudogenes but also of non-eukaryotic taxa among the amplicon data, with the simultaneous absence of the latter from the most commonly-used reference databases justify this phenomenon to a great extent. To identify such cases the Dark mAtteR iNvestigator (DARN) software was developed; DARN makes use of a COI-oriented tree of life to provide further insight to such known unknown sequences (Chapter 2.2).

Amplicon and shotgun metagenomics approaches along with the rest of the omics technologies, have led to vast amount of data and metadata, recording the *who*, the *what* and the *where*. To enable optimal accessibility and usage of this information, a great

number of databases, ontologies as well as community-standards have been developed. By exploiting data integration techniques to bring such bits of information together, as well as text mining methods to retrieve knowledge "hidden" among the billions of text lines in already published literature, the PREGO knowledge-base returns thousands of *what - where - who* potential associations (Chapter 3).

The driving question though is *how* the different microbial taxa ascertain their endurance as part of a community. Metabolic interactions among the various taxa play a decisive role for the composition of such assemblages. Genome-scale metabolic networks (GEMs) enable the inference of such interactions. Random sampling on the flux space of such metabolic models, provides a representation of the flux values a model can get under various conditions. However, flux sampling is challenging from a computational point of view, especially as the dimension of a metabolic model increases. To address such challenges, a Python library called dingo was developed using a Multiphase Monte Carlo Sampling algorithm (Chapter 4).

Finally, sediment and microbial mat samples as well as microbial aggregates from a hypersaline marsh in Tristomo bay (Karpathos, Greece) were analyzed. Both amplicon (16S rRNA) and shotgun sequencing data were used to characterize the microbial structure of the communities and environmental parameters (e.g. salinity, oxygen concentration) were measured at the sampling sites. Key functions supporting life in such environments were identified and metagenome-assembled genomes (MAGs) of novel species found were built (Chapter 5).

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

Περίληψη

Οι μικροβιακές κοινότητες αποτελούν ακρογωνιαίο λίθο για τους περισσότερους τύπους οικοσυστημάτων. Για να διευκρινιστούν οι μηχανισμοί που καθορίζουν τέτοιες κοινότητες είναι καθοριστικής σημασίας η αναγνώριση των τάξων που τις απαρτίζουν (ποιος) καθώς και των διεργασιών που πραγματοποιούνται (τι) στους διάφορους τύπους περιβαλλόντων (που). Χάρη σε μια σειρά τεχνολογικών επιτευγμάτων, ιδιαίτερα μεγάλες ποσότητες πληροφορίας/δεδομένων από όλα τα επίπεδα οργάνωσης της ζωής έχουν σωρευτεί τις τελευταίες δεκαετίες. Σε αυτό το πλαίσιο, οι μελέτες μικροβιακής οικολογίας είναι άρρηκτα συνδεδεμένες και βασίζονται σε βιοπληροφορικές μεθόδους και αναλύσεις. Ωστόσο, έχει προκύψει ένας σημαντικός αριθμός προκλήσεων τόσο από την βιολογική σκοπιά όσο και από αυτήν την επιστήμης υπολογιστών. Μεταξύ αυτών, καθοριστικό ερώτημα αποτελεί το τι μπορούμε να κάνουμε με όλα αυτά τα επιμέρους κομμάτια πληροφορίας·. Το παράδειγμα της Βιολογίας Συστημάτων απαντά σε αυτό το ερώτημα περνώντας από πιο αναγωγικές σε πιο ολιστικές προσεγγίσεις προσπαθώντας να ερμηνεύσει το πως προκύπτουν και συνδέονται οι ιδιότητες ενός συστήματος.

Στόχος αυτής της διδακτορικής διατριβής ήταν να ενισχύσει την ανάλυση δεδομένων από μικροβιώματα αναπτύσσοντας λογισμικά εργαλεία που να απαντούν σε τρέχουσσες υπολογιστικές προκλήσεις για την μελέτη μικροβιακών κοινοτήτων. Επιπλέον, να μελετήσει μικροβιακές κοινότητες σε ακραία περιβάλλοντα εφαρμόζοντας σύγχρονες μεθόδους για την αναγνώριση τάξων και διεργασιών. Για την επίτευξη αυτού του στόχου, το έλος Τριστόμου στο νησί της Καρπάθου, επιλέχθηκε ως περιοχή μελέτης.

Το περιβαλλοντικό DNA και η μέθοδος της μετακωδικοποίησης έχουν χρησιμοποιηθεί σημαντικά για την εκτίμηση της βιοποικιλότητας (ποιος) και τη δομή των κοινοτήτων. Σημαντικός αριθμός αλληλουχικών δεδομένων που στοχεύουν σε ορισμενα γονίδια δείκτες και που εξαρτόνται από τις ταξινομικές ομάδες στόχους, είναι διαθέσιμα χάρη στις τεχνικές αλληλούχισης υψηλής απόδοσης HTS. Ωστόσο, η βιοπληροφορική ανάλυση τέτοιων δεδομένων απαιτούν μεγάλο αριθμό βημάτων και παραμέτρων καθώς και σημαντικούς υπολογιστικούς πόρους. Οι ροές εργασιών σε συνδυασμό με υπολογιστικές υποδομές μπορούν να απαντήσουν σε αυτές τις απαιτήσεις σε σημαντικό βαθμό. Σε αυτό το πλαίσιο αναπτύχθηκε η ροή εργασίας PEMA με στόχο την ανάλυση δεδομένων μετακωδικοποίησης από περιβαλλοντικό DNA. Κεφάλαιο 2.1. Ωστόσο, η μέθοδος μετακωδικοποίησης χαρακτηρίζεται από σειρά περιορισμών. Η υπομονάδα I της κυτοχρωμικής οξειδάσης c (COI), αποτελεί έναν δείκτη που χρησιμοποιείται ευρέως, ειδικά στην περίπτωση ευκαρυωτικών τάξων - στόχων. Είναι γνωστό πως σε μελέτες όπου ο δείκτης αυτός χρησιμοποιείται, ένας μεγάλος αριθμός των λειτουργικών ταξινομικών μονάδων (OTUs) που προκύπτουν, δεν καταφέρνουν να ταυτοποιηθούν. Η παρουσία τοσο

ψευδογονιδίων όσο όμως και μη-ευκαρυωτικών τάξων ανάμεσα σε τέτοια αλληλουχικά δεδομένα, με την ταυτόχρονη απουσία των τελευταίων από τις βάσεις αναφοράς, εξηγεί την μη ταυτοποίησή τους σε σημαντικό βαθμό. Για την αναγνώριση τέτοιων περιπτώσεων, αναπτύχθηκε το υπολογιστικό εργαλείο DARN το οποίο αξιοποιεί ένα φυλογενετικό δέντρο που καλύπτει και τις 3 επικράτειες του δέντρου της ζωής, βασισμένο σε αλληλουχίες του δείκτη Κεφάλαιο COI, Κεφάλαιο 2.2.

Μέθοδοι γονιδίων δεικτών και μεταγονιδιωματικής καθώς και το σύνολο των μεθόδων αλληλούχισης υψηλή απόδοσης, έχουν οδηγήσει στην σώρευση σημαντικά μεγάλου αριθμού δεδομένων και μεταδεδομένων καταγράφοντας τάξα και διεργασίες σε σειρά τύπους περιβαλλόντων. Για να επιτρέψουν την βέλτιστη προσβασιμότητα και αξιοποίηση αυτής της πληροφορίας, έχουν δημιουργηθεί σειρά βάσεων δεδομένων, οντολογιών αλλά και προτύπων - κανόνων για να ακολουθεί η κοινότητα για την καταχώρηση τους. Αξιοποιώντας μεθόδους ενσωμάτωσης/ολοκλήρωσης δεδομένων **data integration** για την εύρεση των διάφορων κομματιών πληροφορίας και την συσχέτισή τους, καθώς και τεχνικών εξόρυξης κειμένου **text mining** για την ανάκτηση γνώσης από το σύνολος της δημόσια διαθέσιμης βιβλιογραφίας αναπτύχθηκε η βάση-γνώσης PREGO, Κεφάλαιο 3. η οποία επιστρέφει χιλιάδες σχέσεις μεταξύ τάξων, περιβαλλόντων και διεργασιών.

Καθοριστικό ερώτημα ωστόσο σε οτι αφορά τις μικροβιακές κοινότητες, αποτελεί το ¨πώς' τα διάφορα μικροβιακά τάξα εξασφαλίζουν την θέση τους ως μέλη της κοινότητας. Μεταβολικές αλληλεπιδράσεις μεταξύ των διάφορων τάξων παίζουν καθοριστικό ρόλο για την συγκρότηση τέτοιων κοινοτήτων. Μεταβολικά δίκτυα στην κλίμακα του γονιδιώματος (GEMs) επιτρέπουν την αναγνώριση τέτοιων αλληλεπιδράσεων. Η τυχαία δειγματοληψία στον χώρο που ορίζεται από τις πιθανές τιμές που μπορεί να πάρουν οι ροές των αντιδράσεων (**flux sampling**) επιτρέπει την αναπαράσταση των τιμών που μπορεί να λάβουν αυτές οι ροές κάτω από συγκεκριμένες συνθήκες. Ωστόσο η μέθοδος **flux sampling** είναι ιδιαίτερα απαιτητική από υπολογιστική σκοπιά, ιδιαίτερα όσο η διάσταση του μεταβολικού μοντέλου αυξάνει. Για τον σκοπό αυτό αναπτύχθηκε η βιβλιοθήκη **dingo** η οποία κάνει χρήση ενός πολυφασικού αλγορίθμου **Monte Carlo**, Κεφάλαιο 4.

Τέλος, αναλύθηκαν δείγματα ιζήματος από το έλος Τριστόμου Καρπάθου, καθώς επίσης δείγματα από μικροβιακούς τάπητες **mat** και από μικροβιακά συσσωματώματα (**aggregates**). Για τον σκοπό αυτό, χρησιμοποιήθηκε τόσο η μέθοδος μετακωδικοποίησης με γονίδιο-δείκτη το 16S όσο και η μέθοδος μεταγονιδιωματικής **shotgun**. Επίσης μετρήθηκαν περιβαλλοντικές παράμετροι (όπως αλατότητα, συγκέντρωση οξυγόνου). Βασικές λειτουργίες που υποστηρίζουν τη ζωή σε τέτοιες συνθήκες εντοπίστηκαν ενώ ακόμη γονιδιώματα τάξων που εντοπίζονται για πρώτη φορά ανασκευαάστηκαν από τις αλληλουχίες του μεταγονιδιώματος (MAGs), Κεφάλαιο 5.

Όπως συμβαίνει και στις μικροβιακές κοινότητες, οι βιοπληροφορικές μέθοδοι σπάνια στέκουν απομονωμένες, αντίθετα τείνουν να συγκροτούν κι αυτές τις δικές τους ¨κοινότητες¨. Οι μέθοδοι που αναπτύχθηκαν στα πλαίσια αυτής της διατριβής επιδιώκουν να συγκροτήσουν ένα πλαίσιο μελέτης από το επίπεδο της κοινότητας σε αυτό του είδους και από εκεί, πίσω πάλι στην κοινότητα. Ένα τέτοιο πλαίσιο αναλύεται για την μελέτη μικροβιακών αλληλεπιδράσεων.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

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

Symbols

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

Chapter 1

Introduction

1.1 Microbial communities: composition , functions & interactions

1.1.1 Microbial diversity: life under extraordinary conditions

Microbes are considered to be omnipresent in the various ecosystems on Earth [Falkowski et al., 2008]. It was only until recently (2019), when Belilla et al. discovered for the first time a place on Earth where no microbial forms of life are present. Extremely low pH, high salt and high temperature had to be at the same place at the same time to stop microbes from "conquering" them. However, microbes are not just abundant but exceedingly variant too. Locey and Lennon using a unified scaling law and a log-normal model of biodiversity, estimated microbial diversity at about 1 trillion species [Locey and Lennon, 2016]. However, despite the extensive studies of the scientific community, less than 1% of the microbial species on Earth have been identified [ism].

Microbes are distinguished by multiple properties. Based on their morphology microbes can be spherical (cocci), rod-shaped (bacilli), arc-shaped (vibrio), and spiral (spirochete) [Dunlap, 2001]. Based on their metabolic characteristics, microbes are further distinguished. More specifically, according to their *energy source*, a microbe can either oxidate inorganic compounds (**chemotrophs**) or sunlight (**phototrophs**). Similarly, microbes can use CO_2 (**autotrophs**) as their *carbon source*, or organic compounds (**heterotrophs**) or both (**mixotrophs**). Finally, based on their *electron source*, microbes are distinguished between those using inorganic compounds (**lithotrophs**) and those using organic compounds (**organotrophs**) [Madigan et al., 2018]. Microbial taxa combine combining alternatives of the aforementioned categories shape a range of microbial profile of all the possible combinations; for example **chemolithoautotrophic** bacteria, e.g. nitrifying and sulfur-oxidizing bacteria, as well as **photoautotrophic** bacteria, e.g. purple bacteria and Green sulfur bacteria. Finally, microbial taxa can also be distinguished by their various ecological distributions and activities, and by their distinct genomic structure, expression, and evolution [Dunlap, 2001].

1.1.2 Functional diversity: shaping the conditions of life

However, it is not only the number of microbial taxa and their massive biomass that make the study of microbial communities essential; it is mostly their functional potentials. Life on Earth would not be as we know it, if existed at all, if it was not for the microbes and their long contribution on ensuring life-supporting conditions. Nevertheless, these are the *biological machines responsible for planetary biogeochemical cycles* [Falkowski et al., 2008]; meaning that biogeochemical cycling to a global extent is powered by the metabolic processes of the microbial taxa [Louca et al., 2016]. In Figure 1.1 the contribution of microbial communities in the cycle of CO_2 is shown.

The biological fluxes of most of the major elements (i.e., carbon, hydrogen, oxygen, nitrogen and sulfur) required for any biological macro-molecule, are driven largely by microbially catalyzed, thermodynamically constrained redox reactions [Falkowski et al., 2008]. Phosphorus the last of the 6 fundamental elements for life, is also included in the metabolic pathways catalyzed by microbes. Thus, microbial communities consist of hundreds or even thousands of metabolically diverse strains and species [Leventhal et al., 2018], and their functions and determine the fitness of most organisms on Earth. In case of human health, specific microbial enzymatic pathways and molecules necessary for health promotion have been well known. Some of these "beneficial factors" are already known for probiotics and species in the human microbiome [Marco, 2021].

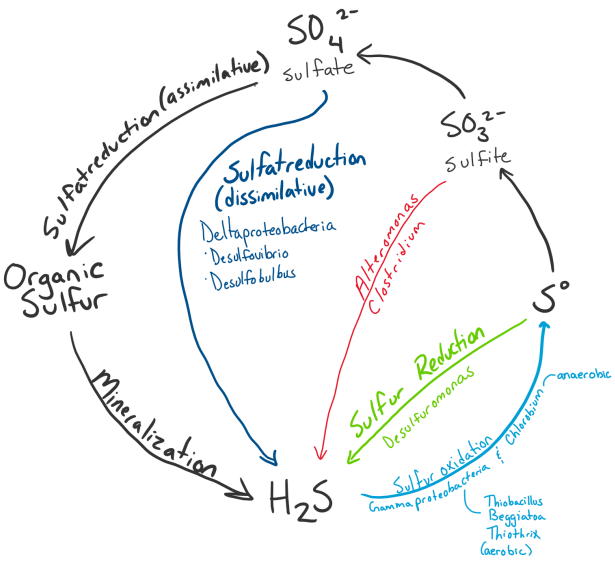
The relationship between the taxonomic and the functional profile of a microbial community has been an open question for scientists; is the *who* or the *what* more important to distinguish communities [Xu et al., 2014]? And how does each of these profiles respond to the various perturbations of an environment; Do they tend to converge [Estrela et al., 2022]? Do perturbations of the taxonomic composition of a community influence the robustness of the community's functional profile [Eng and Borenstein, 2018]? divergence of each under and from an evolutionary point-of-view. Does it matter *who* is doing *what* and how does this affect the niche of a species [Louca et al., 2018]? And what about the rare taxa and their corresponding functions in an assemblage [Chen et al., 2020a, Jousset et al., 2017]?

Microbial Ecology focuses on the study of the following interactions:

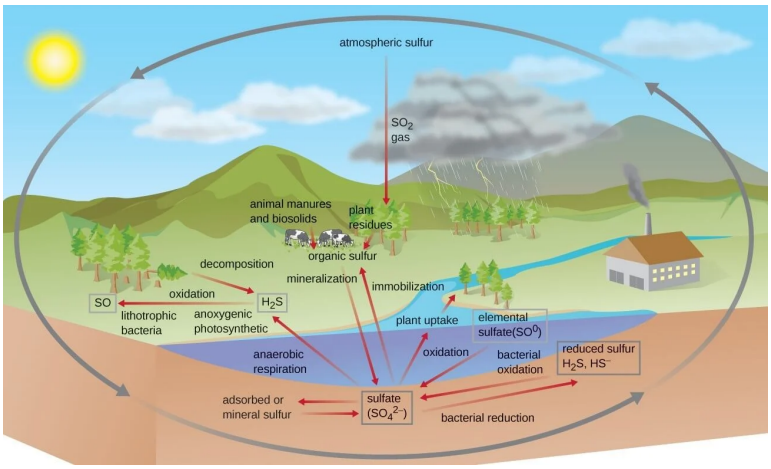
- those between microbial taxa and their environment
- those among the various microbial taxa present in a community, and
- those between microbial taxa and their host [ism]

Microbial ecologists also investigate the role of microbial taxa in biogeochemical cycles [Falkowski et al., 2008] and their interaction with anthropogenic effects e.g. pollution and climate change [Cavicchioli et al., 2019].

Even though HTS has allowed a massive extension of our knowledge in specific enzymatic reactions that regulate these pathways the rules that determine the assembly, function, and evolution of these microbial communities remain unclear. Thus, both in case of environmental and human the underlying mechanisms for how microbial assemblages work and affect their environment, remain to be discovered. Understanding



(A) Basic reactions in sulfur cycle



(B) Sulfur cycle reactions per environmnetal type

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

the underlying governing principles is central to microbial ecology [Giri et al., 2021] and crucial for designing microbial consortia for biotechnological [Giri et al., 2020] or medical applications [Kong et al., 2018].

Studies such as the one of Louca et al. have opened new frontiers in our understanding on microbial assemblages. After building metabolic functional groups and assigning more than 30,000 marine species to these groups, Louca et al. showed that the distribution of these functional groups were influenced by environmental conditions to a great extent, shaping *metabolic niches*. At the same time though, the taxonomic composition within

individual functional groups were not affected by such environmental conditions [Louca et al., 2016].

1.1.3 Ecological interactions in microbial communities

Moreover, to elucidate how these assemblages work the biotic interactions have to be considered too. Microbial interactions play a fundamental role in deciphering the underlying mechanisms that govern ecosystem functioning [Braga et al., 2016, Faust and Raes, 2012]. Microbes secrete costly metabolites (called **byproducts**) to their environment, which other microbes can absorb and exploit [Pacheco et al., 2019]. By exchanging metabolic products, mostly as there are also other ways of interactions e.g. quorum sensing, microbial taxa establish various interactions.

The interaction between two taxa can either be neutral or positive / negative (Figure 1.2). In case of a positive interaction, there is a case where both taxa benefit one from another. This *win-win* relationship is called **mutualism** (or "cooperation") and it can be a result of *cross-feeding*, in which two species exchange metabolic products [Faust and Raes, 2012]. Such is the case in biofilms where multiple bacterial taxa are working together building a structure that provides them antibiotic resistance [Santos-Lopez et al., 2019]. There is also the case where only one of the two taxa benefits without helping or harming the other; this interaction is called **commensalism** [Faust and Raes, 2012]. For example, *Nitrosomonas* oxidize ammonia (NH_3) into nitrite (NO_2^-), so *Nitrobacter* can use it to obtain energy and oxidize it into nitrate (NO_3^-) [Laanbroek et al., 2002]. Such interactions are quite common in microbial communities.

In case of a negative interaction, can harm each other either way (**competition**). That is the case between *Listeria monocytogenes* and *Lactococcus lactis* in the study of Freilich et al. where their resource competition is high enough contributing to their non-overlapping existence [Freilich et al., 2010]. Moreover, similarly to commensalism, there is also the case when a taxon has a negative affect on the other without getting any harm (**amensalism**). Such is the case for *Acidithiobacillus thiooxidant* that produces sulfuric acid (H_2SO_4) by oxidation of sulfur [Bobadilla Fazzini et al., 2013] which is responsible for lowering of pH in the culture media which inhibits the growth of most other bacteria [Jin and Kirk, 2018]. Finally, one of the taxa may have a positive affect (host) on the other, but the latter (parasite) can be harmful to its benefactor (**parasitism**) [Faust and Raes, 2012]. There are multiple cases of parasitism in real-world communities; species of the genus *Bdellovibrio* for example, are parasites of other (gram-negative) bacteria [Stolp, 1979].

However, we have a very limited understanding of such interactions and the ways that are combined to rule community-level behaviors. Thus, we cannot predict community responses to perturbations (community stability) [Venturelli et al., 2018]. Over the years, various methods have been used to infer such interactions. co-occurrence modells [Faust and Raes, 2012], time series data and causal models [Mainali et al., 2019], through metabolic interactions as proxy [Levy and Borenstein, 2012]. Dynamic models, such Ordinary Differential Equations (ODEs) and the Generalized Lotka–Volterra (gLV) model [Gonze et al., 2018] have been also widely used. Finally, in recent years, metabolic networks and constraint-based models have been also used to predict microbial interactions [Heinken et al., 2021, Dukovski et al., 2021]. This last approach allows predictions

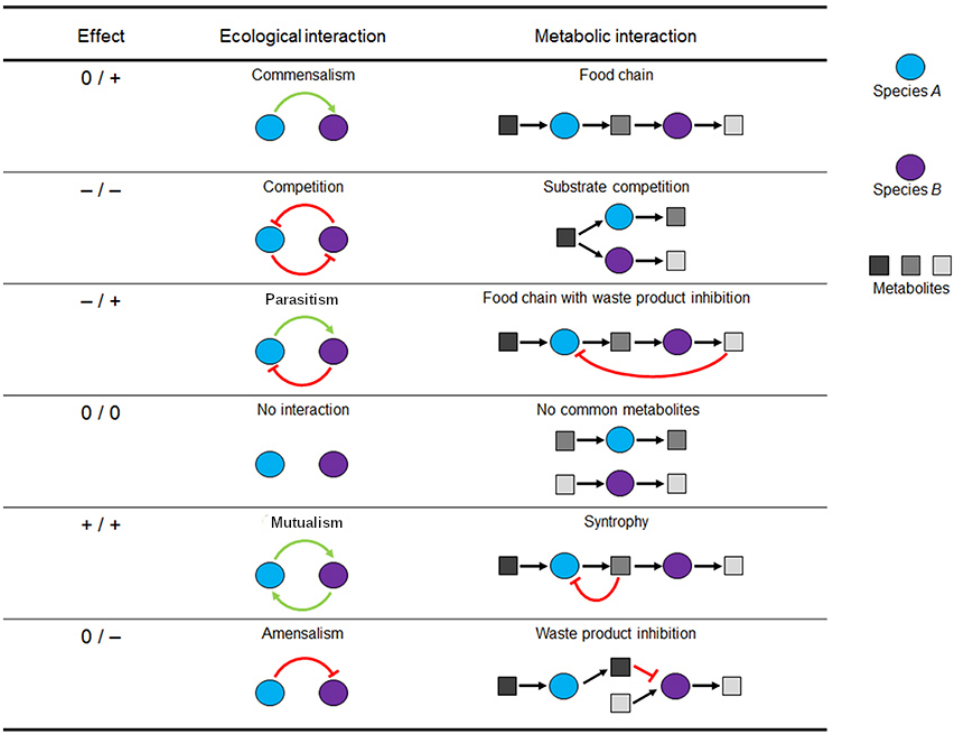


FIGURE 1.2: Microbial interaction types along with their corresponding metabolic ones. Due to certain metabolic interactions, two taxa may have a positive, a negative or a neutral effect one another. Figure based on [Perez-Garcia et al., 2016]

for the metabolic dynamics of the community as well as of the exact set of compounds the taxa of the community exchange [Levy and Borenstein, 2012]. Still though, microbial interactions inference is a challenging task and several questions are still open.

Apparently, the environmental conditions affect the ecological interactions to a great extent. A pair of taxa may be competitors in one case but have a neutral interaction in another one. In addition, evolutionary processes may change certain interactions; for example moving from commensalism to parasitism [Parmentier et al., 2016]. Both ecological and environmental interactions play a part in the composition and the functional potential of microbial assemblages. On top of that, pairwise microbial interactions can be modified by a third organism, leading to higher-order effects that influence community behaviors [Bairey et al., 2016].

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

For decades, *reductionism* has been the main conceptual approach in biological research [Noble, 2008]. Traditionally, for studies relating genetics and ecology scientists first identify an ecological adaptive phenotype and then they try to detect causal genetic variation [Noble, 2008]. However, as described in the previous sections, HTS data have turned the page in Biology research in numerous ways. Therefore, it is nowdays possible to *reverse* this framework and by using the genomic information retrieved, to study the

ecology of a species. The **Reverse Ecology** framework uses advances in both systems biology and genomic metabolic modeling to implement community ecology studies with no a priori assumptions about the organisms under consideration [Cao et al., 2016]. Therefore, Reverse Ecology attempts to interpret HTS (genomic) data as large-scale ecological data [Levy and Borenstein, 2012].

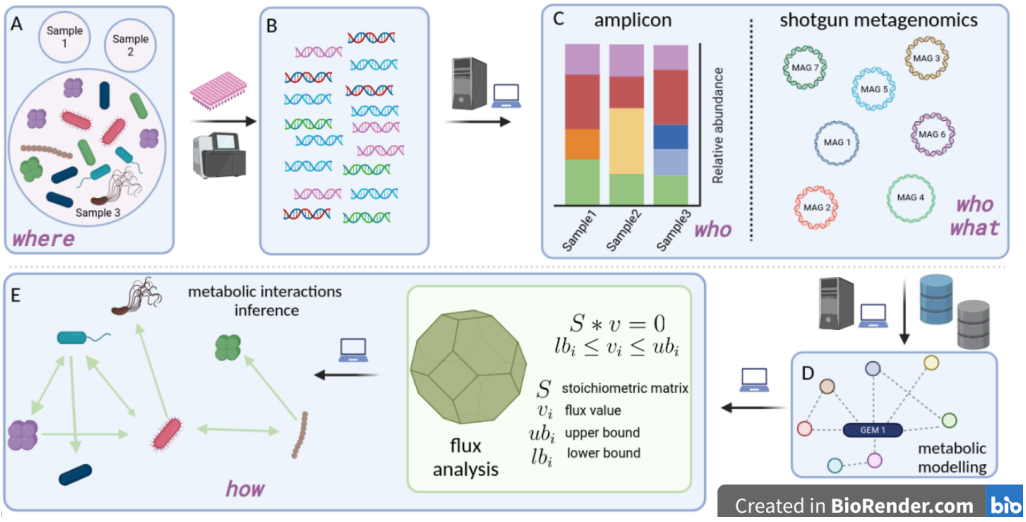


FIGURE 1.3: Without any previous knowledge of the species present in a community (A) and using HTS data (B) one can have an overview of the species present as well as in the functional profile of the community (C). Especially when the complete genome of a species has been retained (either using metagenomics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

As shown in Figure 1.3, the Reverse Ecology framework has multiple alternatives and various methods can exploit this concept. The analysis of metabolic networks (see Section 4) plays a great part in several Reverse Ecology approaches. Most parts of this dissertation have been influenced by this, especially chapters 3, 4 and 5.

1.2 High Throughput Sequencing in Microbial Ecology

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

To discover the microbial taxa present in a sample, scientists have explored multiple ways through the years. Only a particularly limited proportion of the microbial species can be cultured [Steen et al., 2019]. Therefore, mono-cultures and enrichment cultures allow us to observe only a small fraction of the actual diversity. As a consequence, other methods for the taxonomic identification of theses species are required. Based on molecular

characteristics of the microbial taxa, over the last decades, a series of methods have been developed.

Moving from single species to assemblages, molecular-based identification and functional profiling of communities has become available through marker (metabarcoding), genome (metagenomics), or transcriptome (metatranscriptomics) sequencing from environmental samples [Goldford et al., 2018]. To a great extent, these methods address the problem of how to produce and get access to the information on different biological systems and molecules.

In case that the taxonomic assessment of a sample is the aim of a study, *metabarcoding* (amplicon-targeted metagenomics) and *shotgun metagenomics* can be used as alternative options. Metabarcoding studies are common, well-established, cheaper and less computationally demanding than shotgun metagenomics [Bell et al., 2021a]. Its primary drawbacks are the limited information present in the short barcoding sequence and the possible taxonomic bias arising from differential efficiency of PCR primer pairing in different species [Blazewicz et al., 2013]. On the other hand, shotgun metagenomics offers a better taxonomic resolution at the species level by obtaining information from random sampling of virtually all genomic regions, and can address microbiome metabolic functions and entire biochemical pathways [Sharpton, 2014]. Unfortunately, it requires higher sequencing coverage and, consequently, more complex and demanding downstream bioinformatics analysis [Laudadio et al., 2018]. Nevertheless, it has recently been suggested that shotgun metagenomics provides a deeper characterisation of microbiome complexity that metabarcoding recently enabling to profile up to the level of strains, whose non-core genome is responsible for crucial functional differences within the same species, as the fundamental units of the community [Dávila-Ramos et al., 2019, Clooney et al., 2016, Segata, 2018].

Targeting community composition and functional profiles in several ecological niches, microbial ecologists produce vast amount of sequencing data [Harrison et al., 2021]. These approaches enable the study of ecosystems with no prior knowledge of the resident species, while at the same time a number of challenges for their management and bioinformatics analysis is rising.

1.2.2 Bioinformatics challenges in the analysis & management of HTS data

Moving from raw data to taxonomic and functional profiles of a microbial community comes with high computational costs, especially in the case of metagenome studies [Yang et al., 2021]. Sequence pre-processing, assembly, classification, and functional annotation consist of several steps the most of which a significant number of algorithms or/and software tools are available [Breitwieser et al., 2019, Roumpeka et al., 2017]. Tailoring each tool's execution parameters to reflect each experiment's idiosyncrasy is vital for legitimate findings, yet it makes analyses of metagenomics data even more complex.

In addition, there are several challenges on the bioinformatics analysis *per se*. *Taxonomy assignment* in both amplicon- and shotgun metagenomics studies has several issues to meet [Simon et al., 2019]; the taxonomy of microbes is a challenge per on its own [Parks et al., 2020].

In amplicon studies, among the most major issues is the one of the abundances of the taxa found [Fonseca, 2018, Bálint et al., 2016] as well as the presence of pseudo-genes [Song et al., 2008]. In the first case, issues such as the usually unknown number of marker gene copies per cell in the various taxa, PCR - related biases such as primer-template mismatches, length difference of amplicon, artificial base changes, chimeric molecules and library preparation - related issues such as chimera formation by the mix of amplicons from different samples makes hard for the method to have robust quantitative results [Bálint et al., 2016]. Reads on the other hand resulting from pseudo-genes or/and highly divergent nuclear mitochondrial pseudo-genes (NUMTS), nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms [Bensasson et al., 2001], can lead either to false positive taxonomic hits or to non-hits at all, adding extra noise to the amplicon results returned.

In shotgun metagenomics studies there are also several challenges. Meta-genome *assembly* comes with a great number of challenges. Due to the uneven (and unknown) representation of the different organisms within a metagenomic mixture, simple coverage statistics can no longer be used to detect the repeats, while unrelated genomes may contain nearly-identical DNA (inter-genomic repeats) representing, for example, mobile genetic elements [Ghurye et al., 2016]. At the same time, *binning* is a rather tricky step too; several algorithms have been developed to address it [Yue et al., 2020] while approaches combining the output of individual algorithms have been introduced too [Song and Thomas, 2017].

The vast amounts of data that come with metagenomic studies and the computational complexity for implementing multiple steps mentioned earlier imply immense computational requirements for their analysis that usually exceed the capacity of a standard personal computer [Merelli et al., 2014].

For HTS data to be available to the scientific community for further exploitation, it is required to be accompanied by comprehensive metadata [Vangay et al., 2021]. The potential of HTS data is revealed when they are available to the community; this way studies that could never been performed by individual researchers, labs or institutes are now possible. This way, a single researcher can now investigate how a certain environmental type reacts in response to an environmental variable by making use of hundreds of metagenomic samples that fulfill the criteria of his/her study. Finding data of interest however, can be particularly difficult. This is so because of a combination of reasons. HTS data can be particularly heterogeneous based on both the data generation and the data processing methods used. However, it is mostly the vague or even absent metadata accompanying the HTS data set several limitations in their re-usage [Hu et al., 2022].

The concept of FAIR data (Findability, Accessibility, Interoperability & Reuse) and the FAIR principles[1] along with community - driven standards and resources such as the Genomic Standards Consortium (GSC)[2], the Minimal Information about any Sequence (MIxS) [Yilmaz et al., 2011b,a] and the National Microbiome Data Collaborative

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

(NMDC)[3] [Wood-Charlson et al., 2020] aim to address these challenges [Wilkinson et al., 2016].

1.3 Data integration in the service of microbial ecology

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

Over the last decades, based on computational and mathematical analysis and modeling, and by exploiting interdisciplinary data and knowledge, Systems Biology focuses on complex interactions within biological systems [Tavassoly et al., 2018]. The more data becoming available from all the different levels of hierarchy of life, the more feasible for scientists to move from reductionism to more holistic approaches for interpreting how the properties of a system emerge [Noble, 2008].

Microbial ecology as a field would have not been the same if it was not for resources such as Integrated Microbial Genomes (IMG) and GOLD [Chen et al., 2021], SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST) [Overbeek et al., 2014], Pathosystems Resource Integration Center (PATRIC), [Zhulin, 2015] and many more that thousands of researches use in their every day work. All these approaches, regardless on what they focus, they are all based on data aggregation and data integration approaches. *Data aggregation* denotes the gathering of data from diverse sources in a certain scheme that will allow them to be used as a combined data-set for further analysis [Simpson et al., 2010]. In case of microbial ecology, that means that data focusing on the genetic information can be combined with phenotypical data or even with environmental and ecological data. *Data integration* on the other hand, is the process of combining everything retrieved on the data aggregation step, to get a summarization and unified view of all the accumulated data [Schneider and Jimenez, 2012]. Such summarizations may lead researchers to new hypotheses that in turn, will be tested through new experiments (Figure 1.4).

Data integration comes with great challenges. Apparently, data integration methods are based on the existence of primary databases. Each of these database resources come with its own assumptions and schemas. Therefore, it is not a straight-forward task to recognize or assign and maintain the correct names of biological entities across the various databases [Stein, 2003]. Taxonomy is quite an indicative example. As there is no a global taxonomy system, even the species name can be a great challenge in such approaches; how to retrieve information about a species that does not have the same name on the various databases to integrate? Therefore, retrieving and mapping entities can be rather complex. Similarly to taxonomy, most biological databases are constantly changing. Thus, integration approaches need to be periodically so the always keep updated [Stein, 2003]. In addition to the heterogeneity of the data *per se*, further challenges that make data integration even harder in case of biological data, is the lack of unique standards [Triplet and Butler, 2011]. In the case of HTS data, great efforts to address this challenge have been made (see Section 1.2.2).

[3] https://microbiomedata.org/metadata/

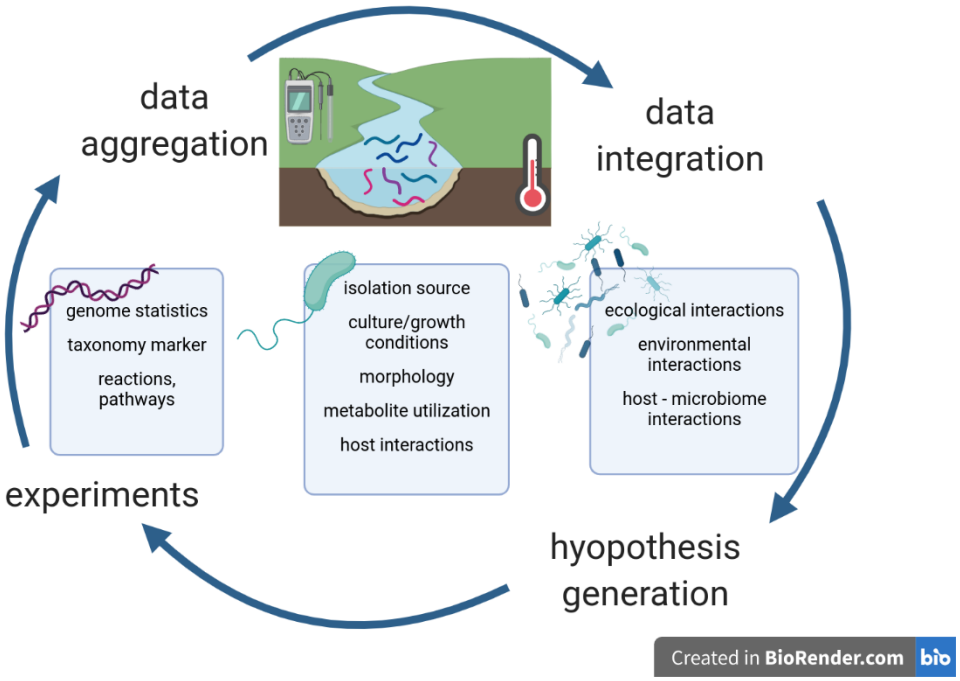


FIGURE 1.4: A data integration scheme for microbial ecology oriented data. Measurements from experiments at every level of organization of life are gathered and their summary provides researchers with new insight. Created with BioRender.com.

One of the most typical examples of data integration and its potential is the STRING database[4], where multiple channels of information are combined to retrieve protein - protein interactions [Mering et al., 2003, Szklarczyk et al., 2021]. In addition to databases of interaction experiments and others of interaction predictions, text-mining methods of the scientific literature enhance further the PPI predictions [Szklarczyk et al., 2021]. Focusing on bacterial information, BacDive[5] [Reimer et al., 2019] is a great example - resource of the added value that data integration methods can provide.

Multiple integration approaches attempt to address the challenges described. The *data warehousing* approach is a widely used data integration approach and has two mains steps; first, a unified data model that can accommodate all types of information from the various source databases is schemed. Then, software is developed aiming at gathering the data from the source databases, convert them to match the unified data model and then load them into the warehouse [Stein, 2003]. Once these steps have been completed, further analysis of the once several bits of information - now a single data-set, can be performed. New insight may come up either from statistical analyses on the unified data-set or from their visualization [Leonelli, 2013].

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

1.3.2 Ontologies & metadata standards: cornerstones for efficient data integration

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

According to Stevens et al. an *ontology* is the *"concrete form of a concepcualisation of a community's knowledge of a domain"* [Stevens et al., 2000]. Ontologies attempt to capture the main concepts in a *knowledge domain*, i.e. a body of knowledge that is often associated with a specialized scientific discipline. For example, considering *where* a species live or *where* a process occurs, one need to describe the environment where the phenomenon under study takes place. Thus, the Environment Ontology (ENVO)[6] aims to provide descriptions of environments [Buttigieg et al., 2016]. Using sets of entities, meaning entities sharing several attributes (*concepts*), descriptions of the interactions between concepts (*relations*), entities - members of a concept (*instances*) and properties of relations that aim to constrain the value a class or an instance may get (*axioms*) aim to create an agreed vocabulary and semantic structure for exchanging information about that domain [Stevens et al., 2000]. A *vocabulary* includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the [Uschold et al., 1998]. Ontologies are fundamental for data integration as they ensure that the knowledge included in a text or in a data set, can be captured by both humans and computers.

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

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

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

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

As mentioned in Section 1.2.2, great efforts have been made on setting HTS - related metadata standards [Yilmaz et al., 2011b,a, Wood-Charlson et al., 2020]. That is so because comprehensive metadata is the only way to ensure:

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

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

Regarding the environmental metadata of a sample, ENVO [Buttigieg et al., 2016] and MIxS [Yilmaz et al., 2011b] are working together to build a solid framework [EnvironmentOntology, 2021]. The *broad-scale environmental context* value is representing the major environmental system a sample came from; thus, *biome*[7] ENVO terms should be used as values. An ENVO *biome* term represents an ecosystem to which resident ecological communities have evolved adaptations. The *local environmental context* value, stands for entities which are in a sample's local vicinity and may have significant causal influences on the sample; ENVO *feature* terms may be used for that. Finally, as values of the *environmental medium* category, environmental *material*[8] (one or more) immediately surrounded the sample prior to sampling. However, other resources use different schemes for describing the environment of a sample. For example in the GOLD database [Mukherjee et al., 2021], a five-level ecosystem classification path that includes Ecosystem, Ecosystem Category, Ecosystem Type, Ecosystem Subtype and Specific Ecosystem has been adopted.

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

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

[7] http://purl.obolibrary.org/obo/ENVO_00000428
[8] http://purl.obolibrary.org/obo/ENVO_00010483

methods is shown as the the ENVO term denoting *oligotrophic water* should be provided as the *feature* and *Deep Chlorophyll Maximum* should be used to describe the *material*.

Sample metadata [-]

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

(A) Poor, non machine readable metadata

Sample metadata [-]

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

(B) Rich, partially machine readable metadata

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

Challenges associated with metadata deposition as the one described above, mean submitters: may lack of training and outreach resulting or they do not fully realise the importance of metadata and how to comply with standards. On top of that, the non-existence of standards in many cases or the use of more than one standards lead to extra complexity. Only by a concerted effort on the part of the database providers, and with the encouragement and support of the research community, will we be able to tame the explosion of biological data [Stein, 2003].

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

1.4.1 Constraint-based modeling for the analysis of metabolic networks

The relationship between genotype and phenotype is fundamental allowing to elucidate mechanisms that govern the physiology of a species as well as those ruling at the community level [Morris et al., 2020]. Meta- bolism penetrates most of the different levels of living entities horizontally [Schramski et al., 2015] and while it reflects the genomic information it indicates what is actually going on on a cell at a certain time as a response to genetic or environmental changes [Lima et al., 2021]. One can use the *Reverse Ecology* framework (Section 1.1.4) to move all the way from genomic information to metabolism and the environment and back. To this end, *metabolic networks* and their analysis are essential. The vast number of reaction taking place in a cell are interlinked (the product of the first acts as the substrate for the next) building up metabolic pathways, while their stoichiometry allows their mathematical representation. The rate of turnover of molecules through a metabolic reaction is called *flux*. The metabolic network of a species consists of the sum of all the reactions that take place in its cell, while *metabolic model* is its representation in a mathematical format (Figure 1.6)[9]. We call *Genome-scale metabolic models (GEMs)* incorporate the vast majority of the processes that occur in a cell or an organism in a mathematical format [Feist et al., 2009].

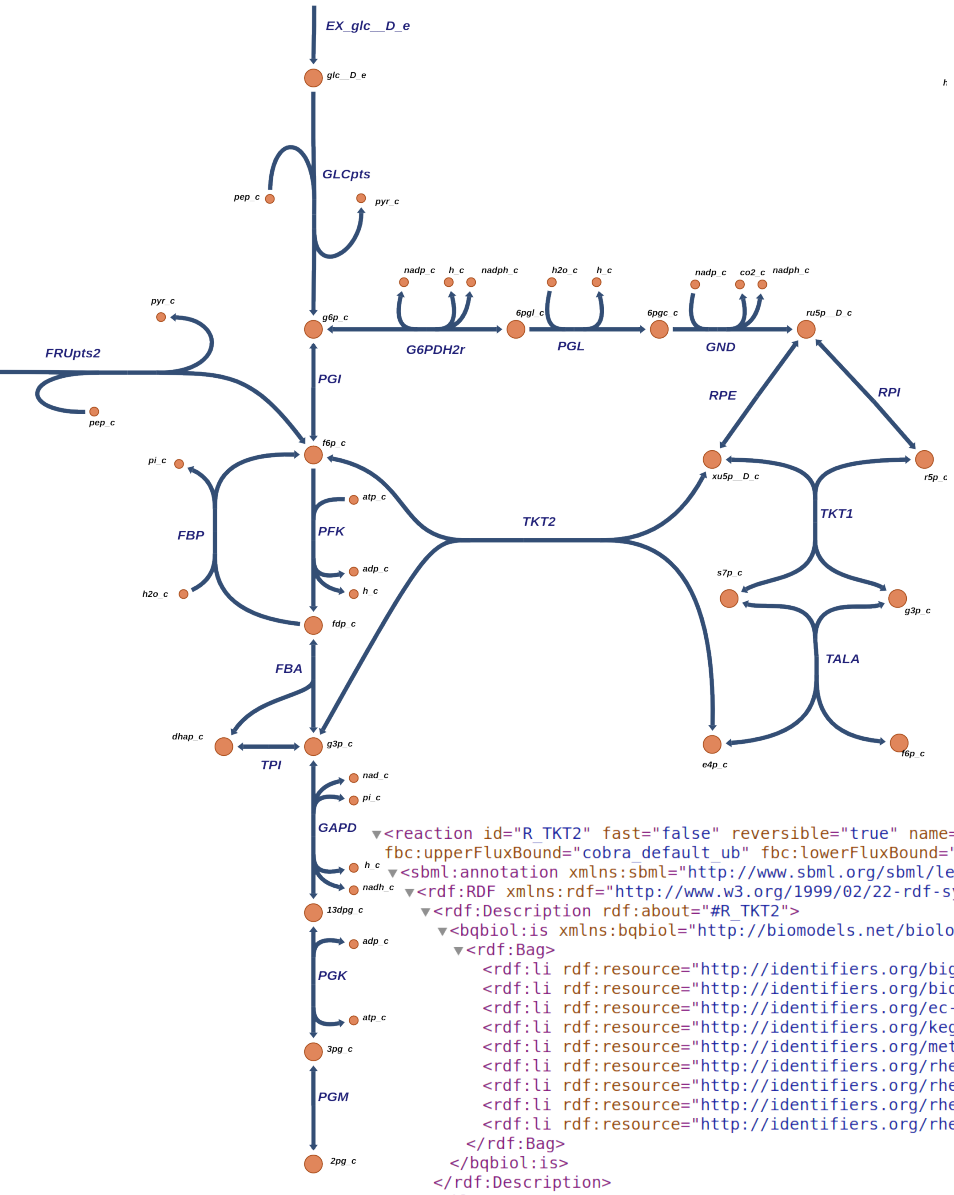
Once the complete genome is retrieved the enzymes and thus the potentially catalyzed by the organism reactions can be listed. However, the reconstruction of a GEM is not a straight forward task and the more the complexity of the species increases, the more effort is required for this task [Thiele and Palsson, 2010]. Thermodynamics, metabolome, physiological and labelling data as well as literature can be also integrated in such models [Saldida et al., 2020].

The analysis of GEMs has been interwoven with constraint-based modeling approachess [Lewis et al., 2012]. As all compounds are finite the concentration of each metabolite is bounded [Palsson, 2015], meaning that the models derived from the metabolic networks have constraints. Likewise, as the laws of thermodynamics need to apply in such systems, the flux of each reaction is also bounded. Therefore, the flux value of each reaction is constraint too. We call *steady state* the condition where the production rate of each metabolite equals its consumption rate [Cakmak et al., 2012]. Equation 1.1 represents the main concept of constraint-based modeling at a steady state.

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

where S is a $m * n$ table (m being the number of metabolites and n the number of reactions of the model) that stands for the *stoichiometric matrix* of the model. The columns of S consists of the *stoichiometric coefficients*, i.e. the number of molecules a biochemical reaction consumes and produces, of the model's reactions. $v \in \mathbb{R}^n$ is the *flux vector*

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




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

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

that contains the fluxes of each chemical reaction of the network. As all the fluxes are bounded, for each coordinate v_i of the vector v, there are constants $v_{ub,i}$ and $v_{lb,i}$ such that $v_{lb,i} \leq v_i \leq v_{ub,i}$, for $i \in [n]$, where n is the number of reactions of the model. The *solution space* of systems such as the one of equation 1.1 "live" in a ***polytope***. Further introductory material on computational geometry can be found at Appendix B.

As discussed in [Reed, 2012] a great range of constraints govern the cells' operations; for a thorough overview on the constraints cells operate under, you may see [Palsson, 2015], Chapter 16.5. (Bio)physico-chemical- (e.g., thermodynamics, nutrient uptake, oxygen availability etc.) as well as connectivity-, capacity- and rates-related constraints are applied on the functions of such a network.metabolic Each of the aforementioned constraint categories include multiple constraints, such as thermodynamics- and gene-expression-oriented constraints that add extra complexity in the model. The more constraints a model incoroporates, the more accurate the flux distributions it returns.

Using constraint-based modeling scientists can predict not only potential interactions, topology-based metabolic models are adequate for this task, but also specific metabolic dynamics in a community [Levy and Borenstein, 2012]. The most commonly used constraint-based methods for the analysis of metabolic networks are ***Flux Balance Analysis (FBA)*** [Orth et al., 2010] and *Flux Variability Analysis (FVA)* [Gudmundsson and Thiele, 2010]. Both have been used to a great number of studies, providing fundamental isngight [Shastri and Morgan, 2005, Chapman et al., 2015]. Models estimate the minimum or the maximum of a specific (linear) ***objective function*** over the polytope. It is common for the *biomass function* of an organism to be used as the objective function. The biomass function aims at representing all metabolites needed for a cell or an organism to double. In this setting the optimization of the biomass function is like optimizing the growth of the organism itself [Feist and Palsson, 2010]. On top of that, *dynamic FBA* approaches have tried to to study the transience of metabolism due to metabolic reprogramming [Mahadevan et al., 2002].

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

As mentioned, constraint-based approaches cover a great range of methods [Lewis et al., 2012]. FBA has been proved particularly useful however, it is a *biased* method due to the selection of the objective function. To study the global features of a metabolic network *unbiased methods* are required. On top of that, FBA is a method that addresses the question of what is the minimum or the maximum of a specific objective function, by identifying only a single optimal flux distribution. However, by construction, there is an infinite number of optimal steady states lie on a certain face of the polytope – which is also a polytope. In addition, there is no guarantee that the system under study would select the optimal steady state that FBA computes.

Using uniformly distributed steady states one could estimate the probability distribution for the flux of any reaction [Herrmann et al., 2019], which can lead to a deep statistical analysis of the metabolic network.

To overcome these obstacles, we sample uniformly from the set of optimal steady states and we express and quantify our uncertainty about each flux by estimating the univariate marginal probability densities [Schellenberger and Palsson, 2009]. Each prob-

ability density corresponds to a reaction flux. With this information at hand we can compute credible confidence intervals, estimate the average flux value, or employ other statistical methods. This procedure relies on collecting, that is sampling, a sufficient number of uniformly distributed points in the interior of the corresponding polytope.

To obtain an accurate picture of the whole solution space, once more, we sample uniformly distributed points. This way instead of a single and optimal solution, the distribution of each each reaction's flux is returned (Figure 1.7). This way, we can now investigate the properties of certain components of the whole network that potentially can lead to biological insights [Palsson, 2015].

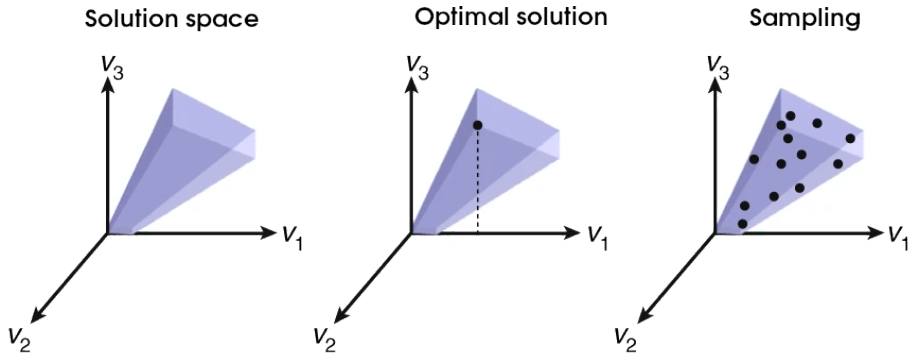


FIGURE 1.7: A visual comparison of the insight FBA ("Optimal solution") and flux sampling ("Sampling") return

Flux sampling has been proved rather valuable for a great range of applications; from design experiments and studying enzymopathies [Price et al., 2004] to the study of metabolism under changing environmental conditions [Herrmann et al., 2019] and the discovery of strain-dependent differences that affect the aroma production in wine yeasts [Scott et al., 2021].

Implementations of Markov Chain Monte Carlo algorithms such as Hit-and-Run (HR) [Smith, 1984], the Artificial Centering Hit-and-Run (ACHR) [Kaufman and Smith, 1998] and Coordinate Hit-and-Run with Rounding (CHRR) [Haraldsdóttir et al., 2017] have been adopted and used to a great extent. Similarly to FBA, flux sampling algorithms, e.g. CHRR, have been integrated in cobra [Heirendt et al., 2019], the most widely used software package for metabolic network analysis.

On top of that, over the last few years, flux sampling has been used in computational approaches for inferring microbial interactions The Computation Of Microbial Ecosystems in Time and Space (COMETS) project and software [Dukovski et al., 2021] first focused on the interactions between a single species and its environment. Nowdays, metabolic modeling moves to the community level. Approaches such as the one of Diener et al. in their MICOM software [Diener et al., 2020], set a new era for the study of microbial ecology.

However, flux sampling is rather challenging from the computational point of view. The "dimensionality curse" is not a problem to a small GEM such as those of single bacterial taxa. However, to more complex species and especially in the case of com-

munity modeling the dimesion of the derived polytope can be notably high. Moreover, polytopes derived from metabolic networks are usually rather skinny, partially due to the great range the various flux values may get, making mixing hard and adding extra complexity [Haraldsdóttir et al., 2017, Schellenberger and Palsson, 2009].

1.5 Aims and objectives

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

The aim of this PhD was double:

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

2. to exploit state-of-the-art methods to identify taxa and functions that play a key part in microbial community assemblages in hypersaline sediments.

All parts of this work are purely computational. Both samples and their corresponding sequencing data used in Chapter 5 have been collected and produced by Dr. Christina Pavloudi [10].

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

In **Chapter 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 metabolic processes for ensuring life at such an extreme environment were identified.

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

Microbial interactions of the assemblages retrieved were also studied by exploiting data integration and reverse ecology approaches.

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 **Chapter 7**, 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, bioinformatics pipelines need to combine state-of-the art technologies and

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

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

Findings: PEMA is a containerized assembly of key metabarcoding analysis tools that requires low effort in setting up, running, and customizing to researchers' needs. Based on third-party tools, PEMA performs read pre-processing, (molecular) operational taxonomic unit clustering, amplicon sequence variant inference, and taxonomy assignment for 16S and 18S ribosomal RNA, as well as ITS and COI marker gene data. Owing to its simplified parameterization and checkpoint support, PEMA allows users to explore alternative algorithms for specific steps of the pipeline without the need of a complete re-execution. PEMA was evaluated against both mock communities and previously published data sets 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 [Pavan-Kumar et al., 2015]. eDNA refers to genetic material obtained directly from environmental samples (soil, sediment, water, etc.) without any obvious signs of biological source material [Thomsen and Willerslev, 2015]. 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 [Ji et al., 2013]. eDNA metabarcoding attempts to turn the page on the way biodiversity is perceived and monitored [Ji et al., 2013]. 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 [Schloss et al., 2009], QIIME 2 [Bolyen et al., 2018], LotuS [Hildebrand et al., 2014]). However, certain limitations accompany each of these and occasionally they can be far from easy-to-use software. Moreover, there is a great need for similarly straightforward and benchmarked approaches for the analysis of other marker genes. With respect to the COI and ITS marker genes, a number of pipelines have been implemented, e.g., Barque [2], ScreenForBio [Axtner

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

et al., 2019], and PIPITS [Gweon et al., 2015]. 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 [3]) to serve the broad biological community via commonly accessible infrastructures.

2.1.3 Contribution

PEMA (Pipeline for Environmental DNA Metabarcoding Analysis) is an open source pipeline that bundles state-of-the-art bioinformatics 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 [Cingolani et al., 2015] 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 [Rad et al., 2017] and Singularity [Kurtzer et al., 2017], 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 [Coissac et al., 2012]. 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 [Coissac et al., 2012].

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 [Callahan et al., 2017]. The use of ASVs instead of OTUs has been suggested [Callahan et al., 2017]; however, the choice of which approach to use should be based on each study's objective(s) [Pauvert et al., 2019].

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 [Rognes et al., 2016] and CROP [Hao et al., 2011], are used for the clustering of reads in (M)OTUs—the former for the case of the 16S/18S rRNA marker

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

genes, the latter for the case of COI and ITS. Swarm v2 [Mahé et al., 2015] allows ASV inference in all cases.

Taxonomic assignment is performed in an alignment-based approach, making use of the CREST LCAClassifier [Lanzén et al., 2012] and the Silva database [Quast et al., 2013] for the case of 16S and 18S rRNA marker genes; the Unite database [Nilsson et al., 2019b] is used for the ITS gene. In the 16S marker gene case, phylogeny-based assignment is also supported, based on RAxML-ng [Kozlov et al., 2019], EPA-ng [Barbera et al., 2019], and Silva [Quast et al., 2013]. For the COI marker gene, the RDPClassifier [Wang et al., 2007] and the MIDORI database [Machida et al., 2017] are used for the taxonomic assignment. In addition, ecological and phylogenetic analysis are facilitated via the phyloseq R package [McMurdie and Holmes, 2013].

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

2.1.4 Methods & Implementation

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

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

First, FastQC [fas, 2015] 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 [Bolger et al., 2014] 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 [Martin, 2011] 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 [Nikolenko et al., 2013], an algorithm of the SPAdes assembly toolkit [Bankevich et al., 2012], revises incorrectly called bases. PANDAseq [Masella et al., 2012] assembles the overlapping paired-end reads, and then the obiuniq program of OBITools [Boyer et al., 2016] groups all the identical sequences in every sample, keeping track of their abundances. The VSEARCH package [Rognes et al., 2016] 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).

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

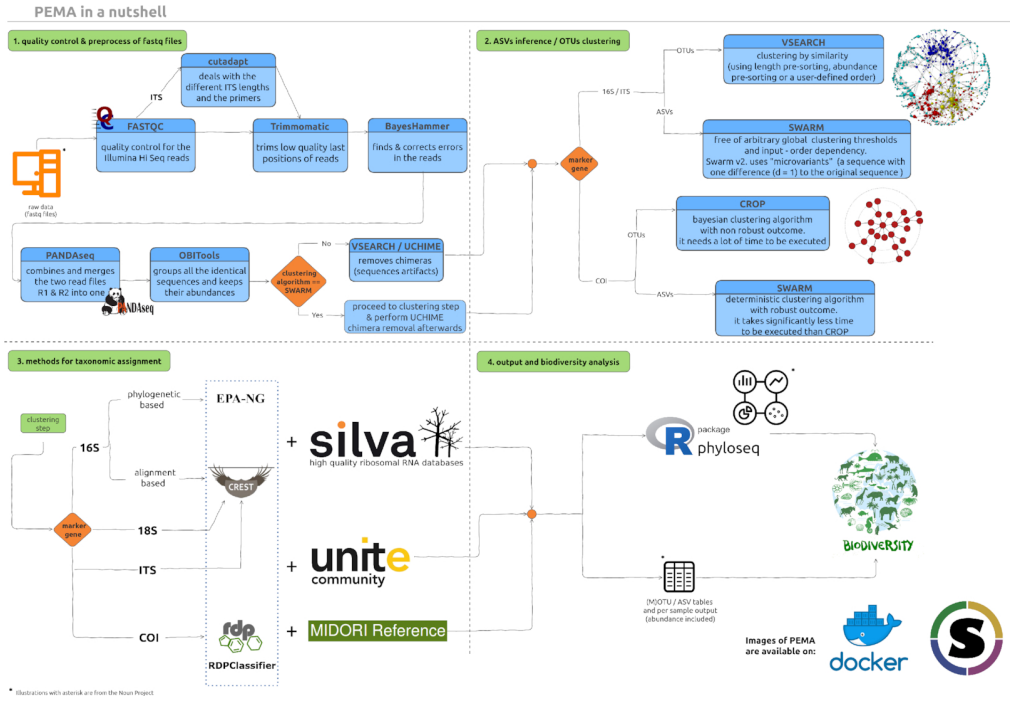


FIGURE 2.1: PEMA comprises 4 parts. The first step (top left) is the quality control and pre-processing of the Illumina sequencing reads. This step is common for both 16S rRNA and COI marker genes. The second step (top right) is the clustering of reads to (M)OTUs or their inferring to ASVs. The third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

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 [Rognes et al., 2016] is used for OTU clustering, while ASVs can be identified by the Swarm v2 algorithm [Mahé et al., 2015]. VSEARCH is an accurate and fast tool that can handle large data sets; at the same time it is a great alternative for USEARCH [Edgar, 2010] because it is distributed under an open source license.

For the ITS and COI marker genes, CROP [Hao et al., 2011], 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 data set 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 [Hao et al., 2011]. ASV inference is conducted by Swarm v2 [Mahé et al., 2015] 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 [Quast et al., 2013] and the Unite [Nilsson et al., 2019b] 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 [Quast et al., 2013] versions new PEMA versions will be available.

For the COI marker gene, PEMA uses the RDPClassifier [Wang et al., 2007] and the MIDORI reference database [Machida et al., 2017] to assign taxonomy of the MOTUs. The MIDORI database contains quality-controlled metazoan mitochondrial gene sequences from GenBank [Benson et al., 2018].

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 [Kozlov et al., 2019] and gappa [phat algorithm] [Czech et al., 2019], Figure 2.2A). PaPaRa [Berger and Stamatakis, 2012] and EPA-ng [Barbera et al., 2019] 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 [Letunic and Bork, 2021] (Figure 2.2B).

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

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

Via the phyloseq R package [McMurdie and Holmes, 2013], 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 [McMurdie and Holmes, 2013] output, a multiple sequence alignment (MSA) and a phylogenetic tree of the OTU/ASVs retrieved can be returned; for the MSA, the MAFFT [Katoh et al., 2002, Nakamura et al., 2018] aligner is invoked while the latter is built by RAxML-ng [Kozlov et al., 2019].

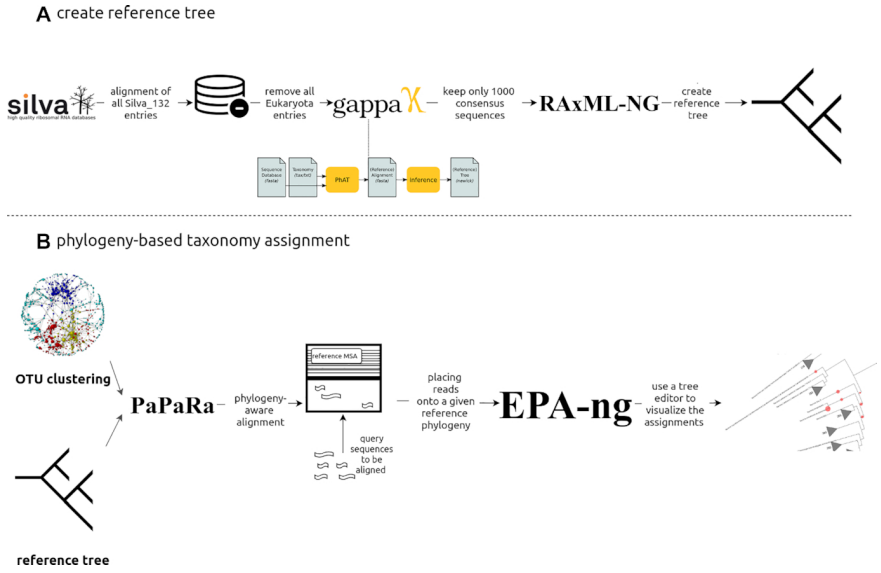


FIGURE 2.2: Phylogeny - based taxonomy assignment. A: Building a reference tree for the phylogeny-based taxonomy assignment to 16S rRNA marker gene OTUs: from the latest edition of Silva SSU, all entries referring to Bacteria and Archaea were used and using the "art" algorithm, 10,000 consensus taxa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

PEMA installation and main output

PEMA container-based installation

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

PEMA output

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

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

2.1.5 Results & Validation

Evaluation

To evaluate PEMA, 2 approaches were followed. First, PEMA was benchmarked against mock community data sets. Second, PEMA was used to analyse previously published data sets. 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. [Gohl et al., 2016] with 20 different bacterial species was studied. Correspondingly, in the case of the 18S rRNA marker gene, a mock community of Bradley et al. [Bradley et al., 2016] with 12 algal species was used; for the ITS, one of Bakker [Bakker, 2018] including 19 different fungal taxa; and for the case of the COI marker gene, a mock community of Bista et al. [Bista et al., 2018] containing 14 metazoan species. More information on the mock communities, their original studies, and the results of PEMA for various combinations of parameters can be found in Additional File 2: Mock Communities.

Complementary to the mock community evaluation, 2 publicly available data sets from published studies were investigated through PEMA. For the 16S rRNA marker gene, the data set reported by Pavloudi et al. [Pavloudi et al., 2017a] was used; the original study aimed at investigating the sediment prokaryotic diversity along a transect river–lagoon–open sea. For the COI case, the data set of Bista et al. [Bista et al., 2017] 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* [Harrison et al., 2019] 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: $F1 = 2 \times (precision \times recall)/(precision + recall)$ [Sammut and Webb, 2011].

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

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)

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.

Evaluation using real-world data

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. [Gohl et al., 2016], there was a discrepancy in the identification of those 2 species that was dependent on the amplification protocol used. It is worth mentioning that as d increases, taxa cannot be identified to species level at all; however, FP assignments decrease. Thus, when d = 30 and strictness = 0.6 for the KAPA samples, Enterococcus was not identified at all; however, PEMA finds its greatest $F1$ value (at the genus level, see Table 2.1) as the FP assignments returned are minimized. When PEMA was run using the VSEARCH clustering algorithm, high precision values were returned in all cases (>0.79). However, the recall values were decreased when using Swarm v2 (0.65–0.68).

18S rRNA

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

ITS

When PEMA was performed using the Swarm v2 algorithm (d = 20) and targeting the ITS2 region, ASVs from 5 of the 19 species of the mock community were assigned to species level, 10 to genus, 2 to family, and 2 to class level. Contrary to the study by Bakker [Bakker, 2018], 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 [Bakker, 2018], 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. data set [Bista et al., 2018] 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. [Bista et al., 2018] 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 —

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

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 data sets.

Comparison with existing software

PEMA's features were compared with those of mothur [Schloss et al., 2009], QIIME 2 [Bolyen et al., 2018], LotuS [Hildebrand et al., 2014] and Barque. Table 2.2 presents a detailed comparison among the 4 tools' features in terms of marker gene support, diversity and phylogeny analysis capability, parameter setting and mode of execution, operation system availability, and HPC suitability. As shown, PEMA is equally feature - rich, if not richer in certain feature categories, compared with the other software packages. In particular, PEMA's support for COI marker gene studies is distinctive; 2 methods for taxonomy assignment are supported, and PEMA's easy parameter setting, step - by - step execution, and container distribution render it user and analysis friendly.

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

TABLE 2.3: OTU predictions and execution time for the different pipelines.
* data from [Pavloudi et al., 2017a]
** ~ 56 if the reference database is already built

Evaluation on real datasets and against other tools

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

16S rRNA marker gene analysis evaluation

To evaluate PEMA's performance, a comparative analysis of the Pavloudi et al. [Pavloudi et al., 2017a] data set with mothur [Schloss et al., 2009], QIIME 2 [Bolyen et al., 2018], LotuS [Hildebrand et al., 2014] and PEMA was conducted.

It is known that the choice of parameters affects the output of each analysis; therefore, it is expected that different user choices might distort the derived outputs. For this reason and for a direct comparison of the pipelines, we have included all the commands and parameters chosen in the framework of this study in Additional File 1: Supplementary Methods. The results of the processing of the sequences by PEMA are presented in Table S1. All analyses were conducted on identical Dell M630 nodes (128 GB RAM, 20 physical Intel Xeon 2.60 GHz cores). LotuS, mothur, and QIIME 2 operated in a single-thread (core) fashion. PEMA, given the BDS intrinsic parallelization [Cingolani et al., 2015], operated with up to the maximum number of node cores (in this case 20).

The execution time and the reported OTU number of each tool are presented in Table 2.3. LotuS and PEMA resulted in a final number of OTUs comparable to that of Pavloudi et. al [Pavloudi et al., 2017a]. Clearly, owing to PEMA's parallel execution support, the analysis time can be significantly reduced (~ 1.5 hours in this case). The execution time depends on the parameters chosen for each software (see Additional File 1: Supplementary Methods).

Owing to the non - full overlap of the sequence reads, mothur resulted in an inflated number of OTUs; thus, it was excluded from further analyses. The results of all the pipelines were analysed with the phyloseq script that is provided with PEMA. The taxonomic assignment of the PEMA - retrieved OTUs is shown in Figure 2.3. The phyla that

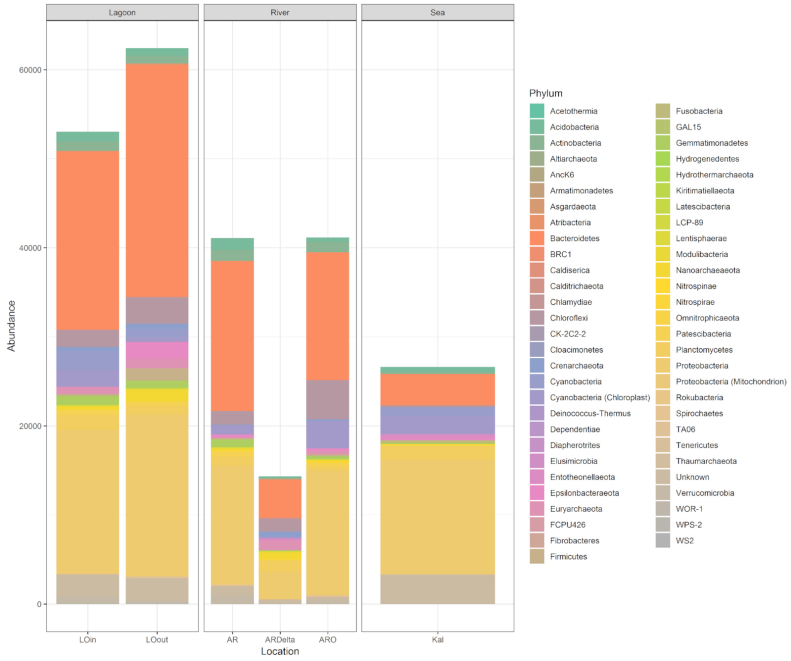


FIGURE 2.3: OTU bar plot at the phylum level. Bar plot depicting the taxonomy of the retrieved OTUs from PEMA for the data set of Pavloudi et al. [Pavloudi et al., 2017a], at the phylum level for the case of the 16S marker gene. AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; LOin: Logarou station inside the lagoon; LOout: Logarou station in the channel connecting the lagoon to the gulf; Kal: Kalamitsi.

were found in the samples are similar to the ones that were found in the original study [Pavloudi et al., 2017a]. 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 [Pavloudi et al., 2017a], the general trend of a decreasing number of OTUs with increasing salinity was observed as in the original study (Supplementary Figure S1). Notably, this result was not observed with the other tested pipelines (Supplementary Table S3). Furthermore, each of the pipelines resulted in a different taxonomic profile (Supplementary Figures S2–S4), with an extreme case of missing the order of Betaproteobacteriales (Supplementary Figures S5–S7).

Moreover, when the PERMANOVA analysis was run for the results of PEMA, LotuS, and DADA2, it was clear that the microbial community composition was significantly different in each of the 3 sampled habitats (i.e., river, lagoon, open sea) (PERMANOVA: F.Model = 7.0718, $P < 0.001$; F.Model = 6.5901, $P < 0.001$; F.Model = 2.2484, $P < 0.05$, respectively), which is in accordance with Pavloudi et al. [Pavloudi et al., 2017a]. However, this was not the case with Deblur (PERMANOVA: $P > 0.05$). Overall, PEMA's output is in accordance with the original study [Pavloudi et al., 2017a], and seen through this perspective PEMA performed equally well with the other tested pipelines, along with having the shortest execution time.

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

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

COI marker gene analysis evaluation

Bista et al. [Bista et al., 2017] 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. [Bista et al., 2017] dataset with Barque [9]; the commands and parameters chosen can be found in Additional File 1: Supplementary Methods. Regarding the creation of the MOTU table, in the Bista et al. [Bista et al., 2017] study VSEARCH [Rognes et al., 2016] was used with a clustering at 97% similarity threshold. Afterwards, the BLAST+ (megablast) algorithm [Camacho et al., 2009] 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 [Bista et al., 2017]. As discussed in the publication, this approach resulted in 138 unique MOTUs of which 73 were assigned to species level. For PEMA's evaluation, the chosen clustering algorithm was Swarm v2, using different options for the cluster radius (d) parameter (Table 2.4); according to Mahé et al. [Mahé et al., 2015], 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 [Machida et al., 2017] using RDPClassifier [Wang et al., 2007]. The results of the processing of the sequences are reported in Supplementary Table S3. For the case of Barque, the BOLD Database was used [Ratnasingham and Hebert, 2007].

As shown in Table 2.4, PEMA resulted in 83 species-level MOTUs with a cluster radius (d) of 2, which is similar to the findings of the published study (i.e., 73 species). Although both the clustering algorithm and the taxonomy assignment methods were different between the original [Bista et al., 2017] and the present study, the results regarding the number of unique species present in the samples are in agreement to a considerable extent.

The computational time required by PEMA for the completion of the analysis is also reported in Table 2.4. Regardless of the value of the d parameter, all analyses were

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

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. [Bista et al., 2017] ; * Taxonomies identical to the published study (species level), ** Taxonomies identical to the published study (genus level).

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 [Nikolenko et al., 2013]), which plays an important part in the enhanced results that PEMA returns, but it also requires a certain computational time; Barque does not have an analogous step, and therefore its overall execution time is shorter.

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

2.1.6 Discussion

OTU clustering vs ASV inference

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

However, the OTU concept is not compulsorily related to the clustering approach; it is widely used to describe results based on its biological meaning but it does not imply clustering. In addition, according to Callahan et al. [Callahan et al., 2017], "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 [Callahan et al., 2017]. Swarm v2 produces all possible *microvariants* of an amplicon to implement an exact-string comparison [Mahé et al., 2015]. 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. [Mahé et al., 2014], which variant-based approaches manage to address. However certain algorithms for OTU clustering such as VSEARCH have been proven to be especially reliable, and they are widely used by many researchers. Furthermore, ASVs intend to improve taxonomic resolution; however, a vast number of inferred ASVs (see here [10] for more) can lead to inflation of diversity estimates, especially in the case of microbial communities, thus making the analysis even more complicated.

ASV or OTU approaches are supported by PEMA, although we have found that similar ecological results are produced by both these methods, as also suggested by Glassman and Martiny [Glassman and Martiny, 2018].

Beyond environmental ecology, ongoing and future work

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

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

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

[11] https://www.elixir-greece.org/

[12] https://www.elixir-greece.org/

[13] http://www.embrc.eu

Conclusions

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

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

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

2.1.7 Advances and PEMA modules added since its publication

PEMA has been under continuous development and testing. Custom databases can be now used to train both classifiers used in the PEMA framework, thus the taxonomy assignment step is not limited by the reference databases included on PEMA. With the release of the v.2.1.3 [14] version, PEMA was re-architectured completely aiming at an easier way for people to contribute. On top of that, several modules have been added, mostly in an attempt to address requests from users and e-infrastructures (e.g., LifeWatch ERIC IJI [15])(see Figure 2.4). Similar efforts have been done so PEMA will be integrated in the HYPATIA [16], the Cloud infrastructure of the ELIXIR-GR community.

On its current version (v.2.1.5) it now supports the analysis of one extra marker gene, the 12S rRNA gene, by exploiting the 12S Vertebrate Classifier v2.0.0-ref database [Porter, 2021]. For the case of 18S rRNA marker gene, the PR2 database [Guillou et al., 2012] was integrated so now the user may select between Silva and PR2, while Silva v.138 has been also added. Furthermore, thanks to the ncbi-taxonomist tool [Buchmann and Holmes, 2020], PEMA now provides an extended OTU/ASV table where in the last column the NCBI Taxonomy Id for the taxonomic level closer to the species name rank for which there is one, is available. Last but not least, a new version of the parameters file has been made to provide a machine-readable version of it so the values set by the user can be parsed for potential errors in an automatic way.

The potential of the eDNA metabarcoding method as well as the valid PEMA output were emphasized in a recent study where Autono-mous Reef Monitoring Structures

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

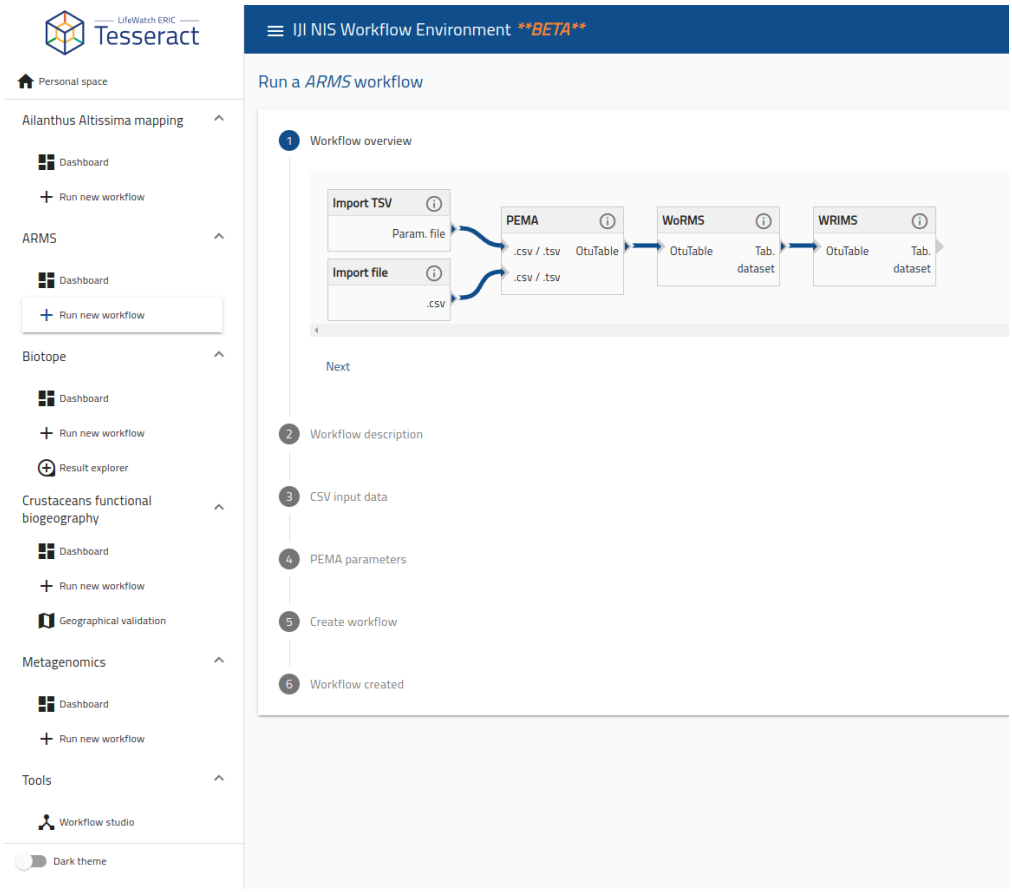


FIGURE 2.4: PEMA is now available through the LifeWatch ERIC portal called Tesseract which is currently under a beta version. A web interface is now available allowing users that are not familiar with Unix to use PEMA. Most importantly, users that have no access to computing resources required for their analyses can now use the capacity of Tesseract.

(ARMS) data were combined with amplicon studies to record, for the first time in Greek waters, the nudibranch *Anteaeolidiella lurana* (Ev. Marcus & Er. Marcus, 1967) [Bariche et al., 2020].

Supplementary Material

You may find the Supplementary files of this study through PEMA's publication [17]

Additional File 1: Supplementary Methods: Description of tools invoked by PEMA and their licenses. Description of the commands, along with their parameters, used to run PEMA, mothur, LotuS, and QIIME 2.

Additional File 2: Mock Communities: Details about the mock communities chosen and their corresponding studies, as well as the returned output of PEMA for each for a number of sets of parameters.

[17] https://academic.oup.com/gigascience/article/9/3/giaa022/5803335#supplementary-data

Supplementary Table S1: Number of sequences after each pre-processing step for the case of 16S rRNA gene.

Supplementary Table S3: Number of sequences after each pre-processing step for the case of COI, dataset from Bista et al. [Bista et al., 2017].

Supplementary Table S2: Diversity indices of the samples.

Supplementary Figure S1: Linear regression between the number of OTUs (averaged per sampling station) and the salinity of the sampling stations. L: Lagoon; S: Sea; R: River; 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.

Supplementary Figure S2: Bar plot depicting the taxonomy of the retrieved OTUs from LotuS at the phylum level.

Supplementary Figure S3: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using Deblur at the phylum level.

Supplementary Figure S4: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using DADA2 at the phylum level.

Supplementary Figure S5: Bar plot depicting the taxonomy of the retrieved OTUs from LotuS at the class of Betaproteobacteriales.

Supplementary Figure S6: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using Deblur at the class of Betaproteobacteriales.

Supplementary Figure S7: Bar plot depicting the taxonomy of the retrieved OTUs from PEMA at the class of Betaproteobacteriales.

2.2 The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data[18]

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

2.2.1 Abstract

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

2.2.2 Introduction

Metabarcoding: concept and caveats

DNA metabarcoding is a rapidly evolving method that is being more frequently employed i n a range of fields, such as biodiversity, biomonitoring, molecular ecology and others [Deiner et al., 2017, Ruppert et al., 2019]. Environmental DNA (eDNA) metabarcoding, targeting DNA directly isolated from environmental samples (e.g., water, soil or sediment, [Taberlet et al., 2012a]), is considered a holistic approach (Stat et al. 2017) in terms of biodiversity assessment, providing high detection capacity. At the same time, it allows

[18] For author contributions and supplementary material please refer to the relevant sections.

wide-scale rapid bio-assessment [Stat et al., 2017] at a relatively low cost as compared to traditional biodiversity survey methods [Ji et al., 2013].

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 [Deagle et al., 2014]. 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 [Taberlet et al., 2012b]. However, it is essential that comprehensive public databases containing well curated, up-to-date sequences from voucher specimens are available [Schenekar et al., 2020]. This way, sequences generated by universal primers can be compared with the ones in reference databases, assessing sample OTU composition. The taxonomy assignment step of the eDNA metabarcoding method and thus, the identification via DNA-barcoding, is only as good and accurate as the reference databases [Cilleros et al., 2019].

Nevertheless, there is not a truly "universal" genetic marker that is capable of being amplified for all species across different taxa [Kress et al., 2015]. Different markers have been used for different taxonomic groups [Deiner et al., 2017]. 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 [Coissac et al., 2012]. 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 [Hebert et al., 2003]. There are cases where COI has been the standard marker for metabarcoding, such as in the assessment of freshwater macroinvertebrates [Elbrecht and Leese, 2017] even though not all taxonomic groups can be differentiated to the species level using this locus [Deiner et al., 2017]; for example, in case of fish other loci are widely used such as 12S rRNA gene (hereafter referred to as 12S rRNA) [Miya et al., 2020].

The COI locus

The mitochondrial cytochrome c oxidase subunit I (also called cox1 or/and COI) is a gene fragment of 700 bp, widely used for metazoan diversity assessment. Here we present some of the reasons that microbial eukaryotes and prokaryotes are also amplified in such studies, raising the issue of the known unknown sequences. COI is a fundamental part of the heme aa3-type mitochondrial cytochrome c oxidase complex: the terminal electron acceptor in the respiratory chain. Even if aa3-type Cox have been found in bacteria, there are also other cytochrome c oxidase (Cox) groups, such as the cbb3-type cytochrome c oxidases (cbb3-Cox) and the cytochrome ba3 [Ekici et al., 2012, Schimo et al., 2017].

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 [Song et al.,

2008]. Numts are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms [Bensasson et al., 2001].

Thus, as Mioduchowska et al. (2018) [Mioduchowska et al., 2018] highlight, when universal primers are used targeting the COI locus, it is possible to co-amplify both non-target numts and prokaryotes [Siddall et al., 2009]. 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 [Mioduchowska et al., 2018].

Even though there are various known issues [Deagle et al., 2014], COI is indeed considered as the "gold standard" for community DNA metabarcoding of bulk metazoan samples [Andújar et al., 2018]; bulk is an environmental sample containing mainly organisms from the taxonomic group under study providing high quality and quantity of DNA [Taberlet et al.]. However, as highlighted in the same study, this is not the case for eDNA samples. As Stat et al. (2017) [Stat et al., 2017] 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 [Yang et al., 2013, 2014, Collins et al., 2019]. Cold-adapted marine gammaproteobacteria are an indicative example for this case as shown by Siddall et al. (2009) [Siddall et al., 2009].

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 [Aylagas et al., 2016]) 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 [Siddall et al., 2009], 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., [Sinniger et al., 2016, Haenel et al., 2017]).

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) [Bernard et al., 2018]. 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) [Barbera et al., 2019] (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. [Jamy et al., 2020].

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) [Machida et al., 2017]. 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 [Ratnasingham and Hebert, 2007]. 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 [19]), the respective DNA sequences were extracted from their corresponding genomes. This way an additional 217 archaeal and 9,154 bacterial sequences were obtained (see Table 1). In total, sequences from 15 archaeal, 371 bacterial families and 60 taxonomic groups of higher level not assigned in the family level, were gathered. An overview of the approach that was followed is presented in Figure 2.5.

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 [Czech et al., 2019]; 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 [Katoh et al., 2002, Nakamura et al., 2018].

Resources	**bacteria**		**archaea**	
	# of sequences	**# of strains**	**# of sequences**	**# of strains**
BOLD	3,917	2,267	117	117
PFam-oriented	9,154	4,532	217	115

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

In the case of Eukaryotes, the alignment of the corresponding sequences would be impractically long because of their large number (183K sequences). To address this challenge, a two-step procedure was followed; a sequence subset of 500 sequences (*reference set*) was selected and aligned and then used as a backbone for the alignment of all the remaining eukaryotic COI sequences. All sequences were considered reliable as they were retrieved from curated databases (Midori2 and BOLD). To build the reference set, a number (n) of the longest sequences from each of the various phyla were chosen, proportionally to the number (m) of sequences of each phylum (see Supplementary Table 2.6). The –min-tax-level parameter of the PhAT algorithm corresponded to the class level, for the case of eukaryotes and to the family level for archaea and bacteria. This parameter

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

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 [Czech et al., 2020].

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 [Steenwyk et al., 2020] to keep only phylogenetically informative sites. The final MSA is available on GitHub;
see the trimmed_all_consensus_aligned_adjust_dir.aln file.

The reference tree was then built based on this trimmed MSA using the IQ-TREE2 software [Hoang et al., 2018a, Minh et al., 2020]. ModelFinder was invoked through IQ-TREE2 and the GTR+F+R10 model was chosen based on the Bayesian Information Criterion (BIC) among 286 models that were tested. The phylogenetic tree was then built using 1,000 bootstrap replicates (-B 1,000) and 1,000 bootstrap replicates for Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) (1,000 1000).

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

A thorough description of all the implementation steps for building the reference tree is presented in this Google Collab Notebook [20] . The computational resources of the IMBBC High Performance Computing system, called Zorba [Zafeiropoulos et al., 2021c], 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 [Rognes et al., 2016] 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 [Berger and Stamatakis, 2012]. The query sequences are then split to build a discrete query MSA. Finally, the Evolutionary Placement Algorithm EPA-ng [Barbera et al., 2019] is used to assign the query sequences to the reference tree.

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

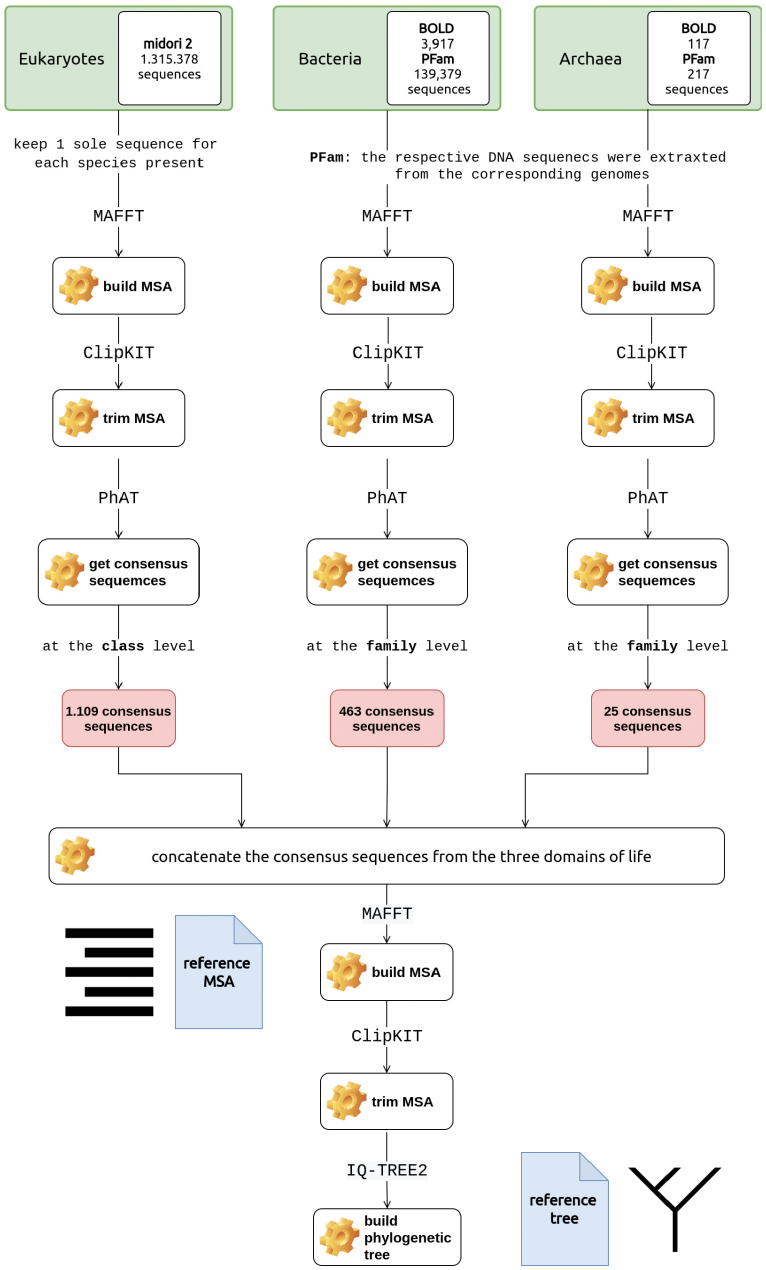


FIGURE 2.5: Overview of the approach followed to build the COI reference tree of life. Sequences were retrieved from Midori 2 (eukaryotes) and BOLD (bacteria and archaea) repositories. Consensus sequences at the family level were built for each domain specific dataset. MAFFT and consensus sequences at 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.

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 [Ondov et al., 2011]. 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 [21].

2.2.5 Results & Validation

Evaluation of the phylogenetic tree

The inferred phylogenetic tree is shown in Figure 2.6, with the bacterial (light blue) and archaeal (dark green) branches highlighted; in Supplementary material 3: Figure S1 the distribution of the eukaryotic phyla on the tree is presented. As shown, bacteria and archaea can be distinguished from eukaryotes. Scattered bacterial branches that are present among eukaryotic ones represent the diversity of the COI locus. To evaluate the phylogenetic tree, the set of consensus sequences were placed on it using the EPA-ng algorithm. The placements (see .jplace through a phylogenetic tree viewer, e.g. iTOL) verified that the phylogenetic tree built is valid, as the consensus sequences have been placed in their corresponding taxonomic branches (Supplementary material 4: Figure S2; the figure was built using the heat-tree module of the gappa examine tool).

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

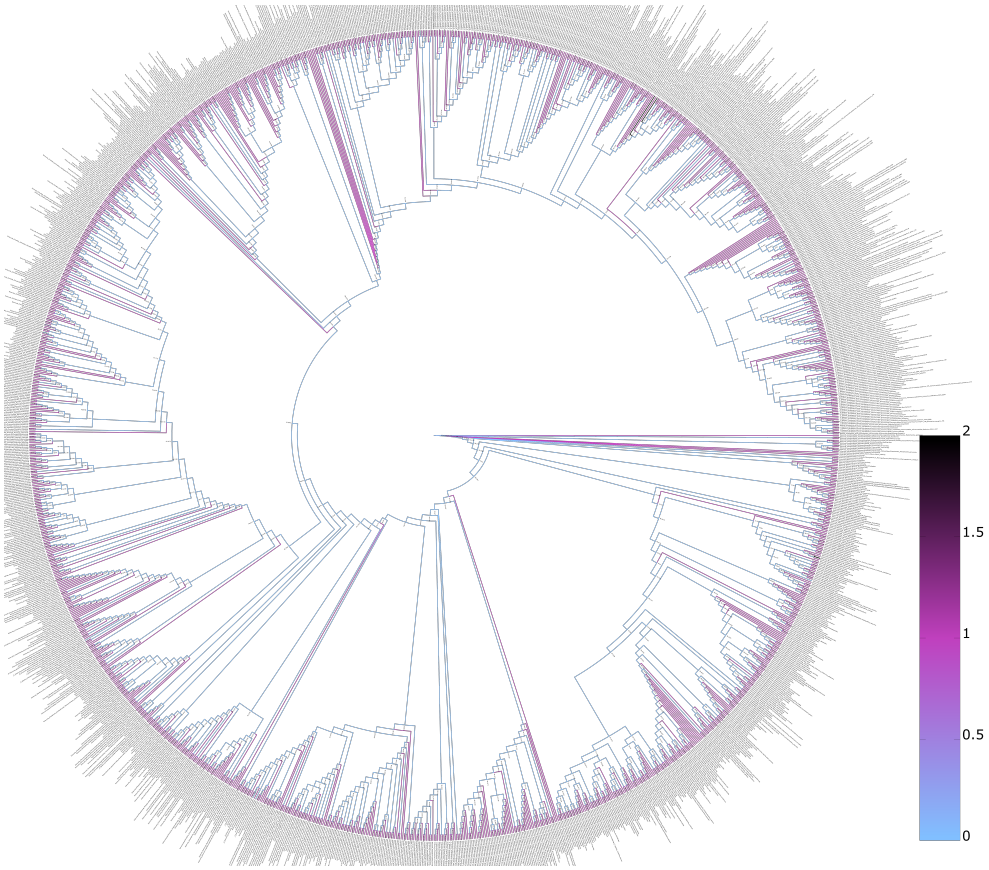


FIGURE 2.6: Phylogenetic tree of the consensus sequences retrieved; the tree that DARN makes use of. Light blue: bacterial branches. Dark green: archaeal branches. White: eukaryotic branches.

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

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

DARN using real community data

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

More specifically, 57 marine, surface water, eDNA samples from Ireland were analysed through a. QIIME2 [Bolyen et al., 2018] and DADA2 [Callahan et al., 2016] and, b. PEMA [Zafeiropoulos et al., 2020]. Similarly, 18 mangrove and 18 reef marine eDNA samples from Honduras, were analyzed using a. JAMP v0.74 [23] and DnoisE [Antich et al., 2021] 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) [Obst et al., 2020] 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 [Mahé et al., 2015] that PEMA invokes was set equal to 2 and 10 to cover a great range of different cases [Kamenova, 2020]. DARN was then executed using the ASVs retrieved in each case as input. All the DARN analyses and the PEMA runs were performed on an Intel(R) Xeon(R) CPU E5649 @ 2.53GHz server of 24 CPUs and 142 GB RAM in the Area52 Research Group at the University College Dublin.

The number of sequences returned, using various bioinformatic analyses, ranged from circa 3k to 214k (Table 2.7) in the different amplicon datasets used. A coherent visual representation of the DARN outcome for all the datasets is available here[24]. The visual and interactive properties of the Krona plot allow the user to navigate through the taxonomy. Furthermore, DARN also supports a thorough investigation per OTU/ASV, as it returns a .json file with all the OTUs/ASVs ids that have been assigned in each of the four categories (Bacteria, Archaea, Eukaryotes and distant).

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

[23] https://github.com/VascoElbrecht/JAMP
[24] https://hariszaf.github.io/darn/

Sample(s) accession number	Sample type	Primer set	Amplicon length (bp)	Bioinfo pipeline(s)	# of ASVs	~% of sequence assignments per domain (if PEMA[a])			
						Eukaryotes	Bacteria	Archaea	distant
ERS6449795-ERS6449829	eDNA	jgHCO2198 - jgLCO1490 & LoboF1 - LoboR1	658	QIIME2 - Dada2	13,376	11	88.0	0.02	0.003
				PEMA	39,454	25	75.0	0.1	0.4
ERS6463899–ERS6463901				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

assigned as Eukaryotes, suggesting that non-eukaryotic sequences are more abundant in environmental samples. This could be expected since prokaryotes are amplified as whole organisms from environmental samples, while metazoa that are usually the targeted taxa in COI studies, are amplified from DNA traces or/and other parts of biological source material.

2.2.6 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. [Yang et al., 2013, Aylagas et al., 2016] 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 [Yang et al., 2013]. Furthermore, dark matter seems to be particularly common in eDNA as compared to bulk samples [Andújar et al., 2018]. However, it should be mentioned that the high number of prokaryotic sequences in COI metabarcoding data is also reflecting known issues with contamination [Kumar et al., 2013, Dittami and Corre, 2017, De Simone et al., 2020], incorrectly labeled reference sequences [Steinegger and Salzberg, 2020] and holobionts [Gilbert et al., 2012, Salvucci, 2016] 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 [Weigand et al., 2019, Miya et al., 2020] 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 [Huys et al., 2008].

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

Chapter 3

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

Citation:
Zafeiropoulos H, Paragkamian S, Ninidakis S, Pavlopoulos GA, Jensen LJ, Pafilis E. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. Microorganisms. 2022; 10(2):293.
DOI: 10.3390/microorganisms10020293[1]

3.1 Abstract

To elucidate ecosystem functioning, it is fundamental to recognize *what* processes occur in which environments (*where*) and which microorganisms carry them out (*who*). Here, we present PREGO, a one-stop-shop knowledge base providing such associations. PREGO combines text mining and data integration techniques to mine such what-where-who associations from data and metadata scattered in the scientific literature and in public omics repositories. Microorganisms, biological processes, and environment types are identified and mapped to ontology terms from established community resources. Analyses of co-mentions in text and co-occurrences in metagenomics data/metadata are performed to extract associations and a level of confidence is assigned to each of them thanks to a scoring scheme. The PREGO knowledge base contains associations for 364,508 microbial taxa, 1090 environmental types, 15,091 biological processes, and 7,971 molecular functions with a total of almost 58 million associations. These associations are available through a web interface (https://prego.hcmr.gr), an Application Programming Interface (API), and bulk download. By exploring environments and/or processes associated with each other

[1] For author contributions and supplementary material please refer to the relevant sections. This is a modified version of the published version, in terms of relevance, coherence and formatting.

or with microbes, PREGO aims to assist researchers in design and interpretation of experiments and their results. To demonstrate PREGO's capabilities, a thorough presentation of its web interface is given along with a meta-analysis of experimental results from a lagoon-sediment study of sulfur-cycle related microbes.

3.2 Introduction

Microbes are omnipresent and impact global ecosystem functions [Falkowski et al., 2008] through their abundance [Bar-On et al., 2018], versatility [Delgado-Baquerizo et al., 2016], and interactions [Röttjers and Faust, 2018]. These facts have inspired microbiologists from diverse scientific fields to study their genotype and phenotype [Morris et al., 2020], their metabolism [Biggs et al., 2015], and their interactions with the environment [Hall et al., 2018]. All this work has resulted in a wealth of knowledge available in the forms of literature and experimental data. Literature contains vast amounts of information in the free text form that overwhelms researchers. Advanced text mining methods [Jensen et al., 2006] have been developed to assist this issue. Experimental data and their metadata require mining [Delmont et al., 2011] as well for their integration, mostly through metagenomic mining from online repositories. Hence, the combination of this knowledge about microbial life (who), their metabolic functions (what), and the environment they influence (where) is an important step to study ecosystem function [Raes and Bork, 2008].

High Throughput Sequencing (HTS) has turned the page on microbial ecology studies [Nilsson et al., 2019a]. Over the past 20 years, both the taxonomic and the functional profiles of microbial communities from both local and large-scale regions (e.g., Tara Oceans [Pesant et al., 2015], Earth Microbiome [Gilbert et al., 2014]) are being accumulated at a higher and higher rate. Extreme environments, i.e., areas with high salinity, low pH, etc., are being studied, providing us with unprecedented insight [Shu and Huang, 2021]. Both amplicon and shotgun metagenomics studies have played a crucial part in this development. Latest technological breakthroughs, such as Metagenome-Assembled Genomes (MAGs) and Single Amplified Genomes (SAGs), are enhancing the assessment of the taxonomic and functional repertoire of microbiomes even further. However, the mass use of these technologies and their consequent data have led to a number of needs and challenges, with metadata curation being among the most crucial ones.

Standards-promoting communities, like Genomic Standards Consortium (GSC)[2], their efforts, like Minimum Information about any (x) Sequence (MIxS) [Yilmaz et al., 2011b], and projects endorsing those, like National Microbiome Data Collaborative (NMDC) [Wood-Charlson et al., 2020, Vangay et al., 2021], offer guidelines and best-practices to assist the annotation of microbial ecology samples. Controlled vocabularies and ontologies contribute to these efforts as they describe each subject area with formal terms [Walls et al., 2014]. Environment types, for example, are described by the Environment Ontology (ENVO) [Buttigieg et al., 2016]. Other key biological aspects that have been captured include molecular functions (Gene Ontology Molecular Function (GOmf) [Ashburner et al., 2000, gen, 2021], Enzyme Commission nomenclature [noa, 1999], etc.), and the pathways carrying out different biological processes (GO Biological Process (GObp),

[2] https://gensc.org/

MetaCyc [Caspi et al., 2020], etc.). These knowledge structures, along with taxonomic centralized resources like the National Center for Biotechnology Information (NCBI) Taxonomy [Schoch et al., 2020] and LPSN (List of Prokaryotic names with Standing in Nomenclature) [Parte et al., 2020], provide the means for a standardized representation of, for example, environments, process-oriented terms, and microbial taxa, respectively. Global-scale public resources (like MGnify [Mitchell et al., 2020], JGI/IMG [Chen et al., 2021], MG-RAST [Wilke et al., 2015]) combine some of the aforementioned resources to support the collection, analysis, and distribution of multiple types of HTS data (e.g., amplicon, metagenomics, metatranscriptomics, etc.).

Besides the data and the analyses *per se*, the related scientific literature stores valuable information in billions of text lines. PubMed [Schoch et al., 2020] and PubMed Central (PMC) [Roberts, 2001] are gateways to relationships among microbes (*who*), the environments they live in (*where*) and their associated processes and functions (*what*) hidden in text [Harmston et al., 2010]. Text mining (on both literature and metadata) can serve the extraction of these relationships. Named Entity Recognition (NER) can, for example, locate organism names [Pafilis et al., 2013], ENVO and GO terms [Pafilis et al., 2016] mentioned in text and map them to their corresponding identifiers. Association statistics, like co-mention analysis, can subsequently suggest ranked association among such entities [Von Mering et al., 2005, Franceschini et al., 2012]. The new era of omics has been interwoven with data integration [Gomez-Cabrero et al., 2014] by bringing together scattered and fragmented pieces of information.

The time is ripe for tools that integrate all this knowledge and henceforth assist researchers to tackle major challenges like climate change [Cavicchioli et al., 2019], sustainability [D'Hondt et al., 2021], and synthetic ecology [Conde-Pueyo et al., 2020]. Many resources have emerged in this realm [Baltoumas et al., 2021a], each one serving a specific purpose, such as BacDive [Reimer et al., 2019]. BacDive is a large-scale curated database with prokaryotic information about phenotypic, morphological, and metabolic information. Other resources like Microbe Directory [Shaaban et al., 2018], Web of Microbes (WoM) [Kosina et al., 2018], and Microbial Interaction Network Database (MIND[3]) focus on microbial environmental conditions, metabolite interactions with microbes and microbe-microbe interactions, respectively. In addition, taking advantage of aforementioned resources, novel pipelines, e.g., [Tang et al., 2020], are emerging with the aim to explore the network associations of who (microbial taxa) is performing what (microbial processes) and where (environments) directly using graph theory [Koutrouli et al., 2020]. These analyses and resources are important because microbiologists can enrich their data to explore hypotheses but also to identify potential gaps in knowledge regarding these associations [Li et al., 2021].

Here, we present PREGO, a hypothesis generation web resource that is designed to be useful to microbiologists—in particular microbial ecologists and environmental microbiologists. Its specific aims include: (a) the gathering of source data, metadata, and literature followed by the extraction of microorganism, process, environment associations contained therein, (b) making such a mined knowledge base available to life sciences researchers via an easy to use and explore web portal. As such, PREGO can be useful also to

[3] http://www.microbialnet.org/mind_home.html

system microbiologists and large-scale data analysts through bulk download and programming access. We document the principles, analysis methodology, and contents behind PREGO. Last but not least, we demonstrate PREGO's capabilities for researcher-support related to the above through a case study involving sulfate-reducing microorganisms.

3.3 Methods & Implementation

PREGO is a resource designed to assist molecular ecologists in acquiring a single point overview of what-where-who process–environment–organism associations. The system is comprised of two main parts: (a) a server that periodically harvests data and extracts process-environment-organism associations from the scientific literature, environmental samples, and genome annotation sequences (Figure 3.1, step 1 to 5) and (b) a web-based interface as well as an Application Programming Interface (API) that provides users and programmers with a friendly way to extract and navigate across the process–environment–organism associations (Figure 3.1, step 6).

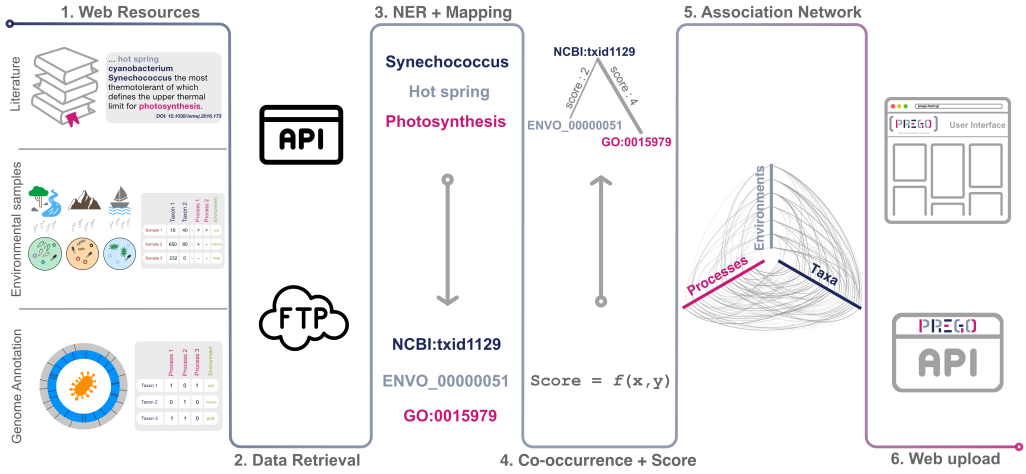


FIGURE 3.1: PREGO analysis methodology: PREGO periodically retrieves three distinct types of data from open access resources. Scientific text, environmental sample data, and genomic annotations are handled with respective methodologies in order to standardize their entities. Named Entity Recognition and Comention/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 A.2). Icons used from the Noun Project released under CC BY: Books by Shakeel Ch., Bacteria by Maxim Kulikov, ftp by DinosoftLab, Mountain by Diane, Ship on Sea by farra nugraha, River by Chanut is Industries.

3.3.1 Entity Types, Channels, and Associations

PREGO supports three entity types: *Process*, *Environment*, and *Organism*. For interoperability and consistency, an ontology or taxonomy is adopted for each type of entity. Processes are represented as Gene Ontology (GO) terms and are grouped either as Biological processes (GObp) or as Molecular functions (GOmf). In addition, Environments are represented by terms from the Environmental Ontology. Organisms are represented by the microbial NCBI Taxonomy Ids (Bacteria, Archaea, and unicellular eukaryotes). For the unicellular eukaryotes, a custom list was populated with the unicellular eukaryotic taxa using a curated list. PREGO's contents are mainly divided into three distinct channels of information based on data origin and format (Figure 3.1, step 1). The *Literature* channel exploits scientific publications, i.e., abstracts and full text open access scientific publications (Table 3.1 and Section 3.3.2). Through the *Annotated Genomes and Isolates* channel, PREGO retrieves genome annotations and their accompanying metadata (Table 3.1 and Section 3.3.3). Finally, the *Environmental Samples* channel supports the integration of metagenomic analyses from both amplicon and shotgun studies. These include taxonomic and functional profiles along with their corresponding metadata (Table 3.1, more details in Section 3.3.4).

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 a Creative Commons license available for commercial and noncommercial use. JGI has its own license, the same applies for BioProject, MEDLINE®, and PubMed® as well.

In cases in which the retrieved data and metadata are in text form, they are standardized to the aforementioned identifiers and taxonomies using Named Entity Recognition (NER) tools, namely the EXTRACT tagger [Pafilis et al., 2016, Jensen, 2016]. In cases where data contain KEGG Orthology terms and/or Uniref identifiers, they are mapped to the respective GOmf using the mapping files available from the UniProt (see Appendix A.1). Associations are extracted after the mapping and standardization of the entities from each resource (Figure 3.1, step 3). The association extraction pipeline is distinct for each chan-

nel and resource because of differences in the data type origin (see prego_gathering_data in the Availability of Supporting Source Codes section). By the means of navigation, the large number of associations returned to the user require a type of sorting; ideally, one that ranks the most trustworthy associations at the top. For those reasons, each channel of PREGO has a dedicated scoring scheme bounded within the (0,5] space for consistency. In Appendix A.1 , the scoring scheme of each channel is elaborated.

3.3.2 Text Mining of Scientific Literature

PREGO implements a text mining methodology to extract associations of the aforementioned entities from literature. The origin of text mining is a corpus that comprises scientific abstracts and full text articles from MEDLINE® and PubMed® and PubMed Central® Open Access Subset (PMC OA Subset) [Sayers et al., 2021], respectively. The building and periodic update of the corpus is possible through the NCBI File Transfer Protocol (FTP) services. PREGO also has a dedicated text-mining dictionary (see Availability of Supporting Source Codes section) that contains the entities ids, names, synonyms, and neglected words (stop words). PREGO dictionary incorporates the ORGANISMS [Pafilis et al., 2013] and ENVIRONMENTS [Pafilis et al., 2015] evaluated dictionaries as well as the experimental dictionaries of Gene Ontology Biological Process and Molecular Function. Text mining is subsequently performed on the corpus using the dictionary through the EXTRACT tagger [Pafilis et al., 2016, Jensen, 2016]. The tagger recognizes the entities of the dictionary in each abstract and full text article and assigns their co-mentions with a score. The score is sensitive to the text structural level of co-mention; higher to lower scoring when co-mention appears in the same sentence, then, in the same paragraph, and lastly in the same article. All these are integrated and normalized to a single score for each association, as implemented in STRING 9.1 [Franceschini et al., 2012] (see Appendix A.3 for more details). In addition, the tagger extracts each mention in every article to provide the origin of each association it extracts.

3.3.3 Annotated Genomes and Isolates

Annotated genomes and isolates comprise the most trustworthy data in PREGO's knowledge base because they refer to a single species/strain and also have manually curated metadata. Among other data types, JGI-IMG [Chen et al., 2021, Mukherjee et al., 2021] includes millions of genes from isolated genomes (isolates), SAGs and MAGs. Such annotations, along with their corresponding metadata, were collected using web-parsing technologies. Their metadata, describing their related environment/ecosystem, were tagged using the EXTRACT tagger to infer organisms—environments associations. The annotated KEGG terms were mapped to GOmf terms (see Appendix A). The GOmf terms were then used to extract organisms—processes associations.

The Struo pipeline [de la Cuesta-Zuluaga et al., 2020] and its outcome when using the Genome Taxonomy DataBase (GTDB) (v.03-RS86) [Parks et al., 2020] was exploited to enrich organisms—processes associations. A set of 21,276 representative genomes, accompanied by UniRef50 annotations, was retrieved using the provided FTP server. The annotations were then mapped to GOmf terms (see Appendix A.1). Related GTDB

genomes were mapped to their corresponding NCBI taxa (see Appendix A.1). All associations extracted from these resources were assigned arbitrarily a confidence level of four out of five. This score choice reflects the high-quality of these data and metadata.

In addition, BioProject data were integrated to PREGO using the NCBI FTP/e-utils services [Sayers et al., 2021]. The BioProject ids that were integrated are the ones that have been assigned a PubMed abstract, a unicellular taxon, and Genome sequencing as data type. Then, using the text mining pipeline, associations were extracted connecting the assigned taxon with the rest of the entities that appear in the abstracts. This method resulted in associations that were assigned a confidence level of three (out of five) because of the combined method of curated data with text mining.

3.3.4 Environmental Samples

MGnify [Mitchell et al., 2020] and MG-RAST [Wilke et al., 2015] repositories provide a great number of public metagenomic records. In the PREGO framework, both amplicon and shotgun metagenomic analyses are retrieved periodically along with their corresponding metadata. Data retrieval from these resources is possible from their Application Programming Interfaces (APIs). Marker gene analyses are retrieved and by measuring the co-occurrence of taxa present in the various environmental types (e.g., biomes, materials, features, etc.) organisms—environments associations are extracted. These associations emerge when a taxon is reported together with a certain environmental type, being mentioned in the metadata of a sample (metadata based co-occurrence). Similarly, analyses of metagenomic samples along with their corresponding metadata and annotations are also retrieved and organisms—environments, organisms—processes and processes—environments are extracted. The processes—environments associations are possible through co-occurrence of the functional annotations of metagenomes with the environmental metadata of the samples.

In all cases, the EXTRACT tagger is used on the microorganism names and the corresponding metadata of each sample to identify their identifiers (NCBI ids, ENVO terms, GOmf, GObp). All associations in this channel are scored based on the number of samples the entity of interest co-occurs with specific sample metadata (e.g., environmental type) or annotations (functional annotations or taxonomic annotations). The same scoring scheme was implemented across the channel resources (see Appendix A.3 for more details), which ranks these associations with a value in the (0,5] space.

3.3.5 Sequence Search

In the case of organisms, PREGO enables sequence-based queries, meaning a sequence (amplicon) can be used as an entry point like it was a taxon name. To this end, a custom database was built using a set of reference custom databases for four commonly used marker genes. For 16S and 18S rRNA, the SILVA database (v.138) [Quast et al., 2013] and the PR2 database (version_4.14.0) [Guillou et al., 2012, Del Campo et al., 2018] were used. Cytochrome c oxidase I (COI) [Suter et al., 2021] is another commonly used marker gene; for this reason, Midori 2 (version GB243) [Leray et al., 2018] was integrated in PREGO's custom database. Finally, for the Internal transcribed spacer (ITS), common in studies

focusing on Fungi, the Unite (version 8.3, accessed 10.05.2021) [Nilsson et al., 2019b] database was added.

3.3.6 Back-End Server and Front-End Implementation

PREGO is a multi-tier web-based application. It is hosted on a 64 GB RAM DELL R540, 20 core, Debian server. Custom API clients (written in Python) are responsible for retrieving the data and metadata from each source (Figure 3.1, step 2). These clients as well as the subsequent methodology (Figure 3.1, step 3 to 6) are updated in regular cycles using custom daemons (see Appendix A.2, Figure A.1). The *mamba/blackmamba* web framework underlies communication to the Postgres association-holding database and the client-side communication. HTML 5, Ajax, JQuery, and custom Javascript enhance the user web experience. PREGO supports widely used browsers (e.g., Chrome, Firefox, Safari, Edge) in various operating systems, such as Windows 10, Linux (Ubuntu 18), and MacOS (10.12, 11).

3.4 Results & Validation

3.4.1 The PREGO Web Resource

Users can access the PREGO contents through its web User Interface (UI) (Figures 3.2 and 3.3), its Application Programming Interface (API) (Figure 3.4), or bulk download of all associations (Appendix D). The User Interface comes with two search fields: a plain text search and a sequence search (Figure 3.2a). The latter is used when the user wants to search for a taxon sequence (see Section 3.3.5 for supported sequence databases). The plain text search supports three types of entry points; the user can search for a taxon name, e.g., *Methanosarcina mazei*, an environmental type, e.g., lagoon, or a biological process e.g., methanogenesis. In all entry points, PREGO returns an overview page consisting of tabs with associations of the entity of interest with the entities of the two other types (Figure 3.2b–d) as well as Documents and Downloads tabs (Figure 3.2e,f).

Regarding the association tabs, when a taxon is used as a query, PREGO returns an overview page consisting of tabs for environments, biological processes, and molecular functions. When an environmental type is used as input, PREGO returns the organisms that have been found to be related to it, as well as the Biological Processes observed in the given environment. Lastly, if a biological process is under study, PREGO returns a tab with the organisms along with another tab with the Environments related to the process. Notably, only the associations with scores higher than 0.5 are presented in the web platform and are sorted in descending order based on their score. The score is visualized with a five-star system (see Appendix A.3 for the scoring scheme). Every association tab contains three tables with associations derived from the PREGO channels (see Section 3.3) along with their supported evidence. The user can both search and scroll through these tables, which makes knowledge extraction easier in cases where, for example, Isolate data contain hundreds of associations. In the *Literature* channel, each association is supported by the scientific articles with text-mining identified co-mentions. When a user clicks on an association, a popup window appears. This window displays

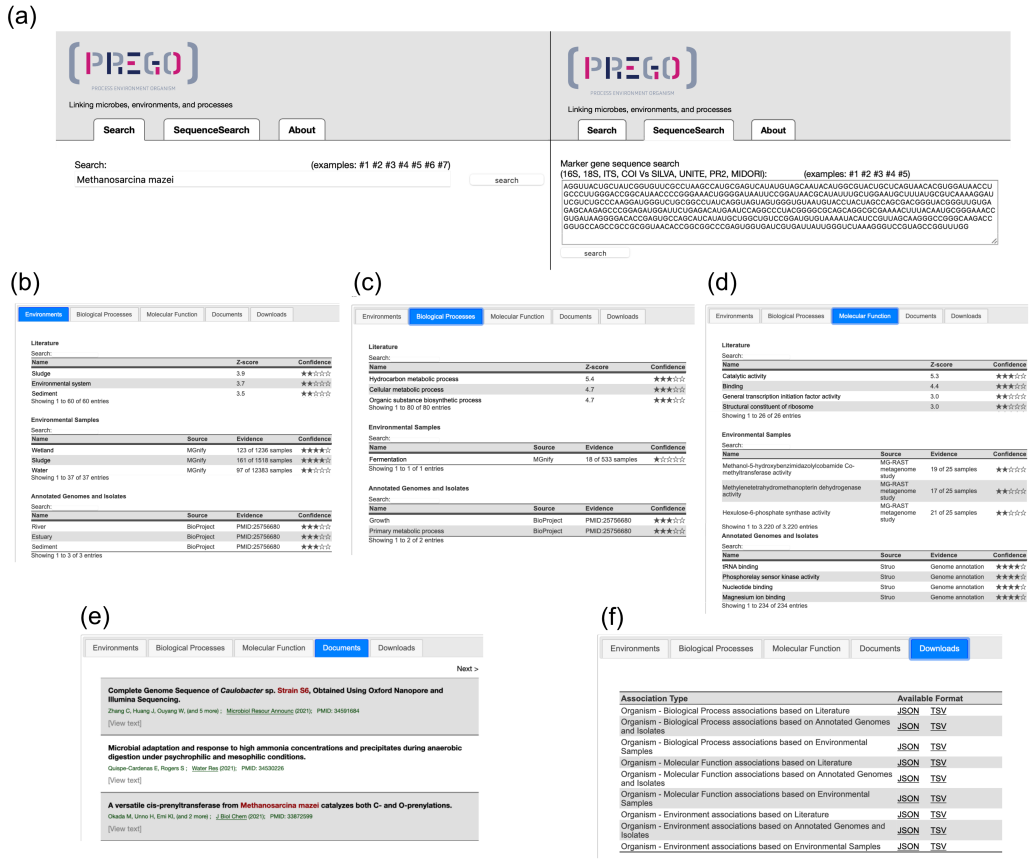


FIGURE 3.2: PREGO web user interface. (a) There are two search fields, plain text and taxa sequences. (b-d) three associations tabs each one presenting associations of the querred entity with the respective entities, Environments (b), Biological Process (c) and Molecular Function (d). Three channels 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.

abstracts or excerpts of full text with the associated entities highlighted (Figure 3.3a). Additionally, the Environmental Samples and Genome annotations and Isolates channels support evidence for each association by providing links to more detailed information. In the former channel, when the users click on an association, they are redirected to pertinent sample pages of MGnify (Figure 3.3b). Similarly, the latter redirects users to JGI and NCBI genomes when the associations originated from JGI—IMG and Struo, respectively (Figure 3.3c).

The *Documents* tab includes a list of scientific publications where the queried entity is mentioned. Through the *Downloads* tab, users are able to get all of the PREGO associations found for their query, per entity type (e.g., all the environments found related to an organism) and per channel (e.g., all the Environments found related to an organism

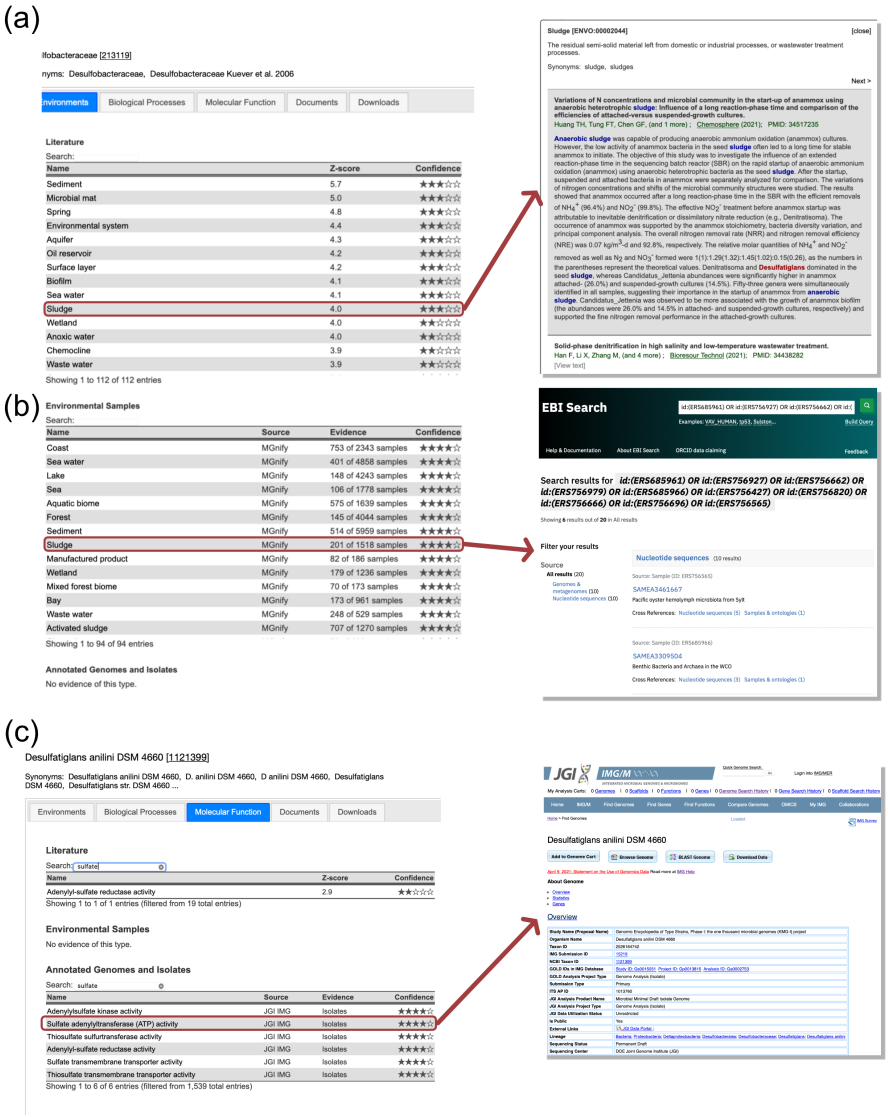


FIGURE 3.3: Each association is supported by original data. (a) Literature channel has a pop-up functionality that displays the scientific articles that each specific association occurs with highlighted color. (b) Environmental Samples channel redirects to the samples that support the specific association (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).

through the *Literature* channel). This data retrieval functionality is also available via the PREGO API (syntax described in Figure 3.4). Finally, all PREGO associations are available for bulk download from each channel (see Table A.1).

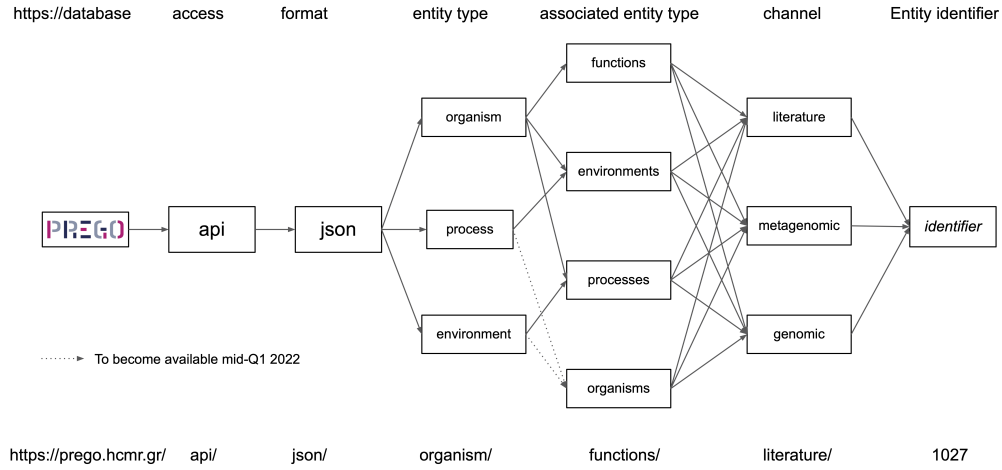


FIGURE 3.4: The PREGO API schema.

3.4.2 PREGO in Action

To demonstrate PREGO's potential, we present four different ways that PREGO can assist molecular ecologists. The demo focuses on the sulfate-reducing microorganisms (SRMs) as well as the processes and environments that relate to sulfate reduction. Through this demo, we highlight how the different channels may provide complementary insights regarding different taxonomic levels and different association types.

Which Environments Are Related to a Taxon?

Based on Pavloudi et al. (2017) [Pavloudi et al., 2017b], several bacterial and archaeal SRM were found in lagoonal sediments, after amplifying and sequencing the dissimilatory sulfite reductase β-subunit (dsrB). Using PREGO for the case of Desulfobacteraceae, the family in which the majority of the observed OTUs of the study belonged to, several environmental types similar to lagoons were retrieved from both the *Literature* and the *Environmental samples* channels (Figure 3.3a,b). Moreover, most of them had a high z-score, such as *"sediment"*, *"sludge"*, and *"activated sludge"*. Several dissimilar environmental types were associated with Desulfobacteraceae, e.g., *"oil reservoir"* indicating them as potential environments where sulfate reduction takes place. However, the presence of taxa within that family in different environments, from *"sea water"* to *"forest"* and *"Wastewater treatment plant"*, may suggest that this family has ubiquitous representatives in diverse conditions.

Searching for *Desulfatiglans anilini* (example1[4], accessed on 24 December 2021) at the species level, the most abundant species in Pavloudi et al. (2017) and, for *Desulfatiglans anilini* DSM 4660 strain (example 2[5], accessed on 24 December 2021), PREGO

[4] https://prego.hcmr.gr/example1
[5] https://prego.hcmr.gr/example2

provides associations with the *"Anaerobic sediment"*, *"Marine oxygen minimum zone"*, and *"Anaerobic digester sludge"* terms. These associations further corroborate the relationship between the species and sulfate reduction. More specifically, the *"sulfur spring"* ENVO term was retrieved from the Environmental samples channel as well.

Which Biological Processes and Molecular Functions Are Related to a Taxon?

According to Pavloudi et al. (2017), *Desulfatiglans anilini* plays an important role in sulfate reduction. The Biological Processes provided by PREGO's Literature channel are the GO terms *"Sulfate reduction"*, *"Sulfide oxidation"*, and *"Sulfide ion homeostasis"*, which support this claim. In addition, the *"Denitrification pathway"* term was also retrieved. This is rather interesting as it is in line with what Pavloudi et al. (2017) discussed about the SRMs and their ability to use various electron acceptors, e.g., nitrate and nitrite.

Furthermore, PREGO's Molecular Function tab provides more insight on this example. Several GO terms related to sulfate reduction (e.g., terms related to *"sulfite reductase"*) were associated with DSM 4660 strain and Desulfatiglans anilini species in multiple channels. Interestingly, in the case of the strain query, the Annotated Genomes channel returned many GO terms related to the nitrogen fixation, e.g., *"nitric oxide dioxygenase activity"*.

Which Taxa Are Related to a Biological Process?

PREGO can be also used to report organisms that relate to a certain biological process. Searching for *"dissimilatory sulfate reduction"* associations with taxa (example 3[6], accessed on 24 December 2021) resulted in several taxa that were mentioned in the Pavloudi et al. (2017) study. For example, taxa such as Thermodesulfobacteria and Thermodesulfovibrio were found among the entries with the highest score (e.g.,) based on the Literature channel. The other two channels did not contain any associations. Using the *"Sulfate assimilation"* (example 4[7], accessed on 24 December 2021) as the biological process input, PREGO results showed several genera that were missing from PREGO results concerning the *"dissimilatory sulfate reduction"*. Hence, manual search of GObp terms that describe the actual biological process of interest is more insightful.

Are There Any Associations between Environments and Biological Processes?

Are there other environmental types, except the lagoonal sediments, in which sulfate assimilation occurs? In that question, and in *"dissimilatory sulfate reduction"* (example 3) in particular, PREGO assigns the highest score to "sediment" while, among others, *"anoxic water"*, *"oil reservoir"*, *"mud volcano"*, and *"basalt"* are potentially associated with environments related to sulfate reduction.

Inversely, PREGO is insightful about occurring processes in a specific environmental type. For example, searching for the biological processes that take place in *"basalt"* (example 5[8], accessed on 24 December 2021), processes like *"Nitrogen fixation"* and

[6] https://prego.hcmr.gr/example3
[7] https://prego.hcmr.gr/example4
[8] https://prego.hcmr.gr/example5

"Reactive nitrogen species metabolic process" stand out. However, sulfate reduction is not among the associations. However, when asking for *"Mafic lava"* (example 6[9], accessed on 24 December 2021), both the *"nitrogen fixation"* and *"Sulfur compound metabolic process"* terms are returned. This highlights the suggestions of Pavloudi et al. (2017), regarding the potential use of various electron acceptors from the different strains present in different environmental types.

3.4.3 PREGO Contents

PREGO contains the literature, environmental samples, and genome annotations of the resources shown in Table 3.1. The extracted contents of these resources have resulted to a knowledge base with 364 K distinct taxonomic groups (out of a pool of 620K Bacteria, Archaea, and microbial eukaryotes, based on NCBI Taxonomy) from which 258K are at the species level (Table 3.2). These taxa are associated with 1 K Environment Ontology terms, 15 K GObp terms, and with 7.9 K GOmf terms. Combining the above, PREGO maintains a knowledge base of entities and associations between them that form a multipartite network with entities as nodes and scored associations between them as weighted links.

As shown in Figure 3.5, in its current version (December 2021), PREGO knowledge base covers 157 bacterial phyla (107 are Candidatus), 23 phyla from archaea (18 are Candidatus), and 22 unicellular eukaryotic phyla described in the NCBI Taxonomy database. The number of bacterial taxa present among the associations of each phylum ranges from the order of 10s, as in the case of *Candidatus Coatesbacteria*, to hundreds of thousands, e.g., Actinobacteriae. The number of environmental types, found among the PREGO associations for each phylum, ranges from just a few to up to 1000. Similarly, the number of biological processes that have been related to the various phyla may range from less than a dozen, e.g., Yanofskybacteria to up to several thousands, e.g., Bacteroidetes. On the contrary, the number of molecular functions found to be related to taxa of each phylum is rather constant in all three domains.

3.5 Discussion

3.5.1 PREGO Contents

On its current version and according to the NCBI Taxonomy that it is based on, PREGO manages to cover a great range of microbial taxa, as most (if not all phyla) are present in the knowledge base (Figure 3.5). The different number of organisms' entities per phylum highlights the diverse number of the members of the various phyla. On the contrary, the similar number of molecular functions in all cases indicates the robustness of the main metabolic processes required for life. With respect to biological processes, their number per phylum varies to some extent, especially for the case of Bacteria and Archaea. That could be observed as, in many cases, phyla that have been recently described using molecular techniques have not been studied extensively yet, e.g., Candidatus Delongbacteria. As expected, the number of environmental types that have been associated with members

[9] https://prego.hcmr.gr/example6

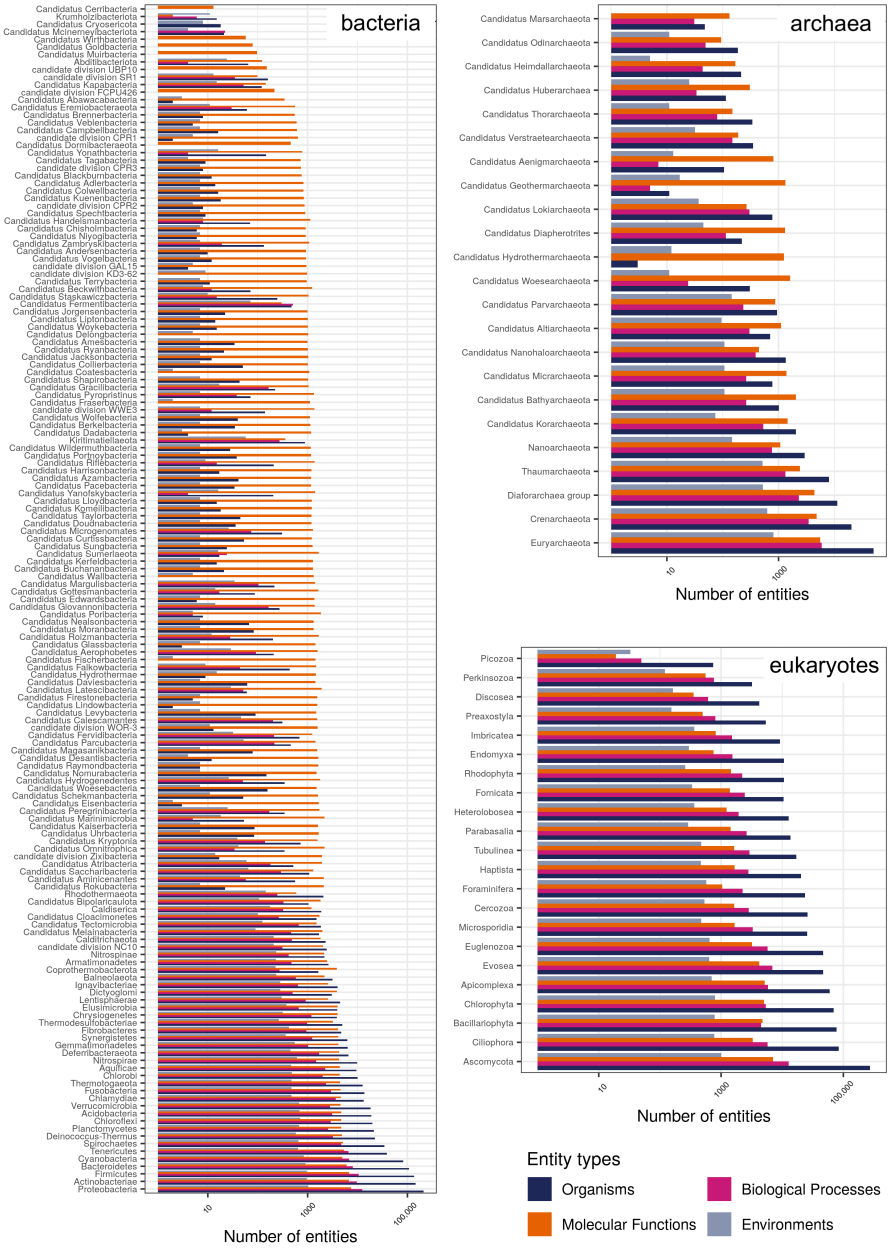


FIGURE 3.5: Summary of all the unique entities per phylum for each of the four entity types (in log10 scale) that appear in PREGO. Phyla are grouped based on their superkingdom (in log10 scale). Only phyla for which associations are available in the PREGO platform are mentioned.

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).

of each phylum varies, as a phylum may be universally present, while others could be strongly niche-specific (e.g., Hydrothermarchaeota).

Because of its three different channels, PREGO manages to extract associations both in the species and higher taxonomic levels. The Isolates channel supports explicit associations at the species level (Table 3.3 and Figure S3). Interestingly, the number of such genomes seems to have reached a plateau for now, as PREGO-like platforms include the same order of magnitude. The *Literature* channel, on the other hand, promotes the extraction of associations at higher taxonomic levels (Table 3.3 and Figure S1). This also applies to environment—organisms associations derived from the Environmental Samples channel (Table 3.3 and Figure S2). Associations regarding biological processes, though, are strongly enhanced by the Literature channel and the massive increase of literature.

Additionally, the text mining methodology of the Literature channel has retrieved most

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

TABLE 3.3: The associations between entities of PREGO after co-occurrence analysis: The supported entity types of associations are Environments—Biological Processes, Environments—Molecular Functions, Taxa—Environments, Taxa—Biological Processes, Taxa—Molecular Functions.

of the entities present in PREGO knowledge base (Table 3.2). A significant contribution to the taxa with associations is due to the PMC OA processing by the text mining pipeline of the Literature channel. This is in-line with reports in other applications of text mining when using full text articles [Westergaard et al., 2018]. However, the resulting associations are suggestive because of the text mining nature, and therefore subject for further review by the users.

3.5.2 Related Tools' Functionality and Content

There is an emerging niche for tools similar to PREGO to bring forward microbe associations and metadata. Table 3.4 summarizes the common and different features of BacDive, WoM, NMDC data portal, and PREGO. All of them commonly share the environmental associations and biological/metabolic processes with the microbes.

BacDive is a well-established platform with a focus on phenotype and cultivation information for about 100,000 prokaryotes, bacteria, and archaea. It has a high level of curation for most of its input types, like literature, internal databases, and personal collections. The NMDC data portal has published the scheme, the user interface, and a demonstrative collection of samples that will be populated later on. Standout features are the spatial visualization with coordinates and the detailed information of the samples, e.g., sequencing instruments and methodology. An alternative approach is facilitated by WoM, which aims to bind chemistry to microbes. An environment, in particular, is defined as the starting metabolite pool that is transformed by an organism. Another tool is The Microbe Directory that contains fully curated metadata for about 8000 microbes from all superkingdoms. This tool focuses on conditions of growth and on host taxa.

Complementary to these tools, PREGO contains associations of bacteria, archaea, and eukaryotes. Distinctive features are the associations of environments with processes/functions and the large-scale literature integration with text mining. Most importantly, most of the tools are complementary to each other with minimum overlap, an indication of the opportunities for further innovative synergies.

3.5.3 PREGO Next Steps

PREGO is a user-friendly association mining and sharing platform. Its modular web-architecture grants it the flexibility for further improvements in the aforementioned aspects, namely: source datasets, user interface, entity, and association scope expansion. Regarding datasets, additional data, such as transcriptomes from MGnify and other records annotated with metadata from studies in EuroPMC, accessed on 24 December 2021) [Ferguson et al., 2021], could be incorporated. Similarly, the NMDC data platform standards-compliant annotated records[10] (accessed on 24 December 2021) could serve as an additional resource with its high-quality metadata [Wood-Charlson et al., 2020, Vangay et al., 2021]. Reciprocally, if requested, pertinent literature and association summaries could be programmatically offered to interested third parties.

Furthermore, the entity types supported by the PREGO system could be expanded. For example, GOmf terms could be upgraded as a search-entry point to the system. In ad-

[10] https://data.microbiomedata.org/

Functionality	BacDive	Web of Microbes	NMDC	PREGO
manual curation	high	high	intermediate	low
literature integration	limited	no	no	yes
environment—taxa associations	yes	yes	yes	yes
environment—process/ function associations	no	no	no	yes
process/function—taxa associations	yes	yes	yes	yes
phenotypic data	yes	no	no	no
data origin	original integration	original	original integration	integration
spatial coordinates	yes	no	yes	no
application programming interface	yes	no	yes	yes
bulk download	limited	yes	yes	yes

TABLE 3.4: Feature comparison among platforms that facilitate knowledge discovery and integration of microbial data.

dition, disease and tissue describing terms, already supported by the PREGO-underlying EXTRACT system [Pafilis et al., 2016], could enter the PREGO ecosystem of associated entities. From a statistics perspective, the calculation of a combined association score, when an association is reported by more than one channel of information, could be another feature to add.

The user interface can be enhanced to support multiple entity and/or sequence queries, instead of single ones. Sequences can be processed by taxonomy assignment pipelines (e.g., PEMA [Zafeiropoulos et al., 2020]) and be converted into searching PREGO for associations. In addition, network visualization tools, like Arena3Dweb [Karatzas et al., 2021], could allow interactive browsing of associations through multi-layered graphs. Enrichment analyses, like those performed by OnTheFly2.0 [Baltoumas et al., 2021b] or Flame [Thanati et al., 2021], can be incorporated. Omics data analysis pipelines, like MiBiOmics [Zoppi et al., 2021], environment associations with sequences using SeqEnv [Sinclair et al., 2016] and biogeochemical associations with metagenomic data with DiTing [Xue et al., 2021] could be enabled by comparing the associations pertinent to different groups of entities. The computationally intensive tasks of multiple queries, taxonomy assignments to sequences and enrichment analysis could be offered by our in-house High Performance Computing facility (https://hpc.hcmr.gr/, accessed on 24 December 2021) [Zafeiropoulos et al., 2021c] in synergy with pertinent Research Infrastructures like ELIXIR[11] (accessed on 24 December 2021) and LifeWatch ERIC[12] (accessed on 24 December 2021).

[11] https://elixir-europe.org
[12] https://www.lifewatch.eu/

Availability of Supporting Source Codes:

The PREGO software modules are available under BSD 2-Clause “Simplified” License. Scripts, where additional libraries have been used, are subject to their individual licenses. More information on each module can be found as listed below:

- prego_gathering_data github.com/lab42open-team/prego_gathering_data
- prego_daemons github.com/lab42open-team/prego_daemons
- prego_mappings github.com/lab42open-team/prego_mappings
- prego_statistics github.com/lab42open-team/prego_statistics

Additional software and curated lists along with their individual license are:

- tagger: https://github.com/larsjuhljensen/tagger, BSD 2-Clause "Simplified" License
- mamba: https://github.com/larsjuhljensen/mamba, BSD 2-Clause "Simplified" License
- tagger dictionary: https://download.jensenlab.org/ and there in: https://download.jensenlab.org/prego_dictionary.tar.gz, CC-BY 4.0 license

Chapter 4

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

Citation: Chalkis, A., Fisikopoulos, V., Tsigaridas E. and Zafeiropoulos H. Geometric Algorithms for Sampling the Flux Space of Metabolic Networks. 37th International Symposium on Computational Geometry (SoCG 2021)
DOI: 10.4230/LIPIcs.SoCG.2021.21 [1]

4.1 Abstract

Systems Biology is a fundamental field and paradigm that introduces a new era in Biology. The crux of its functionality and usefulness relies on metabolic networks that model the reactions occurring inside an organism and provide the means to understand the underlying mechanisms that govern biological systems. Even more, metabolic networks have a broader impact that ranges from resolution of ecosystems to personalized medicine.

The analysis of metabolic networks is a computational geometry oriented field as one of the main operations they depend on is sampling uniformly points from polytopes; the latter provides a representation of the steady states of the metabolic networks. However, the polytopes that result from biological data are of very high dimension (to the order of thousands) and in most, if not all, the cases are considerably skinny. Therefore, to perform uniform random 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 in metabolic networks. Its backbone is a Multiphase Monte Carlo Sampling (MMCS) algorithm that unifies rounding and sampling in one pass, obtaining both upon termination. It exploits an improved variant of the Billiard Walk that enjoys faster arithmetic complexity per step. We demonstrate the efficiency of our approach by performing extensive experiments on various metabolic networks. Notably, sampling on the most complicated human metabolic

[1] Authors' names are in alphabetical order. This is a modified version of the published version, in terms of relevance, coherence and formatting. Proofs for the lemmas mentioned and parameter tuning can be found in the original publication. The dingo Python library, a wrapper of the C++ code of the MMCS algorithm, is available at https://github.com/geomscale/dingo and a relative publication is under preparation.

network accessible today, Recon3D, corresponding to a polytope of dimension 5 335, took less than 30 hours. To our knowledge, that is out of reach for existing software.

4.2 Introduction

4.2.1 The field of Systems Biology

Systems Biology establishes a scientific approach and a paradigm. As a research approach, it is the qualitative and quantitative study of the systemic properties of a biological entity along with their ever evolving interactions [Klipp et al., 2016, Kohl et al., 2010]. By combining experimental studies with mathematical modeling it analyzes the function and the behavior of biological systems. In this setting, we model the interactions between the components of a system to shed light on the system's *raison d'être* and to decipher its underlying mechanisms in terms of evolution, development, and physiology [Ideker et al., 2001].

Initially, Systems Biology emerged as a need. New technologies in Biology accumulate vast amounts of information/data from different levels of the biological organization, i.e., genome, transcriptome, proteome, metabolome [Quinn et al., 2016]. This leads to the emerging question *"what shall we do with all these pieces of information"?* The answer, if we consider Systems Biology as a paradigm, is to move away from reductionism, still the main conceptual approach in biological research, and adopt holistic approaches for interpreting how a system's properties emerge [Noble, 2008]. The following diagram provides a first, rough, mathematical formalization of this approach.

$$\textit{components} \rightarrow \textit{networks} \rightarrow \textit{in silico models} \rightarrow \textit{phenotype}$$ [Palsson, 2015].

Systems Biology expands in all the different levels of living entities, from the molecular, to the organismal and ecological level. The notion that penetrates all levels horizontally is *metabolism*; the process that modifies molecules and maintains the living state of a cell or an organism through a set of chemical reactions [Schramski et al., 2015]. The reactions begin with a particular molecule which they convert into some other molecule(s), while they are catalyzed by enzymes in a key-lock relationship. We call the quantitative relationships between the components of a reaction *stoichiometry*. Linked reactions, where the product of the first acts as the substrate for the next, build up metabolic pathways. Each pathway is responsible for a certain function. We can link together the aggregation of all the pathways that take place in an organism (and their corresponding reactions) and represent them mathematically using the reactions' stoichiometry. Therefore, at the species level, metabolism is a network of its metabolic pathways and we call these representations *metabolic networks*.

4.2.2 From metabolism to computational geometry

The complete reconstruction of the metabolic network of an organism is a challenging, time consuming, and computationally intensive task; especially for species of high level of complexity such as *Homo sapiens*. Even though sequencing the complete genome of

a species is becoming a trivial task providing us with quality insight, manual curation is still mandatory and large groups of researchers need to spend a great amount of time to build such models [Thiele and Palsson, 2010]. However, over the last few years, automatic reconstruction approaches for building genome-scale metabolic models [Machado et al., 2018] of relatively high quality have been developed. Either way, we can now obtain the metabolic network of a bacterial species (single cell species) of a tissue and even the complete metabolic network of a mammal. Biologists are also moving towards obtaining such networks for all the species present in a microbial community. This will allow us to further investigate the dynamics, the functional profile, and the inter-species reactions that occur. Using the stoichiometry of each reaction, which is always the same in the various species, we convert the metabolic network of an organism to a mathematical model. Thus, the metabolic network becomes an *in silico* model of the knowledge it represents.

In metabolic networks analysis mass and energy are considered to be conserved [Palsson, 2009]. As many homeostatic states, that is steady internal conditions [Shishvan et al., 2018], are close to steady states (where the production rate of each metabolite equals its consumption rate [Cakmak et al., 2012]) we commonly use the latter in metabolic networks analysis.

Stoichiometric coefficients are the number of molecules a biochemical reaction consumes and produces. The coefficients of all the reactions in a network, with m metabolites and n reactions ($m < n$), form the *stoichiometric matrix* $S \in \mathbb{R}^{m \times n}$ [Palsson, 2015]. The nullspace of S corresponds to the steady states of the network:

$$S \cdot x = 0, \tag{4.1}$$

where $x \in \mathbb{R}^n$ is the *flux vector* that contains the fluxes of each chemical reaction of the network. Flux is the rate of turnover of molecules through a metabolic pathway.

All physical variables are finite, therefore the flux (and the concentration) is bounded [Palsson, 2015]; that is for each coordinate x_i of the x, there are $2n$ constants $x_{ub,i}$ and $x_{lb,i}$ such that $x_{lb,i} \leq x_i \leq x_{ub,i}$, for $i \in [n]$. We derive the constraints from explicit experimental information. In cases where there is no such information, reactions are left unconstrained by setting arbitrary large values to their corresponding bounds according to their reversibility properties; i.e., if a reaction is reversible then its flux might be negative as well [Lularevic et al., 2019]. The constraints define a n-dimensional box containing both the steady and the dynamic states of the system. If we intersect that box with the nullspace of S, then we define a polytope that encodes all the possible steady states and their flux distributions [Palsson, 2015]. We call it the steady-state *flux space*. Figure 4.1 illustrates the complete workflow from building a metabolic network to the computation of a flux distribution.

Using the polytopal representation, a commonly used method for the analysis of a metabolic network is Flux Balance Analysis (FBA) [Orth et al., 2010]. FBA identifies a single optimal flux distribution by optimizing a linear objective function over a polytope [Orth et al., 2010]. Unfortunately, this is a *biased* method because it depends on the selection of the objective function. To study the global features of a metabolic network we need *unbiased methods*. To obtain an accurate picture of the whole solution space we exploit

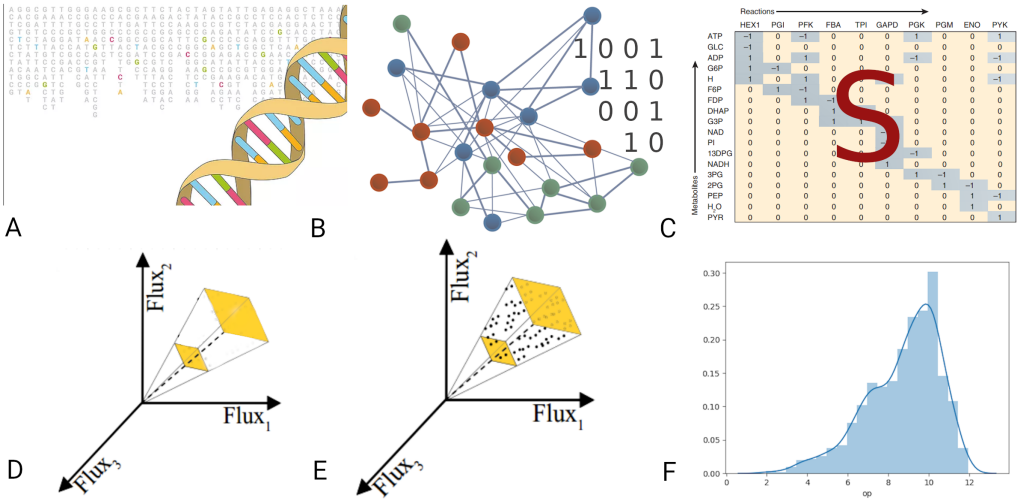


FIGURE 4.1: From DNA sequences to distributions of metabolic fluxes. (A) The genes of an organism provide us with the enzymes that it can potentially produce. Enzymes are like a blueprint for the reactions they can catalyze. (B) Using the enzymes we identify the reactions in the organism. (C) We construct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

sampling techniques [Schellenberger and Palsson, 2009]. If collect a sufficient number of points uniformly distributed in the interior of the polytope, then the biologists can study the properties of certain components of the whole network and deduce significant biological insights [Palsson, 2015]. Therefore, efficient sampling tools are of great importance.

4.2.3 Metabolic networks through the lens of random sampling

Efficient uniform random sampling on polytopes resulting from metabolic networks is a very challenging task both from the theoretical (algorithmic) and the engineering (implementation) point of view. First, the dimension of the polytopes is of the order of certain thousands. This requires, for example, advanced engineering techniques to cope with memory requirements and to perform linear algebra operations with large matrices; e.g., in Recon3D [Brunk et al., 2018] we compute the null space of a $8\,399 \times 13\,543$ matrix. Second, the polytopes are rather skinny (Section 4.5); this makes it harder for sampling algorithms to move in the interior of polytopes and calls for novel practical techniques to sample.

There is extended on-going research concerning advanced algorithms and implementations for sampling metabolic networks over the last decades. Markov Chain Monte Carlo algorithms such as Hit-and-Run (HR) [Smith, 1984] have been widely used to address the

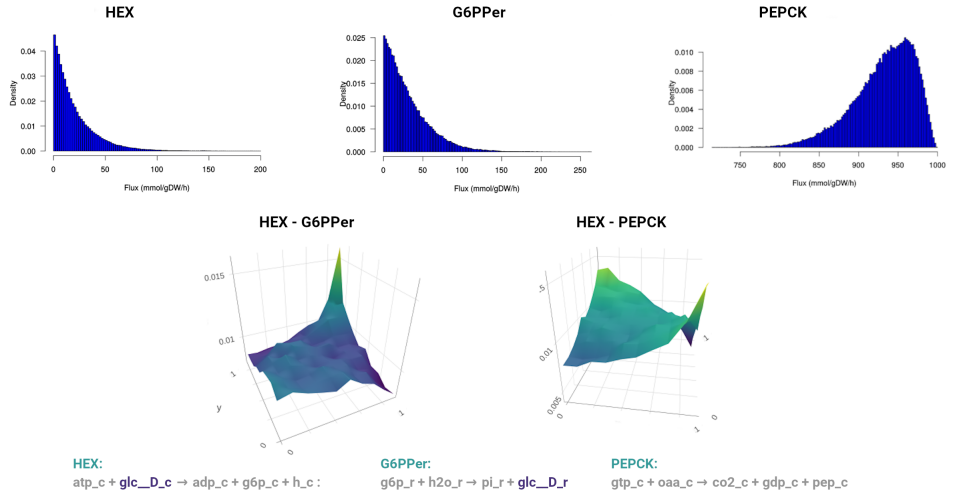


FIGURE 4.2: Flux distributions in the most recent human metabolic network Recon3D [Brunk et al., 2018]. We estimate the flux distributions of the reactions catalyzed by the enzymes Hexokinase (D-Glucose:ATP) (HEX), Glucose-6-Phosphate Phosphatase, Edoplasmic Reticular (G6PPer) and Phosphoenolpyruvate 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.

challenges of sampling. Two variants of HR are the non-Markovian Artificial Centering Hit-and-Run (ACHR) [Kaufman and Smith, 1998] that has been widely used in sampling metabolic models, e.g., [Saa and Nielsen, 2016], and Coordinate Hit-and-Run with Rounding (CHRR) [Haraldsdóttir et al., 2017]. The latter is part of the cobra toolbox [Heirendt et al., 2019], the most commonly used software package for the analysis of metabolic networks. CHRR enables sampling from complex metabolic networks corresponding to the highest dimensional polytopes so far. There are also stochastic formulations where the inclusion of experimental noise in the model makes it more compatible with the stochastic nature of biological networks [MacGillivray et al., 2017]. The recent study in [Fallahi et al., 2020] offers an overview as well as an experimental comparison of the currently available samplers.

These implementations played a crucial role in actually performing in practice uniform sampling from the flux space. However, they are currently limited to handle polytopes of dimension say $\leq 2\,500$ [Fallahi et al., 2020, Haraldsdóttir et al., 2017]. This is also the order of magnitude of the most complicated, so far, metabolic network model built,

Recon3D [Brunk et al., 2018]. By including 13 543 metabolic reactions and involving 4 140 unique metabolites, Recon3D provides a representation of the 17% of the functionally of annotated human genes. To our knowledge, there is no method that can efficiently handle sampling from the flux space of Recon3D.

Apparently, the dimension of the polytopes will keep rising and not only for the ones corresponding to human metabolic networks. Metabolism governs systems biology at all its levels, including the one of the community. Thus, we are not only interested in sampling a sole metabolic network, even if it has the challenges of the human. Sampling in polytopes associated to network of networks are the next big thing in metabolic networks analysis and in Systems Biology [Bernstein et al., 2019, Perez-Garcia et al., 2016].

Regarding the sampling process, from the theoretical point of view, we are interested in the convergence time, or *mixing time*, of the Markov Chain, or geometric *random walk*, to the target distribution. Given a d-dimensional polytope P, the mixing time of several geometric random walks (e.g., HR or Ball Walk) grows quadratically with respect to the sandwiching ratio R/r of the polytope [Lovász et al., 1997, Lovász and Vempala, 2006]. Here r and R are the radii of the smallest and largest ball with center the origin that contains, and is contained, in P, respectively; i.e., $rB_d \subseteq P \subseteq RB_d$, where B_d is the unit ball. It is crucial to reduce R/r, i.e., to put P in well a rounded position where $R/r = \widetilde{\mathcal{O}}(\sqrt{d})$; the $\widetilde{\mathcal{O}}(\cdot)$ notation means that we are ignoring polylogarithmic factors. A powerful approach to obtain well roundness is to put P in *near isotropic position*. In general, $K \subset \mathbb{R}^d$ is in isotropic position if the uniform distribution over K is in isotropic position, that is $\mathbb{E}_{X\sim K}[X] = 0$ and $\mathbb{E}_{X\sim K}[X^T X] = I_d$, where I_d is the $d \times d$ identity matrix. Thus, to put a polytope P into isotropic position one has to generate a set of uniform points in its interior and apply to P the transformation that maps the point-set to isotropic position; then iterate this procedure until P is in c-isotropic position [Cousins and Vempala, 2016, Lovász and Vempala, 2006], for a constant c. In [Adamczak et al., 2010] they prove that $\mathcal{O}(d)$ points suffice to achieve 2-isotropic position. Alternatively in [Haraldsdóttir et al., 2017] they compute the maximum volume ellipsoid in P, they map it to the unit ball, and then apply to P the same transformation. They experimentally show that a few iterations suffice to put P in John's position [John, 2014]. Moreover, there are a few algorithmic contributions that combine sampling with distribution isotropization steps, e.g., the multi-point walk [Bertsimas and Vempala, 2004] and the annealing schedule [Kalai and Vempala, 2006].

An important parameter of a random walk is the walk length, i.e., the number of the intermediate points that a random walk visits before producing a single sample point. The longer the walk length of a random walk is, the smaller the distance of the current distribution to the stationary (target) distribution becomes. For the majority of random walks there are bounds on the walk length to bound the mixing time with respect to a statistical distance. For example, HR generates a sample from a distribution with total variation distance less than ϵ from the target distribution after $\widetilde{\mathcal{O}}(d^3)$ [Lovász and Vempala, 2006] steps, in a well rounded convex body and for log-concave distributions. Similarly, CDHR mixes after a polynomial, in the diameter and the dimension, number of steps [Laddha and Vempala, 2020, Narayanan and Srivastava, 2020] for the case of uniform distribution. However, extended practical results have shown that both CDHR and HR converges after $\mathcal{O}(d^2)$ steps [Chalkis et al., 2020, Cousins and Vempala, 2016, Haraldsdóttir

et al., 2017]. The leading algorithms for uniform polytope sampling are the Riemannian Hamiltonian Monte Carlo sampler [Lee and Vempala, 2018] and the Vaidya walk [Chen et al., 2018], with mixing times $\widetilde{\mathcal{O}}(md^{2/3})$ and $\widetilde{\mathcal{O}}(m^{1/2}d^{3/2})$ steps, respectively. However, it is not clear if these random walks can outperform CDHR in practice, because of their high cost per step and numerical instability.

Billiard Walk (BW) [Gryazina and Polyak, 2014] is a random walk that employs linear trajectories in a convex body with boundary reflections; alas with an unknown mixing time. The closest guarantees for its mixing time are those of HR and stochastic billiards [Dieker and Vempala, 2015]. Interestingly, [Gryazina and Polyak, 2014] shows that, experimentally, BW converges faster than HR for a proper tuning of its parameters. The same conclusion follows from the computation of the volume of zonotopes [Chalkis et al., 2020]. It is not known how the sandwiching ratio of P affects the mixing time of BW. Since BW employs reflections on the boundary, we can consider it as a special case of Reflective Hamiltonian Monte Carlo [Chevallier et al., 2018].

For almost all random walks the theoretical bounds on their mixing times are pessimistic and unrealistic for computations. Hence, if we terminate the random walk earlier, we generate samples that are usually highly correlated. There are several *MCMC Convergence Diagnostics* [Roy, 2020] to check if the quality of a sample can provide an accurate approximation of the target distribution. For a dependent sample, a powerful diagnostic is the *Effective Sample Size* (ESS). It is the number of effectively independent draws from the target distribution that the Markov chain is equivalent to. For autocorrelated samples, ESS bounds the uncertainty in estimates [Geyer, 1992] and provides information about the quality of the sample. There are several statistical tests to evaluate the quality of a generated sample, e.g., potential scale reduction factor (PSRF) [Gelman and Rubin, 1992], maximum mean discrepancy (MMD) [Gretton et al., 2012], and the uniform tests [Cousins, 2017]. Interestingly, the copula representation we employ in Figure 4.2 to capture the dependence between two fluxes of reactions was also used successfully in a geometric framework to detect financial crises capturing the dependence between portfolio return and volatility [Calès et al., 2018].

4.3 Contribution

We introduce a Multi-phase Monte Carlo Sampling (MMCS) algorithm (Section 4.4.2 and Algorithm 2) 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. This improves the efficiency of the random walk in the next phase, see Figure 4.3. For sampling, we propose an improved variant of Billiard Walk (BW) (Section 4.4.1 that enjoys faster arithmetic complexity per step. We also handle efficiently the potential arithmetic inaccuracies near to the boundary, see [Chevallier et al., 2018]. 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[2] provides a complete software framework to handle efficiently sampling 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 as cobra (Section 4.5). 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 (Table 4.1). The quality of the sample our software produces is measured with two widely used diagnostics, i.e., ESS and potential scale reduction factor (PSRF) [Gelman and Rubin, 1992]. 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 4.2 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. Last, MMCS algorithm is quite general sampling scheme and so it has the potential to address other hard computational problems like multivariate integration and volume estimation of polytopes.

4.4 Methods & Implementation

4.4.1 Efficient Billiard walk

The geometric random walk of our choice to sample from a polytope is based on Billiard Walk (BW) [Gryazina and Polyak, 2014], which we modify to reduce the per-step cost.

For a polytope $P = \{x \in \mathbb{R}^d \mid Ax \leq b\}$, where $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$, BW starts from a given point $p_0 \in P$, selects uniformly at random a direction, say v_0, and it moves along the direction of v_0 for length L; it reflects on the boundary if necessary. This results a new point p_1 inside P. We repeat the procedure from p_1. Asymptotically it converges to the uniform distribution over P. The length is $L = -\tau \ln \eta$, where η is a uniform number in $(0,1)$, that is $\eta \sim \mathcal{U}(0,1)$, and τ is a predefined constant. It is useful to set a bound, say ρ, on the number of reflections to avoid computationally hard cases where the trajectory may stuck in corners. In [Gryazina and Polyak, 2014] they set $\tau \approx diam(P)$ and $\rho = 10d$. Our choices for τ and ρ depend on a burn-in step that we detail in Section 4.5.

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

[2] https://github.com/GeomScale/volume_approximation/tree/socg21

compute the point $\partial P \cap \ell$ where the first reflection of a BW step takes place, we solve the following m linear equations

$$a_j^T(p_0 + t_j v_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T v_0,\ j \in [k], \tag{4.2}$$

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 $\mathcal{O}(dk)$. A straightforward approach for BW 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* by storing computations from the previous iteration combined with a preprocessing step. The preprocessing step involves the normal vectors of the facets, that takes $m^2 d$ operations, and the amortized per-step complexity of BW becomes $\mathcal{O}((\rho + d)k)$.

Lemma 1. *The amortized per step complexity of BW is $\mathcal{O}((\rho + d)k)$ after a preprocessing step that takes $\mathcal{O}(k^2 d)$ operations, where ρ is the maximum number of reflections per step.*

The use of floating point arithmetic could result to points outside P due to rounding errors when computing boundary points. To avoid this, when we compute the roots in Equation (4.2) we exclude the facet that the ray hit in the previous reflection.

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 $\partial P \cap \ell = O$ **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

$$a_j^T(p_0 + t_j u_0) = b_j \Rightarrow t_j = (b_j - a_j^T p_0)/a_j^T u_0,\ j \in [k], \tag{4.3}$$

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 1.

4.4.2 Multiphase Monte Carlo Sampling algorithm

To sample steady states in the flux space of a metabolic network, with m metabolites and n reactions, we introduce a Multiphase Monte Carlo Sampling (MMCS) algorithm; it is multiphase because it consists of a sequence of sampling phases.

Let $S \in \mathbb{R}^{m\times n}$ be the stoichiometric matrix and $x_{lb},\ x_{ub} \in \mathbb{R}^n$ bounds on the fluxes. The flux space is the bounded convex polytope

$$\text{FS} := \{x \in \mathbb{R}^n \mid Sx = 0,\ x_{lb} \le x \le x_{ub}\} \subset \mathbb{R}^n. \tag{4.4}$$

The dimension, d, of FS is smaller than the dimension of the ambient space; that is $d \le n$. To work with a full dimensional polytope we restrict the box induced by the inequalities $x_{lb} \le x \le x_{ub}$ to the null space of S. Let the H-representation of the box be $\left\{x \in \mathbb{R}^n \,\middle|\, \begin{pmatrix} I_n \\ -I_n \end{pmatrix} x \le \begin{pmatrix} x_{ub} \\ x_{lb} \end{pmatrix}\right\}$, where I_n is the $n \times n$ identity matrix, and let $N \in \mathbb{R}^{n\times d}$ be the matrix of the null space of S, that is $SN = 0_{m\times d}$. Then $P = \{x \in \mathbb{R}^d \mid Ax \le b\}$, where $A = \begin{pmatrix} I_n N \\ -I_n N \end{pmatrix}$ and $b = \begin{pmatrix} x_{ub} \\ x_{lb} \end{pmatrix} N$, is a full dimensional polytope (in $\mathbb{R}^d$). After we sample (uniformly) points from P, we transform them to uniformly distributed points (that is steady states) in FS by applying the linear map induced by N.

MMCS generates, in a sequence of sampling phases, a set of points, that is almost equivalent to n independent uniformly distributed points in P, where n is given. At each phase, it employs Billiard Walk (Section 4.4.1) to sample approximate uniformly distributed points, rounding to speedup sampling, and uses the Effective Sample Size (ESS) diagnostic to decide termination. The pseudo-code of the algorithm appears in

Algorithm 2.

Overview.

Initially we set $P_0 = P$. At each phase $i \geq 0$ we sample at most λ points from P_i. We generate them in chunks; we also call them *chain* of sampling points. Each chain contains at most l points (for simplicity consider $l = \mathcal{O}(1)$). To generate the points in each chain we employ BW, starting from a point inside P_i; the starting point is different for each chain. We repeat this procedure until the total number of samples in P_i reaches the maximum number λ; we need $\frac{\lambda}{l}$ chains. To compute a starting point for a chain, we pick a point uniformly at random in the Chebychev ball of P_i and we perform $\mathcal{O}(\sqrt{d})$ burn-in BW steps to obtain a warm start.

After we have generated λ sample points we perform a rounding step on P_i to obtain the polytope of the next phase, P_{i+1}. We compute a linear transformation, T_i, that puts the sample into isotropic position and then $P_{i+1} = T_i(P_i)$. The efficiency of BW improves from one phase to the next one because the sandwiching ratio decreases and so the average number of reflections decreases and thus the convergence to the uniform distribution accelerates (Section 4.5). That is we obtain faster a sample of better quality. Finally, the (product of the) inverse transformations maps the samples to $P_0 = P$. Figure 4.3 depicts the procedure.

Termination.

There are no bounds on the mixing time of BW [Gryazina and Polyak, 2014], hence for termination we rely on ESS. MMCS terminates when the minimum ESS among all the univariate marginals is larger than a requested value. We chose the marginal distributions (of each flux) because they are essential for systems biologists, see [Bordel et al., 2010] for a typical example. In particular, after we generate a chain, the algorithm updates the ESS of each univariate marginal to take into account all the points that we have sampled in P_i, including the newly generated chain. We keep the minimum, say n_i, among all marginal ESS values. If $\sum_{j=0}^{i} n_j$ becomes larger than n before the total number of samples in P_i reaches the upper bound λ, then MMCS terminates. Otherwise, we proceed to the next phase. In summary, MMCS terminates when the sum of the minimum marginal ESS values of each phase reaches n.

Rounding step.

This step is motivated by the theoretical result in [Adamczak et al., 2010] and the rounding algorithms [Lovász and Vempala, 2006, Cousins and Vempala, 2016]. We apply the linear transformation T_i to P_i so that the sandwiching ratio of P_{i+1} is smaller than that of P_i. To find the suitable T_i we compute the SVD decomposition of the matrix that contains the sample row-wise [Artstein-Avidan et al., 2020].

Updating the Effective Sample Size.

The effective sample size of a sample of points generated by a process with autocorrelations ρ_t at lag t is function (actually an infinite series) in the ρ_t's; its exact value is unknown. Following [Geyer, 1992], we efficiently compute ESS employing a finite sum

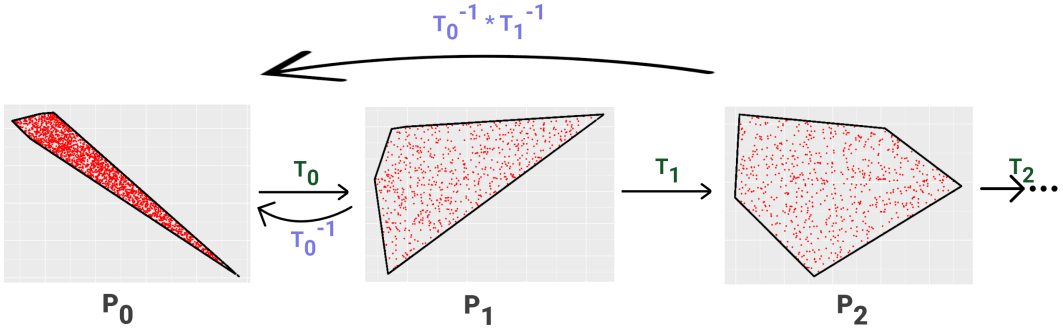


FIGURE 4.3: An illustration of our Multiphase Monte Carlo Sampling algorithm. The method is given an integer n and starts at phase $i = 0$ sampling from P_0. In each phase it samples a maximum number of points λ. If the sum of Effective Sample Size in each phase becomes larger than n before the total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

of monotone estimators $\hat{\rho}_t$ of the autocorrelation at lag t, by exploiting Fast Fourier Transform. Furthermore, given M chains of samples, the autocorrelation estimator $\hat{\rho}_t$ is given by, $\hat{\rho}_t = 1 - \frac{C - \frac{1}{M}\sum_{i=1}^{M} \hat{\rho}_{t,i}}{B}$, where C and B are the within-sample variance estimate and the multi-chain variance estimate given in [Gelman and Rubin, 1992] and $\hat{\rho}_{t,i}$ is an estimator of the autocorrelation of the i-th chain at lag t. To update the ESS, for every new chain of points the algorithm generates, we compute $\hat{\rho}_{t,i}$. Then, using Welford's algorithm we update the average of the estimators of autocorrelation at lag t, as well as the between-chain variance and the within-sample variance estimators given in [Gelman and Rubin, 1992]. Finally, we update the ESS using these estimators.

To update the ESS, for every new chain of points the algorithm generates, we compute the estimator of its autocorrelation. Then, using Welford's algorithm we update the average of the estimators of autocorrelation at lag t, as well as the between-chain variance and the within-sample variance estimators [Gelman and Rubin, 1992]. Finally, we update the ESS using these estimators.

Lemma 2 (Complexity of MMCS per phase)**.** *Let $P = \{x \in \mathbb{R}^d \mid Ax \leq b\}$, where $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$, be a full dimensional polytope in $\mathbb{R}^d$. To sample n points (approximately) uniformly distributed in P, MMCS (Algorithm 2) performs $\mathcal{O}(W(\rho + d)k\lambda + \lambda^2 d + d^3)$ arithmetic operations per phase, where W is the walk length of Billiard Walk, ρ is an upper bound on the number of reflections, and λ and upper bound on the points generated at each phase.*

In Section 4.5 we discuss how to tune the parameters of MMCS to make it more efficient in practice. We also comment on the (practical) complexity of each phase, based on the tuning.

Algorithm 2 Multiphase Monte Carlo Sampling$(P, n, l, \lambda, \rho, \tau, W)$

Require: A full dimensional polytope $P \in \mathbb{R}^d$;
requested effectiveness $n \in \mathbb{N}$ (number of sampled points);
l length of each chain;
λ upper bound of the number of generated points in each phase λ;
upper bound on the number of reflections ρ;
parameter τ to adjust the length of the trajectory; walk length W.

Ensure: a set n of approximate uniformly distributed points $S \in P$

Set $P_0 \leftarrow P$, $sum\_ess \leftarrow 0$, $S \leftarrow \emptyset$, $i \leftarrow 0$, $T_0 = I_d$
while $sum\_ess < n$ **do**
  $sum\_point\_phase \leftarrow 0$, $U \leftarrow \emptyset$
  while $sum\_point\_phase < \lambda$; **do**
    Set $Q \leftarrow \emptyset$; Generate a starting point $q_0 \in P_i$;
    for $j = 1, \ldots, l$ **do**
      $q_j \leftarrow$Billiard_Walk$(P_i, q_{j-1}, \rho, \tau, W)$, Store the point q_j to the set Q
    end for
    $S \leftarrow S \cup T_i^{-1}(Q)$, $U \leftarrow U \cup Q$, $sum\_point\_phase \leftarrow sum\_point\_phase + l$ Update ESS n_i of this phase
    if $sum\_ess + n_i \geq n$ **then**
      break
    end if
  end while
  $sum\_ess \leftarrow sum\_ess + n_i$, Compute T such that $T(U)$ is in isotropic position, $P_{i+1} \leftarrow T(P_i)$, $T_{i+1} \leftarrow T_i \circ T$, $i \leftarrow i + 1$
end while
return S

4.5 Results

This section presents the implementation of our approach and the tuning of various parameters. We present experiments in an extended set of BiGG models [King et al., 2016], including the most complex metabolic networks, the human Recon2D [Swainston et al., 2016] and Recon3D [Brunk et al., 2018]. We end up to sample from polytopes of thousands of dimensions and show that our method can estimate precisely the flux distributions. We analyze various aspects of our method such as the run-time, the efficiency, and the quality of the output.

We compare against the state-of-the-art software for the analysis of metabolic networks, which is the Matlab toolbox of cobra [Heirendt et al., 2019]. Our implementation for low dimensional networks is two orders of magnitude faster than cobra. As the dimension grows, this gap on the run-time increases. The workflow of cobra for sampling first performs a rounding step and then samples using Coordinate Directions Hit-and-Run

(CDHR).

In [Jadebeck et al., 2020] they provide a C++ implementation of the sampling method that cobra uses and they show that their implementation is approximately 6 times faster than cobra. Nevertheless, we choose to compare against cobra, since it additionally provides efficient preprocessing methods that are crucial for the experiments, and give an implicit comparison with [Jadebeck et al., 2020].

The fast mixing of billiard walk allow us to use all the generated samples to approximate each flux distribution and so we compute a better flux distribution estimation. To estimate each marginal flux distribution, using the samples, we exploit Gaussian kernel density estimation. This is a non-parametric way to estimate the probability density function of a random variable. For more details we refer to [Jones et al., 1996].

We provide a complete open-source software framework to handle big metabolic networks. The framework loads a metabolic model in some standard file formats (e.g., mat and json files) and performs an analysis of the model, e.g., it estimates the marginal distributions of a given reaction flux. All the results are reproducible using our publicly available code

The core of our implementation is in C++ to optimize performance while the user interface is implemented in R. The package employs [Guennebaud et al., 2010] for linear algebra, number generation, [Chalkis and Fisikopoulos, 2020], an open-source package for high dimensional sampling and volume approximation.

All experiments were performed on a PC with Intel Core i7-6700 3.40GHz × 8 CPU and 32GB RAM. In the sequel, MMCS refers to our implementation.

We test and evaluate our software on 17 models from the BIGG database [King et al., 2016] as well as Recon2D and Recon3D from [Noronha et al., 2019]. In particular, we sample from models that correspond to polytopes of dimension less than 100; the simplest model in this setting is the well known bacteria *Escherichia Coli*. We also sample from models that correspond to polytopes of dimension a few thousands; this is the case for Recon2D and Recon3D. We do not employ parallelism for any implementation, thus we report only sequential running times.

We assess the quality of our results by employing both the Effective Sample Size (ESS) and the potential scale reduction factor (PSRF) [Gelman and Rubin, 1992]. In particular, we compute the PSRF for each univariate marginal of the sample that MMCS outputs. Following [Gelman and Rubin, 1992], a convergence is satisfying according to PSRF when all the marginals have PSRF smaller than 1.1.

In Table 4.1, we report the results of MMCS and cobra. For cobra, we report only the run-time of the sampling phase (we do not add to it the preprocessing time). We run MMCS until we get a value of ESS equal to 1 000; i.e. we stop when the sum over all phases of the minimum values of ESS among all the marginals is larger than 1 000. All the marginals of the MMCS samples reported in Table 4.1 have PSRF < 1.1. This is a strong statistical evidence on the quality of the generated sample.

The marginal flux distribution of reaction Thioredoxin in Recon2D was estimated also in [Haraldsdóttir et al., 2017] and used as an evidence for the quality of the sample. In Figure 4.2, we employ the copula representation to capture the dependency between two fluxes of reactions and confirm a mutually exclusive pair of biochemical pathways.

				MMCS		cobra	
model	m	n	d	Time (sec)	N	Time (sec)	N
e_coli_core	72	95	24	6.50e-01	3.40e+03 (8)	7.20e+01	4.61e+06
iLJ478	570	652	59	9.00e+00	5.40e+03 (5)	4.54e+02	2.79e+07
iSB619	655	743	83	1.70e+01	8.20e+03 (5)	9.56e+02	5.51e+07
iHN637	698	785	88	2.00e+01	6.80e+03 (4)	1.03e+03	6.19e+07
iJN678	795	863	91	2.50e+01	8.10e+03 (4)	1.17e+03	6.62e+07
iNF517	650	754	92	1.70e+01	6.20e+03 (4)	1.33e+03	6.77e+07
iJN746	907	1054	116	5.70e+01	8.70e+03 (3)	2.22e+03	1.07e+08
iAB_RBC_283	342	469	130	5.20e+01	1.07e+04 (5)	7.85e+03	4.05e+08
iJR904	761	1075	227	2.98e+02	1.62e+04 (4)	8.81e+03	4.12e+08
iAT_PLT_636	738	1008	289	3.25e+02	1.04e+04 (2)	1.73e+04	6.68e+08
iSDY_1059	1888	2539	509	2.813e+03	2.31e+04 (3)	6.66e+04	2.07e+09
iAF1260	1668	2382	516	6.84e+03	5.33e+04 (6)	7.04e+04	2.13e+09
iEC1344_C	1934	2726	578	4.86e+03	3.95e+04 (4)	9.42e+04	2.67e+09
iJO1366	1805	2583	582	6.02e+03	5.14e+04 (5)	9.99e+04	2.71e+09
iBWG_1329	1949	2741	609	3.06e+03	4.22e+04 (4)	1.05e+05	2.97e+09
iML1515	1877	2712	633	4.65e+03	5.65e+04 (5)	1.15e+05	3.21e+09
Recon1	2766	3741	931	8.09e+03	1.94e+04 (2)	3.20e+05	6.93e+09
Recon2D	5063	7440	2430	2.48e+04	5.44e+04 (2)	~ 140 days	$1.57e{+}11$
Recon3D	8399	13543	5335	1.03e+05	1.44e+05 (2)	–	–

TABLE 4.1: Several, 17, metabolic networks from [King et al., 2016]; also Recon2D and Recon3D from [Noronha et al., 2019]. The semantics of the tables are as follows: (m) the number of Metabolites, (n) the number of Reactions, (d) the dimension of the polytope; (N) is the total number of sampled points × walk length; for MMCS we stop when the sum of the minimum value of ESS among all the univariate marginals in each phase is 1 000 (we report the number of phases in parenthesis); for cobra we set the walk length to $8d^2$ and $1.57e{+}08$ for Recon2D stop when all marginals have PSRF < 1.1; the run-time of cobra for Recon2D is an estimation of the sequential time and we report it to have a rough comparison with our implementation.

Comparing runtime performance, MMCS is one or two orders of magnitude faster than cobra and this gap becomes much larger for higher dimensional models such as Recon2D and Recon3D. Considering the experiments reported in [Jadebeck et al., 2020], they report the run-time of CDHR for each model until it generates a sample with PSFR 1.2; for Recon3D they report ~ 1 day. Interestingly, for Recon3D, MMCS achieves PSRF 1.2 after ~ 1 hour while reach PSRF 1.1 after ~ 1 day.

For some models –we report them in Table 4.2– we introduce a further improvement to obtain a better convergence. If there is a marginal in the generated sample from MMCS that has a PSRF larger than 1.1, then we do not take into account the k first phases, starting with $k = 1$ until we get both ESS equal to 1 000 and all the PSRF values smaller than 1.1 for all the marginals. By "we do not take into account" we mean that we neither store the generated sample –for the first k phases– nor we sum up its ESS to the overall ESS considered for termination by MMCS. Note that for these models it is not practical to repeat MMCS runs for different k until we get the required PSRF value. We can obtain the final results –reported in Tables 4.1– in one pass. We simply drop a phase when the

model	k	Time (sec)	PSRF < 1.1	M	N
iAF1260	0	6955	41%	6	56100
	1	6943	56%	6	54100
	2	6890	76%	6	55200
	3	6867	95%	6	53200
	4	6840	100%	6	53300
iBWG_1329	0	3067	50%	4	42100
	1	3189	97%	5	48800
	2	4652	100%	5	56500
iEC1344	0	4845	77%	4	41100
	1	4721	96%	4	42500
	2	4682	100%	4	39500
iJO1366	0	3708	66%	5	51500
	1	6022	100%	5	51400

TABLE 4.2: During our experiments we do not take into account the sample of the k first phases, thus we do not also count the value of the Effective Sample Size (ESS) in these phases, before we start storing the generated sample and sum up the ESS of each phase. In all cases MMCS stops when the sum of ESS reaches 1000. For each case we report the total run-time, the percentage of the marginals that have PSRF smaller than 1.1, the total number of phases (M) needed (including the k first phases), and the total number of Billiard Walk steps (N), including those performed in the k first phases.

ESS reaches the requested value but the PSRF is not smaller than 1.1 for all the marginals. In Table 4.2, we separately report the MMCS runs for different k just for performance analysis reasons.

Interestingly, the total number of Billiard Walk steps –and consequently the run-time– does not increase as k increases in Table 4.2. This means that the performance of our method improves for these models when we do not take into account the k first phases of MMCS. This happens because the performance of Billiard Walk improves as the polytope becomes more rounded from phase to phase.

In Table 4.3, we analyze the performance of Billiard Walk for the model iAF1260. We sample $20d$ points per phase with walk length equal to 1 and we report the average number of reflections, the ESS, the run-time, and the ratio $\sigma_{\max}/\sigma_{\min}$ per phase. The latter is the ratio of the maximum over the minimum singular value of the point-set. The larger this ratio is the more skinny the polytope of the corresponding phase is. As the method progresses from the first to the last phase, the average number of reflections and the run-time decrease and the ESS increases. This means that as the polytope becomes more rounded from phase to phase, the Billiard Walk step becomes faster and the generated sample has better quality. This explains why the total run-time does not increase when we do not take into account the first k phases: the initial phases are slow and they contribute poorly to the quality of the final sample; the last phases are fast and contribute with more accurate samples.

4.6 Conclusions and future work

We propose a novel method for sampling that can sample from a convex polytope in a few thousands of dimensions within a day on modest hardware. This way, we are able, for the first time, to perform accurate sampling from the latest human metabolic network, Recon3D.

Sampling from iAF1260				
Phase	Avg. #reflections	ESS	$\frac{\sigma_{\max}}{\sigma_{\min}}$	Time (sec)
1st	7819	67	43459	2271
2nd	4909	68	922	1631
3rd	3863	77	582	1278
4th	3198	71	360	1080
5th	1300	592	29	454
6th	1187	4821	3.5	417
7th	1181	4567	2.8	415

TABLE 4.3: We sample $20d = 10320$ points per phase with Billiard Walk and walk length equal to 1, where $d = 516$ is the dimension of the corresponding polytope. For each phase we report the average number of reflections per BIlliard Walk step, the minimum value of Effective Sample Size among all the univariate marginals, the ratio between the maximum and the minimum singular value of the SVD decomposition of the generated sample, and the run-time.

Regarding future work, parallelism could lead to a speedup in the run-time of our method as the algorithm is rather straightforward to parallelize. An additional improvement would be to exploit the sparsity of the stoichiometric matrix S and sample directly from the low dimensional polytope in $\mathbb{R}^n$ without projecting to a lower dimensional space.

Moreover, our method could be extended to any log-concave distribution restricted to the flux space and combined with bayesian metabolic flux analysis, to sample from multivariate, possibly multi-modal target distribution [Heinonen et al., 2019] addressing multiple challenges of the method from the biological point of view (e.g., unrealistic assumptions, uncertainty etc.). Last but not least, flux sampling in metabolic models built out from multiple metabolic networks, e.g., representing a microbial community, could also lead to important biological insights.

The scenario presented in Figure 4.2 demonstrates an essential issue of steady-state oriented methods, that leads to obscure, non-viable flux vectors. As mentioned, flux sampling like FBA, it also assumes that the system is at a steady state. That means that no matter of the flux of each reaction, the total concentration of each metabolite is constant. Therefore, in the example shown in Figure 4.2 (left copula) glucose is converted to G6P by HEX at the expense of ATP. At the same time, glucose is produced again by the G6PPer phosphatase. Such a scenario leads to a dead cell but allows us to make sure that the sampling has been implemented the way it should. To use flux sampling in actual scenarios, there must be an uptake flux for glucose and a production of some end product which is excreted in order to force the sampling to produce a viable flux and some more meaningful result. In the near future, several tests will be added in the dingo Python

library ensuring the viability of the samples returned. Our main goal is to sample on the flux space of a microbial community to infer microbial interactions (see Conclusion 7.4) and maybe, start getting some answers on the *how* and *why* questions that have made an *"entangled bank"* out of the mechanisms that govern such assemblages.

Chapter 5

Deciphering the functional potential of a hypersaline marsh microbial mat community

Citation:
Pavloudi C, Zafeiropoulos H, Deciphering the functional potential of a hypersaline marsh microbial mat community.
Under review in FEMS Microbiology Ecology.
Shared co-first authorship and correspondance.

5.1 Abstract

Microbial mats are vertically stratified communities of microorganisms characterised by pronounced physiochemical gradients allowing for high species diversity and a wide range of metabolic capabilities. High Throughput Sequencing has the potential to reveal the biodiversity and function of such ecosystems in the cycling of elements and organic matter recycling.

The present study combines 16S rRNA amplicon sequencing and shotgun metagenomics on sediments and microbial mats from a hypersaline marsh in Tristomo bay (Karpathos, Greece). Sampling was conducted in July 2018 and November 2019. Samples were collected from the microbial mats and the deeper sediment; orange and pink microbial aggregates observed in the water overlying the sediment were also collected, as well as sediment samples with no apparent layering.

Metagenomic assembly and binning in the sample level, revealed 250 bacterial and 39 archaeal metagenome-assembled genomes, with completeness estimates higher than 70% and contamination less than 5%. Halobacteria and Bacteroidetes were among the most abundant taxa in the microbial mats. Photosynthesis was most likely performed by purple sulphur and non-sulphur bacteria.

Overall, both the sequencing methodologies seemed to result in similar taxonomic compositions. All samples had the functional capacity for sulphate reduction, dissimilatory arsenic reduction and conversion of pyruvate to oxaloacetate.

5.2 Introduction

Microbial mats are vertically stratified communities of functional groups of microorganisms embedded in an organic matrix, which may also contain minerals such as silicates and carbonates [Stal, 2012, Bolhuis et al., 2014, Prieto-Barajas et al., 2018]. They grow on a solid substrate (e.g. sand) and the vast majority of microbial mats utilise inorganic carbon as carbon source, hence they are autotrophic [Bolhuis et al., 2014]. Microbial mats are characterised by pronounced physiochemical gradients which allow for the presence of high species diversity, encompassing a wide range of metabolic capabilities; thus, mats are ideal models to study a whole ecosystem [Al-Thani et al., 2014] and are considered as natural laboratories [Villanueva et al., 2007]. These physicochemical gradients provide microenvironments for various microbial functional groups, which exhibit a certain physiology with which they fulfil a specific function [van Gemerden, 1993].

Microbial mats comprise millions of microorganisms belonging to different species which are embedded in a matrix of extracellular polymers (EPS) and exchange signals and nutrients, thus enabling a flow of resources and energy for the survival of the overall community [Ruvindy et al., 2016, Prieto-Barajas et al., 2018]. The role of microbial mats has been vital throughout Earth's history since they produced and released reduced gases, e.g. O_2, H_2, CH_4, in the early earth's atmosphere [Hoehler et al., 2001]. In addition, they constitute the first ecosystems, along with stromatolites [Santoyo, 2021], and probably are the oldest structured ecosystems on earth [van Gemerden, 1993].

Regardless of the vertical structure, marine microbial mats are comprised of four main functional groups: i) oxygenic phototrophs (CYN) (primarily Cyanobacteria), ii) aerobic heterotrophic bacteria (HET), iii) sulphate-reducing bacteria (SRB) and iv) sulphide-oxidising bacteria (SOB) [Visscher and Stolz, 2005]. Microbial mats function as a consortium where coupling of biogeochemical cycles and processes occurs [Paerl et al., 2000], allowing the products of the metabolism of one group to be available and used by another [Santoyo, 2021]. In addition, the metabolic rates of mat microorganisms are so high that the community production per unit mass competes with that of rainforests [Jørgensen, 1994, Krumbein et al., 2003].

Microbial mats can be distinguished in six categories [Bolhuis et al., 2014, Prieto-Barajas et al., 2018]: i) intertidal or coastal, ii) hypersaline, iii) hot spring, iv) mats in oligotrophic environments, v) psychrophile and vi) acid microbial mats. Intertidal mats are formed on beaches with low slopes and fine sandy sediments [Stal, 2012] and they experience strong salinity fluctuations, large temperature changes [Bolhuis et al., 2014] and irregular floods [Prieto-Barajas et al., 2018]. On the other hand, hypersaline microbial mats are found in natural occurring salt lakes and man-made salterns [Bolhuis et al., 2014] and are exposed to salinities up to the crystallisation point of halite [Jørgensen, 1994], high temperatures and high solar radiation [Bolhuis et al., 2014].

The present study was conducted in the Tristomo marsh in the island of Karpathos (Aegean Sea, Greece) (Figure 5.1 A). The marsh is located at the northern end of Karpathos. The study area is included in the Natura 2000 network (site GR4210003) and also in the catalogue of small island wetlands (Government Gazette Issue on Compulsory Expropriations and City-Planning 229/19.6.2012) with the code Y421KAR001 (total area: 1.9 ha) (Figure 5.1 B). It is a seasonal brackish water marsh formed at the edge of a small plain

where a seasonal stream ends, characterised as an intertidal marsh (type H) according to the Ramsar convention. In the past, it probably occupied a larger area and was connected to the stream. Most of the area is fenced with dry stones that used to be cultivated systematically. Today crops exist in only a small part and there is livestock grazing around the marsh. On the coastal front, the cobbled beach is full of garbage carried by the waves. Freshwater enters the marsh from the precipitation and drainage basin, while the wetland interacts mainly with the sea through the waves but also underground [Greece, 2022]. Due to the close proximity of the marsh with the sea, it occasionally receives saline water, therefore could be characterised as intertidal; however, crystallised salt forms an upper layer above the actual microbial mat, something that is observed in hypersaline mats (Figure 5.1 C and D). High Throughput Sequencing (HTS) technologies and methods have been widely used to study real-world microbial communities. They have enabled the study of ecosystems with no prior knowledge of the resident species, uncovering unknown and uncultivated strains [Hedlund et al., 2014]. Metabarcoding studies are common, well-established and less computationally demanding than shotgun metagenomics [Bell et al., 2021b]. However, taxonomic biases may arise from differential efficiency of PCR primer pairing in different species [van der Loos and Nijland, 2021] while the short barcoding sequences may limit the resolution. On the other hand, shotgun metagenomics by obtaining information from random sampling of virtually all genomic regions, enables profiling up to the level of strains [Clooney et al., 2016, Segata, 2018, Dávila-Ramos et al., 2019]. Therefore, microbiome metabolic functions and entire biochemical pathways that occur in a sample can be explored after processing the metagenomic information [Sharpton, 2014].

Over the recent years, HTS approaches have been used to study the taxonomic and the functional profiles of the microbial communities present in microbial mats [Chen et al., 2020b, Wong et al., 2020, Kindler et al., 2022]. Several novel high-level taxa have been discovered, e.g. Zixibacterial order GN15 [Wong et al., 2020], and a better understanding on both their adaptive responses in such environments has been established. On top of that, further insight on the mechanisms governing such assemblages has been gained, e.g. the role of photoheterotrophy [Kindler et al., 2022]. The aim of the present study was to identify the microbial communities present in samples from the hypersaline Tristomo marsh, as well as their functional and metabolic capabilities.

5.3 Methods

5.3.1 Sample collection

Samples were collected in July 2018 and November 2019 from the Tristomo marsh (Figure 5.1). Details on the sample collection are given in Table 5.1. Sediment samples were collected using cylindrical sampling corers (internal sampling surface 15.90 square centimetres) (Figure 5.1.E). In the cases where microbial mat layers were clearly observed (July 2018), the top layer was collected separately from the bottom layer. In addition, microbial aggregates observed floating in the marsh were also collected. In the cases where microbial mat layers were not clearly formed (November 2019), there was no slicing during sample collection. In November, samples were collected from three different

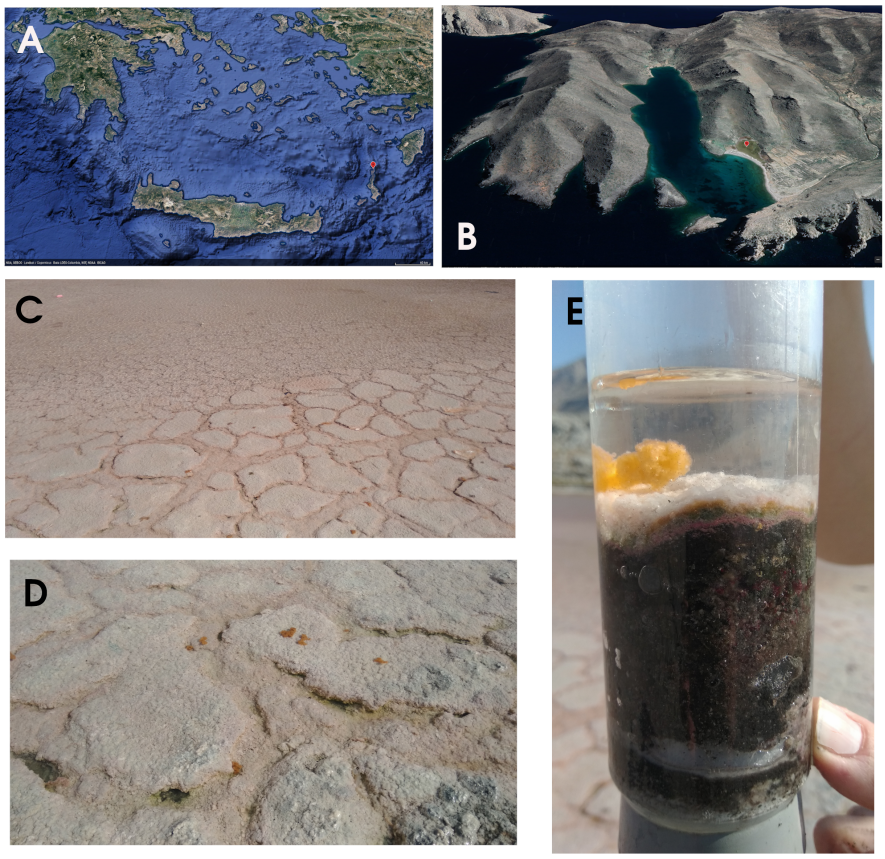


FIGURE 5.1: A) Location of Karpathos island (red pin) in the south east of the Aegean Sea, B) satellite image of Tristomo bay and the Tristomo marsh (red pin), C) overview of the marsh in July 2018, D) overview of the marsh in July 2018, where orange aggregates floating in the water are shown and E) a sediment core from 2018 including salt crust, microbial mat, sediment and an orange aggregate. (Map data: ©2022 Google Earth).

locations in the marsh, distinguished by the colour of the sediment's upper layer (black, purple and orange).

Samples were placed in 50 ml falcon tubes (Sarstedt, Nümbrecht, Germany) and were stored at −20 °C, until further processing in the laboratory. Upon return to the laboratory, they were used for molecular analysis, i.e. DNA extractions, as well as for the measurement of the Particulate Organic Carbon (POC) and chloroplast pigments concentration (chlorophyll-a, phaeopigments and chloroplastic pigment equivalents (CPE)). For the latter, the samples were processed at the Environmental Chemistry Lab of the IMBBC (HCMR), based on standard techniques [Yentsch and Menzel, 1963, Hedges and Stern, 1984]. Water temperature and dissolved oxygen concentration were measured in the water overlaying the sediments by means of a portable multi-parameter (WTW Multi 3420 SET G). Salinity was also measured with the portable multi-parameter but after dilution of samples with dH2O since the initial measurement was out of limits (TetraCon® 925 sensor

Sample	Type	Date	16S rRNA accession number	shotgun accession number (lane 1)	shotgun accession number (lane 2)	latitude	longitude
Elos01	top sediment layer	8/7/2018	ERR9657902	ERR6290772	ERR6290778	35.81936	27.20984
Elos02	bottom sediment layer	8/7/2018	ERR9657903	ERR6290773	ERR6290779	35.81936	27.20984
Elos03	orange aggregate	8/7/2018	ERR9657904	ERR6290774	ERR6290780	35.81936	27.20984
Elos04	top sediment layer	13/7/2018	ERR9657905			35.81936	27.20984
Elos05	bottom sediment layer	13/7/2018	ERR9657906			35.81936	27.20984
Elos06	orange aggregate	13/7/2018	ERR9657907			35.81936	27.20984
Elos07	pink aggregate	13/7/2018	ERR9657908	ERR6290775	ERR6290781	35.81936	27.20984
Elos08	top sediment layer	9/7/2018	ERR9657909			35.81936	27.20984
Elos09	bottom sediment layer	9/7/2018	ERR9657910			35.81936	27.20984
Elos10	sediment (combined layers, black upper layer)	12/11/2019	ERR9657898	ERR6290776	ERR6290782	35.81962	27.2102
Elos11	sediment (combined layers, black upper layer)	12/11/2019	ERR9657899			35.81962	27.2102
Elos12	sediment (combined layers, orange upper layer)	12/11/2019	ERR9657900	ERR6290777	ERR6290783	35.81942	27.21007
Elos13	sediment (combined layers, purple upper layer)	12/11/2019	ERR9657901			35.81943	27.20998

TABLE 5.1: Details on Karpathos marsh sample collection.

range: 0 - 70). Sampling was conducted under authorization from the relevant licensing authority (Directorate General for the Protection and Development of Forests and the Rural Environment, Directorate of Forest Management) of the Ministry of Environment and Energy. Additional authorization was also provided from the Management Agency of Dodecanese Protected Areas.

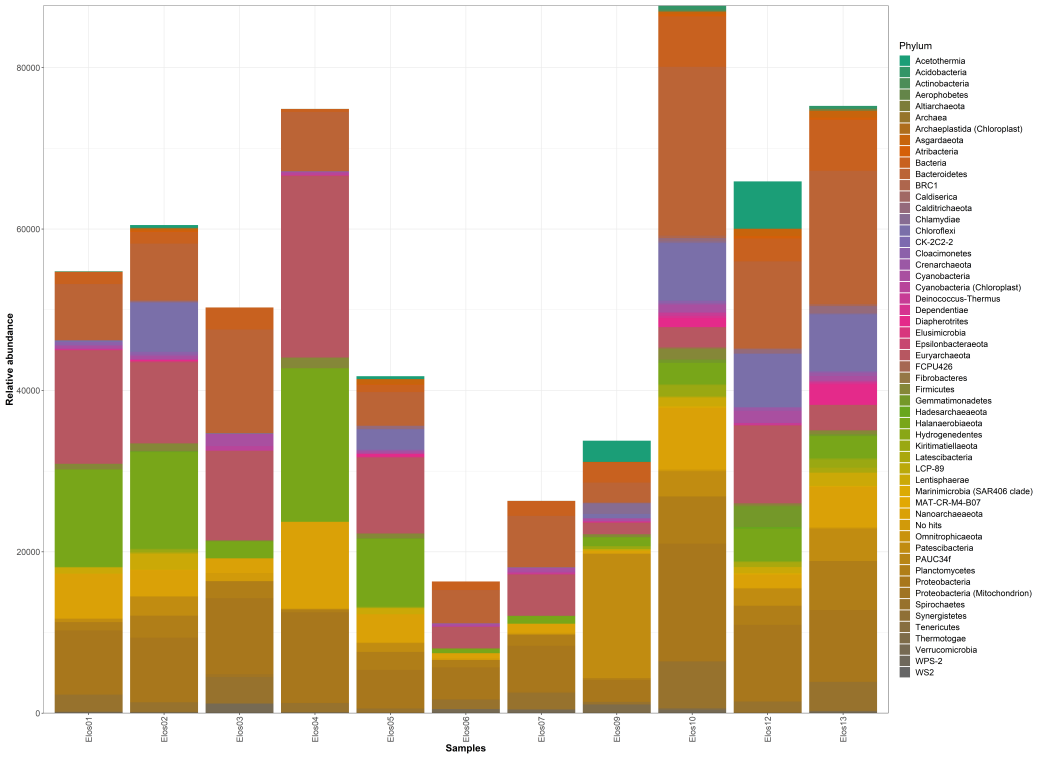


FIGURE 5.2: Bar chart showing the abundances of the main microbial taxa, at the phylum level, at each sample, based on the 16S rRNA amplicon sequencing.

5.3.2 DNA extraction, PCR amplification and 16S rRNA sequencing

DNA was extracted as in [Henckel et al., 1999] and [Lueders et al., 2004]. Approximately 0.7 g of wet sediment were added to a 2-ml screw-cap vial, prefilled with 0.7 g of 0.1 mm (diameter) zirconia/silica beads (11079101z, BioSpec, USA). The vials were filled with 750 µl of 120 mM NaPO4 buffer (pH 8) and 250 µl TNS solution (500 mM Tris-HCl pH 8, 100 mM NaCl, 10 % SDS (w/v)) and placed horizontally in a vortex for 10 minutes at maximum speed. Immediately after that the vials were centrifuged for 10 min at 20,800 rcf and 4 °C and the supernatants were transferred to new 2-ml vials. One volume of phenol/chloroform/isoamylalcohol (P/C/I; 25:24:1; pH 8; Carl-Roth, Karlsruhe, Germany) was added to the aqueous supernatant. Vials were vigorously shaken for 20 s and centrifuged for 5 min at 20,800 rcf and 4 °C. Supernatants were transferred to new 2-ml vials, and one volume of chloroform/isoamylalcohol (C/I; 24:1; Carl-Roth) was added. Vials were again vigorously shaken for 20 s and then centrifuged for 5 min at 20,800 rcf

and 4 °C. Supernatants were transferred to new 2-ml vials and C/I extraction was repeated to successfully remove all phenol remnants. Supernatants were transferred to new 2-ml vials and 1.5 ml of polyethylene glycol (30 % (w/v) polyethylene glycol 6000 in 1.6 M NaCl) was added to precipitate nucleic acids and the vials were centrifuged for 90 min at 20,800 rcf and 4 °C. Supernatants were discarded and the pellets were washed with 1 ml 70% ethanol (4 °C) and centrifuged for 30 min. Supernatants were again discarded, pellets were left for air drying (5 min) to remove leftover ethanol and resuspended with 50 µl 10mM Tris. PCR amplification, library preparation and MiSeq sequencing was performed as in [Pavloudi et al., 2017a]. The PCR negative control sample (blank) was also sequenced, so that possible contamination during the library preparation could be assessed. The raw sequence reads were processed with PEMA (version 2.1.4) [Zafeiropoulos et al., 2020] using VSEARCH for the creation of OTUs. Taxonomic assignment was performed with the SILVA database (version 132) [Quast et al., 2013]. The detailed parameters of the PEMA processing are given in Supplementary File 1. The phyloseq (version 1.36) [McMurdie and Holmes, 2013], vegan (version 2.5.7) [Oksanen et al., 2020] and ggplot2 (version 3.3.5) [Wickham, 2016] packages were used in R (version 4.1.1) (R Core Team 2021) for the creation of barcharts, for the nMDS and PERMANOVA, variation partitioning analysis, db-RDA and mantel test. The scripts of Steinberger (2020) [Steinberger, 2020] were used for the simper and the Kruskal-Wallis tests.

5.3.3 Shotgun metagenomics sequencing

Six samples were selected for shotgun sequencing (Elos01, Elos02, Elos03, Elos07, Elos10 and Elos12). Sample preparation was performed using the Nextera^{TM} DNA Flex Tagmentation and sequencing was done at two lanes of a HiSeq 4000 (2x150bp) at the Norwegian Sequencing Centre (NSC). All the raw sequence files of this study (both 16S rRNA and shotgun metagenomes) were submitted to the European Nucleotide Archive (ENA) [Cummins et al., 2022] with the study accession number PRJEB46254 (available at: http://www.ebi.ac.uk/ena/data/view/PRJEB46254).

5.3.4 Assembly and binning

Since the samples were sequenced in two lanes, the fastq files of each sample were concatenated before proceeding with the analyses. Metagenome raw reads were processed with the MetaWRAP workflow (version 1.3.2) [Uritskiy et al., 2018]. Reads were trimmed and qualified using Trim Galore (version 0.5.0) [Krueger, 2022], which is a wrapper around Cutadapt (version 1.18) [Martin, 2011] and FastQC. The clean reads were concatenated and their co-assembly was implemented through the corresponding MetaWRAP module, using MEGAHIT v1.1.3. The quality of the co-assembly was evaluated using QUAST [Gurevich et al., 2013] (see assembly_report.html). Binning was then performed using the clean reads and the co-assembly. The MetaWRAP module for binning was performed using MetaBAT 2 (version 2.12.1) [Kang et al., 2019] and MaxBin 2 (version 2.2.6) [Wu et al., 2016]. CheckM (version 1.0.12) [Parks et al., 2015] was used by the MetaWRAP module to assess the quality of the bins produced by MetaBAT 2 and MaxBin 2. Bins were then consolidated and refined using Binning_refiner [Song and Thomas, 2017] as wrapped in

the Bin_refinement module of MetaWRAP. The Bin_refinement module was invoked with the default values for minimum completion (70%) and maximum contamination (5%); see binning_results.png. The consolidated bins set was further improved using the reassemble_bins module of MetaWRAP. To this end, bwa (version 0.7.17-r1188) [Li and Durbin, 2009], spades (version v3.13.0) [Nurk et al., 2017] and CheckM were used; see binning_reassembled.png. To estimate bins' abundances in each sample (in genome copies per million reads), the corresponding MetaWRAP module was performed invoking Salmon (version 0.13.1) [Patro et al., 2017]. The refined bin-set was also used for the blobology module of MetaWRAP; taxonomic annotation of the co-assembled contigs was performed using megaBLAST and the nt database of NCBI.

The co-assembled contigs and the refined bins set were then used as input to Anvi'o (version 7.1) [Eren et al., 2015]. Bowtie 2 (version 2.3.5) [Langmead and Salzberg, 2012] was used to build BAM files and mapping and Prodigal (version 2.6.3) [Hyatt et al., 2010] for gene prediction. BAM files were also made out of the clean reads of each sample. A contigs database was built (using the anvi-gen-contigs-database program) after converting the contigs name as Anvi'o suggests (see contigs-per-bin.sh script) and it was decorated with hits from HMM models (anvi-run-hmms). An anvi profile was then built for each of the samples' bam file (anvi-profile) and they were merged (anvi-merge) into a single profile. The refined bins along with their corresponding renamed contigs were imported as a collection in the merged profile database (anvi-import-collection). At this point, a first Anvi'o summary was recovered (anvi-summarize) (see 1st_bins_summary.txt). Bins with a redundancy >10% were manually refined and a second summary of the bins set was made (see SECOND_SUMMARY folder).

5.3.5 Taxonomic composition

Based on the returned co-assembly from MetaWRAP and the clean reads, communities' taxonomic composition was assessed using Kraken2 [Wood et al., 2019] and the standard Kraken 2 database (NCBI: January 2022); Krona plots of the community profiles can be viewed through the kronagram.html. GTDB-Tk (version 1.7.0) [Chaumeil et al., 2020] was used to classify genomes with the Genome Taxonomy Database (GTDB, version r202) [Parks et al., 2022]. GTDB-Tk made use of pplacer (version 1.1.alpha19-0-g807f6f3) [Matsen et al., 2010] and FastANI (version 1.32) [Jain et al., 2018].

5.3.6 Functional annotation

Functions were predicted at two levels: both at the MAG level, as well as at the sample level. For the functional annotation at the MAG level, using the anvio' contigs database and the anvi-run-kegg-kofams program, the anvio' contigs database was annotated with HMM hits from KOfam, a database of KEGG Orthologs (KOs). Likewise, using the anvi-run-ncbi-cogs, NCBI's Clusters of Orthologous Groups (COGs) based annotations were added. The MAGs that correspond to the refined bins as they were retrieved after the MetaWRAP and the anvio refinement steps, were annotated with KEGG modules; manually defined functional units of gene and reaction sets [Kanehisa et al., 2012]. MAGs were "translated" to an anvio collection (i.e., a virtual construct storing bins of items in an Anvi'o profile

database) and this collection was used along with the anvi-estimate-metabolism program to determine which enzymes are present in each MAG and compute the completeness of each metabolic module (scripts can be found under the anvio folder). An nMDS was constructed based on the presence/absence of modules in the MAGs using the jaccard similarity index. For the functional annotation at the sample level, the clean reads as they were returned by the corresponding MetaWRAP module and the DiTing tool [Xue et al., 2021] were used to estimate the contribution of each sample to the biogeochemical cycles incorporated in DiTing. DiTing used MEGAHIT [Li et al., 2015] to build the assembly of each sample separately (so, the co-assembly described in the "Assembly and binning" section was not used for this step) and Prodigal to retrieve the Open Reading Frames (ORFs). KofamScan [Aramaki et al., 2020] was used for the annotation of the ORFs using KEGG ORTHOLOGY terms. The relative abundances of metabolic and biogeochemical functional pathways in each sample were then determined by DiTing (see DiTing folder on the GitHub repository for more information).

5.3.7 MAGs reference phylogenies

The intersection of single copy genes from the Anvi'o [Eren et al., 2015] Bacteria_76 and Archaea_71 sets was used to build the phylogenetic tree of the reconstructed MAGs (n = 25). The anvi-get-sequences-for-hmm-hits program of Anvi'o was used to extract and align the amino acid sequences of each of these genes from all the MAGs independently. This Anvi'o program makes use of MUSCLE (version 3.8.1551) [Edgar, 2004] to return an alignment of the extracted sequences. Once all the amino acid sequence alignments were extracted, they were trimmed using Clipkit (version 1.1.5) [Steenwyk et al., 2020]. A super matrix was then built using the single copy genes of the intersection. In cases where a MAG lacked a gene, gaps were filled with dashes; both the initial and the trimmed per gene alignments as well as the final super matrix alignment are available on the project's GitHub repository under the SCG folder. Using IQ-TREE2 [Hoang et al., 2018b, Minh et al., 2020] the phylogeny of the reconstructed MAGs was built using 1,000 bootstrap replicates (-B 1,000) and 1,000 bootstrap replicates for Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT) (-alrt 1000). The best-fit model (LG+R10) was retrieved using ModelFinder [Kalyaanamoorthy et al., 2017]. Using Barrnap [Seemann, 2014a] the 16S rRNA gene was extracted from the retrieved MAGs. The phylogeny of the MAGs and their relative abundances were integrated and visualised using GraPhlAn [Asnicar et al., 2015]. All bioinformatics analyses were supported by the IMBBC High Performance Computing system [Zafeiropoulos et al., 2021d].

5.4 Results

5.4.1 Taxonomic composition from 16S rRNA amplicon analysis

The results of the processing of the sequences are shown in Table S1. Sequencing of samples Elos08 and Elos11 was not successful and therefore, they were not included in the following analyses. The final number of OTUs, after removal of the OTUs that were also found on the blank sample, was 2,689. The most abundant phyla, as assessed by the rela-

tive abundance percentages of each replicate sample averaged per sampling station, were Bacteroidetes (17%), Euryarchaeota (16%), Proteobacteria (15%) and Halanaerobiaeota (10%, class Halanaerobiia) (Figure 5.2). Among the Bacteroidetes, the most abundant class was Bacteroidia (14%), followed by Rhodothermia (3%). Euryarchaeota had very low abundances in samples Elos09, Elos10 and Elos13 and the higher abundances in the top sediment layers, i.e. in the microbial mat samples (Elos01 and Elos04). In addition, among the Euryarchaeota, the most abundant class was Halobacteria (14%) followed by Thermoplasmata (2%). Proteobacteria were almost equally distributed among the classes Alphaproteobacteria (7%), Gammaproteobacteria (4%) and Deltaproteobacteria (4%). Although Patescibacteria had a low average abundance (6%), they were dominant in Elos09 (46%). Halanaerobiaeota had low abundances in the orange and pink aggregates (Elos03, Elos06, Elos07) as well as in Elos09, Elos10 and Elos13. In addition, Chloroflexi were almost absent from the top layers and the aggregates but they were found in the bottom sediment layer and in the combined sediment samples. Cyanobacteria were about 1% on average of all the samples. The nMDS of the microbial OTUs (Figure S1) showed that their spatial pattern differs both by their type and the year of sampling, which was also confirmed by the PERMANOVA results (Type: F.Model = 2.0396, $p < 0.05$; Year: F.Model = 2.3098, $p < 0.05$).

5.4.2 Co-assembly, binning & taxonomic composition from shotgun metagenomics analysis

Shotgun metagenomic sequencing of the chosen six samples resulted in 744 million reads totalling 112.3 Gbp, with each sample ranging between 16.77 and 21.78 Gbp. Co-assembling of all the samples resulted in 1,5 million contigs totalling 5.04 Gbp. The per-sample assemblies returned a total of 11.2 million contigs with a sum of 10.15 Gbp. Number of reads per sample, before and after the quality control, their length and the corresponding number of contigs are shown in Table S2. Based on the taxonomic profiles retrieved from Kraken2 (Figure S2), after removing sequences belonging to Viruses and sequences that could not be classified (1%), Euryarchaeota (class Halobacteria) represent the majority of the total archaeal taxa (30% on average); however, they are almost absent from sample Elos10 (abundance 1%) while they are dominant in sample Elos01 (59%). As far as bacterial taxa are concerned, the most abundant ones were Alphaproteobacteria (19%), followed by Actinobacteria (13%) and Gammaproteobacteria (10%). Betaproteobacteria, delta/epsilon Proteobacteria subdivisions and Bacteroidetes/Chlorobi group had similar abundances (5%, 4% and 5% respectively). Cyanobacteria were limited in all samples (2% on average). Also Kraken2 analysis did not identify any Nanoarchaeota, it identified in sample Elos01 the other archaeal taxa that are their hosts and namely a) *Ignicoccus* hospitalis, b) *Acidilobus* sp. 7A, c) *Vulcanisaeta* spp., d) *Pyrobaculum* spp., e) *Metallosphaera* spp., f) *Caldivirga* sp. and g) *Sulfolobus* sp.

Krona plots with the taxonomic profiles of each sample are available on GitHub. Prodigal predicted millions of genes per sample ranging from 2.1 millions (Elos03) to 4.3 (Elos10). Their metabolic capacity/potential is further described in the "Biogeochemical cycles" section. Based on the blobology results, among the 1,513,505 co-assembled contigs a set of 102,250 were binned (see Figure S3); according to megaBlast and the nt database

	adjusted R2	p
Oxygen + Temperature	0.64	*
Oxygen with Temperature as condition variable	0.46	**
Temperature with Oxygen as condition variable	0.75	non sign.

TABLE 5.2: The percentage of variation explained of each explanatory physicochemical variable, as well as their combinations. *:<0.1, **:<0.05

of NCBI, among the binned contigs 53,536 were bacterial and 2,230 archaeal while 60 contigs were assigned as viral and 739 as eukaryotic. The corresponding numbers for the case of the unbinned contigs were 1-2 orders of magnitude higher; thus, the number of unbinned contigs were 430,028 bacterial, 126,690 archaeal, 15,828 eukaryotic and 1,805 viral correspondingly (see Figure S4).

5.4.3 MAGs phylogeny, functional annotation and distribution across samples

In line with the quality definitions described in [Bowers et al., 2017], metagenome binning generated a total of 289 MAGs; details are shown in the Supplementary File 2. According to the CheckM software 194 MAGs were reconstructed with a completeness higher than 90% and a contamination lower than 5% and all the rest had a completeness >70% and a contamination score <6%. According to the anvio summary (using the co-assembly as contigs database, the merged samples as profile and the reconstructed MAGs, i.e. the refined bins, as a collection) the redundancy of 10 (bin127, bin114, bin156, bin243, bin268, bin276, bin12, bin269, bin252, bin226) of the reconstructed MAGs was >10%. After the manual refinement of these 10 MAGs, a total of: (i) 178 bacterial high quality (completeness > 90%, contamination <5%), (ii) 70 bacterial and 39 archaeal medium quality (50% < completeness <90% and 5% <contamination <10%), and (iii) 2 bacterial MAGs of low quality (bin263 and bin182 with a completeness score <50%) were retrieved. Combining the anvio summary results (see bin_by_bin folder in SECOND_SUMMARY) and the Barrnap outcome (see arc_rrnas and bac_rrnas on GitHub repo), the 16S rRNA gene was identified in 100 out of the total 250 bacterial MAGs. Likewise, from a total of 39 archaeal MAGs, 16S rRNA gene was found in 28 of them. Contigs included on those MAGs represented 1.03 Gbp of assembled reads. A set of 25 MAGs had a completeness of 100% and contamination less than 5% while 5 bacterial (MAG 143, MAG 66, MAG 129, MAG 189 and MAG 76) and 1 archaeal (MAG 232) MAGs among them had a contamination of 0%. Overall, bacterial MAGs had higher completeness scores.

5.4.4 MAGs phylogenomic placement

The GTDB-Tk returned phylogenetic trees of the GTDB partition and the MAGs assigned to the corresponding domain for the cases of bacteria and archaea including the 2 low quality included (bin_182 : Proteobacteria and bin_263: Verrucomicrobiota). The phylogeny of the reconstructed MAGs (Figure 5.3) was built based on single-copy genes present on both Archaea and Bacteria, using the total number of MAGs even if some of the MAGs did not have all the 25 single-copy genes. Although not all these 25 single-copy

genes were found in every MAG, still, the mean number of occurrences of a gene among the 289 MAGs was 266.68, ranging from 211 to 278. In general, the archaeal MAGs had the fewer single-copy genes, most probably due to their lower completeness. The number of MAGs in which a single-copy gene was found ranged from 211 to 278 (mean = 266.68). Using the total number of MAGs the phylogeny of the reconstructed MAGs highlight the robustness of the method as even for those MAGs the phylogenetic signal was enough to place them among the representatives of their phylum. Thorough investigation of the tree pointed out that the only two discrepancies were that the sole MAG of the RBG-13-61-14 phylum (bin_124) that was placed among the Myxococcota representatives and that a representative of the Patescibacteria phylum (bin_61) was not placed close to the rest of Patescibacteria but as the closest relative of the representatives of the Chloroflexota phylum. The novel candidate phylum (bin_202) was placed within the same clade with Eisenbacteria (bin_31). In general, bootstrap values were >90% with only exceptions a number of clades with representatives of the Nanoarchaeota phylum (of which the completeness and the number of single-copy genes present was relatively lower).

5.4.5 Distribution of MAGs across samples

Based on the taxonomic end results of the shotgun metagenomic survey (Figure S5; MAG abundances per sample (MetaWRAP)), the most phylum was Bacteroidota (28% on average), which almost dominated sample Elos03 (57%) and Elos07 (40%). The second most abundant phylum was Proteobacteria (13% on average), with abundances ranging from 2% in Elos02 to 23% in Elos01 and 22% in Elos07. Planctomycetota and Desulfobacterota were found at about 8% and 7% respectively, with the latter being absent from Elos03 and very rare in Elos07 (2%). The only phylum that was present only in the microbial aggregates, i.e. in Elos03 and Elos07, and was absent from all the other samples was Myxococcota. The most abundant archaeal phylum was Nanoarchaeota (5% on average), which was mostly found in Elos01 (16%) and in much lower abundances in the other samples. Thermoplasmatota and Asgardarchaeota were found in similar abundances (3% and 2% on average respectively) and they were also absent from Elos03 and Elos07. Halobacteriota (2% on average) were not found in Elos10 and Elos12 and were mostly present in Elos01 (6%). Elos10 was the sample with the highest number of bins (Figure S6) even if it was the one with the lowest number of reads. In addition, it seems to be closer to Elos02, the bottom layer sediment sample from July, and to sample Elos12. The microbial aggregates (Elos03 and Elos 07) form another cluster, distinct from the other samples, but closer to Elos01, the microbial mat sample. The MAG 202 that represents a novel phylum is present in samples Elos12, Elos10 and Elos02.

5.4.6 Functional annotation of MAGs

The reconstructed MAGs were annotated with KofamScan with a range of KEGG ORTHOLOGY terms ranging from 354 to 2,879 terms (Figure S7), leading to 1 to 87 complete KEGG modules (Figure S8). The archaeal MAGs had, in general, lower completeness scores, lower number of KO terms assigned and less complete modules. As it is shown in Figure S9, MAGs form distinct clusters both based on the taxonomy, i.e. if they are

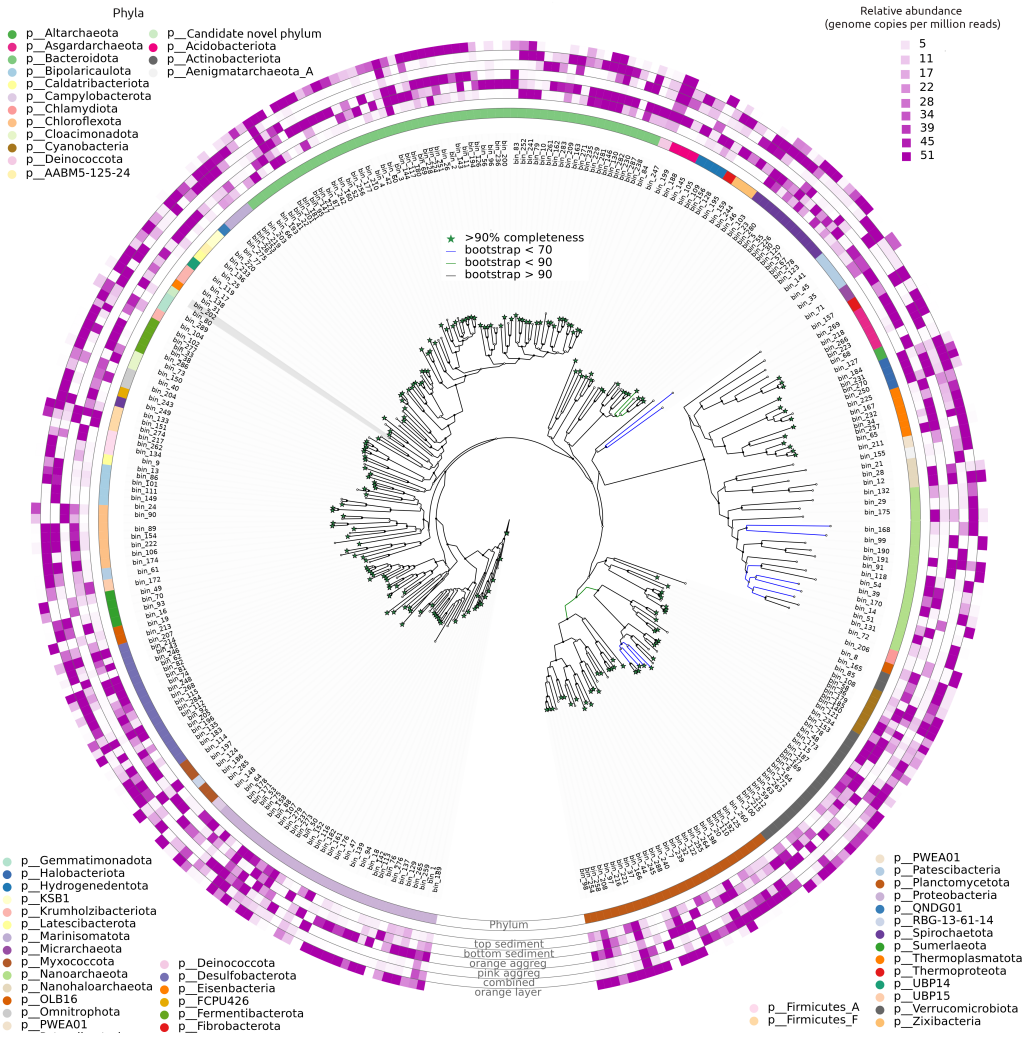


FIGURE 5.3: Concatenated marker gene phylogeny of the Karpathos' marsh MAGs. Phylogeny of the 289 MAGs recovered from a hypersaline marsh in Karpathos island, based on 25 concatenated, single-copy genes present both in Archaea and Bacteria. From inside to outside, the concentric circles around the phylogeny indicate: the bin id, phylum level taxonomy, bin relative abundance in the i) top layer of the sediment (i.e. the microbial mat), ii) bottom layer of the sediment, iii) orange aggregate, iv) pink aggregate, v) combined sediment layers (black upper layer) and vi) combined sediment layers (orange upper layer). Stars indicate high quality bins. The novel phylum (bin_202) is highlighted in grey.

bacterial or archaeal (PERMANOVA: F.Model = 35.767, p<0.001), as well as based on their completeness (PERMANOVA: F.Model = 5.2156, p<0.001). The modules that contribute most to this clustering, as identified by the simper analysis, as well as the significance of any given module's contribution, are shown in Table S3. Examples of these modules are related to oxygenic photosynthesis and nitrogen, sulphur and carbon cycles. When examined separately, again they differ significantly by completeness (PERMANOVA: Bacteria: F.Model = 3.4053, p < 0.001; Archaea: F.Model = 2.4452, p < 0.001).

5.4.7 Comparison of taxonomies between amplicon and metagenomic analysis

As expected, the taxonomic composition of the microbial communities in our samples depends on the analytical procedure that was followed in each case. However, when the similarity matrices of the samples are compared, the pattern deriving from the relative abundances of the microbial OTUs as derived from the amplicon survey, is highly correlated with the one deriving from the shotgun metagenomic survey (Mantel statistic: r=0.83, p<0.001). The pattern deriving from the Kraken2 analysis is also correlated both with the amplicon survey, as well as with the end result of the shotgun metagenomic survey (Mantel statistic: r=0.45, p<0.05; r=0.52, p<0.05, respectively), but on a lesser degree.

5.4.8 Physicochemical analysis

The physicochemical variables are given in Table S4. According to the variation partitioning analysis (Table 5.2), the combination of oxygen and temperature explained 64% of the total variation in the Kraken2 community similarity matrix. For the other cases, i.e the amplicon data matrix and the end result of the shotgun metagenomic analysis, residuals were higher than 0.60 and therefore no significant models were retrieved.

5.4.9 Functional profiles at the sample level

A set of 783,693 unique proteins were predicted from a total of 3,532,725 hits with NCBI COGs. The MEGAHIT assembler as implemented in the framework of DiTing, returned the assembly of each sample ranging from 1,323,538 (Elos03) to 2,773,933 (Elos10) contigs. As shown in Figure 5.4, pathways belonging to the carbon cycle, central metabolism and other metabolism are the most abundant (23%, 19% and 18% of the total pathways on average respectively). More specifically, the Reductive citrate cycle and the Dicarboxylate-hydroxybutyrate cycle dominate the carbon cycle (8% and 5%, 18% respectively). In the central metabolism, pathways like the Embden-Meyerhof glycolysis pathway and tricarboxylic acid (TCA) cycle are most abundant (7% each). In contrast, the Entner-Doudoroff pathway, i.e. an alternative glycolytic pathway, is found in abundances an order of magnitude lower than the Embden-Meyerhof glycolysis pathway (Figure S10). Wood–Ljungdahl pathway that enables the use of hydrogen as an electron donor, is mostly found in bottom sediment layer sample (Elos02), but also in the combined sediments from the November 2019 sampling (Elos10 and Elos12) and to a lesser extent in the other

samples. Bacterial chemotaxis, flagellar assembly and dissimilatory arsenic reduction are also among the most abundant pathways (9% and 7%, 3% respectively). Regarding the methane cycle, methanogenesis pathways were found in the bottom sediment layer sample (Elos02), but also in the combined sediments from the November 2019 sampling (Elos10 and Elos12), as the Wood–Ljungdahl pathway; the most abundant methanogenesis pathway was the formation of methane from acetate. Interestingly, methane oxidation was almost absent in the samples. Regarding the sulphur cycle, the most abundant pathways were the assimilatory and dissimilatory reduction of sulphate to sulphite (1% each). As shown in Figure 5.5, thiosulphate oxidation as well as sulphite oxidation, but to a lesser extent, contribute to the sulphate pool. Sulphur disproportionation to sulphide and sulphite is absent in the aggregate samples, i.e. Elos03 and Elos07. In addition, although DMSO reduction is an abundant pathway in all the samples, DMS oxidation is very rare and mostly found in Elos03, i.e. the orange aggregate (Figure S11). As shown in Figure 5.6, dissimilatory nitrate reduction to nitrite and nitrite to ammonia (DNRA) is prevalent in all samples, but it is mostly found in the combined sediment samples (Elos10 and Elos12). Denitrification (i.e. nitrite to nitric oxide and nitric oxide to nitrous oxide) is mostly found in sample Elos12. Nitrification, i.e. conversion of hydroxylamine to nitrite, is almost absent from samples Elos01 (the microbial mat) and the microbial aggregates (Elos03 and Elos07). Nitrogen fixation is mostly abundant in the sample Elos07 (the pink aggregate). Anaplerotic genes were also very abundant in our samples (2%), and in particular the Pyruvate Carboxylase Pathway which produces oxaloacetate from pyruvate and replenishes the intermediates of the TCA cycle.

5.5 Discussion

Overall, there seems to be an agreement in all three ways that taxonomic composition has been derived for our samples. Despite the fact that six out of the eleven samples were sequenced using shotgun metagenomics, still we can compare them with the amplicon data and derive conclusions. In fact, the amplicon survey produced results that were highly correlated to the results of the classification of the retrieved bins; such congruence between the two methodologies has been reported in recent studies [Chan et al., 2015, Regalado et al., 2020] although there is also evidence for the opposite [Tessler et al., 2017]. The results from Kraken2 were also correlated to the amplicon and the results of the classification of the retrieved MAGs, although the correlation was not as strong. However, Kraken2 is not restricted in the short amplicon length [Johnson et al., 2019] and therefore it was able to reach species level resolution [Lu and Salzberg, 2020].

The top sediment layer, i.e. the microbial mat, was characterised by the presence of Halobacteria/Halobacteriota, Halanaerobiia and Nanoarchaeota which are all halophilic and were therefore able to withstand the salt crust that was on top of the mat [Norton et al., 1993, Casanueva et al., 2008, Cinar and Mutlu, 2020, Akpolat et al., 2021] They are almost absent in the deeper sediment layers and they are absent in the winter, when there was no obvious microbial mat formed, despite the high salinity in sample Elos12. Therefore, the limiting factor for their presence could have been the lower temperature during the winter season, since the optimum temperature for representatives of the Halobacteria is

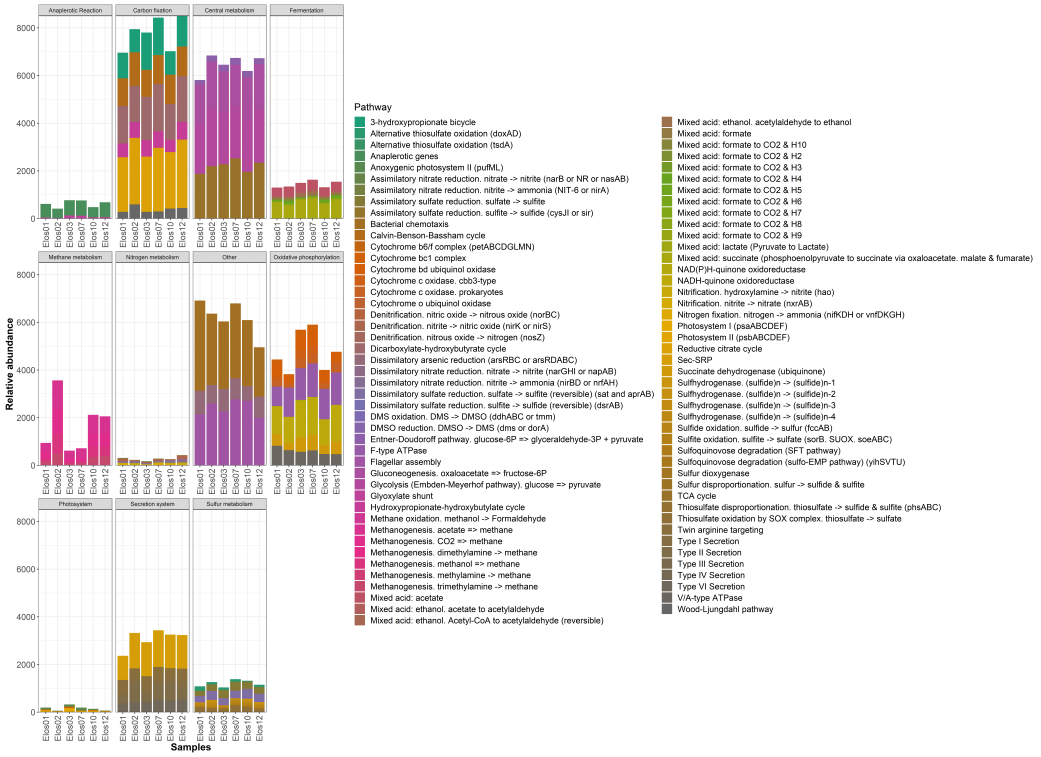


FIGURE 5.4: Bar chart showing the abundances of the metabolic pathways per biogeochemical cycle, at each sample, as retrieved from DiTing.

higher than 35°C [Grant, 2015]. Halobacteria, and their bacteriorhodopsin, are most likely responsible for the purple layer in the microbial mat, since they have been shown to cause striking red colours in salt flats [Stoeckenius et al., 1979]. Nanoarchaeota are obligate symbionts and they have found in association with Crenarchaeota/Thermoproteota [Huber et al., 2002, Podar et al., 2013, Munson-McGee et al., 2015, Wurch et al., 2016, Merkel et al., 2017]. Out of the known host-symbiont pairs and the putative hosts that have been proposed, we identified two of the known hosts (*Ignicoccus hospitalis* and *Acidilobus* sp. 7A) and five of the putative hosts [Jarett et al., 2018]. However, the Nanoarchaeotal representatives were assigned at taxonomies higher than the species level. Thus, we can only presume that the symbionts *Nanoarchaeum equitans* and *Nanopusillus acidilobi* of the aforementioned hosts are present in our data.

It was hypothesised that the microbial aggregates would be more closely related to the microbial mat samples and this was confirmed by our results. However, the microbial aggregates were mostly characterised by the presence of Bacteroidetes/Bacteroidota and Myxococcota and the absence, or very low abundance, of Desulfobacterota, Thermoplasmatota and Asgardarchaeota.

In the photic zone of the microbial mats, phototrophic microorganisms transduce light into energy using either pigments (chlorophyll and/or bacteriochlorophylls) or retinal-based rhodopsins [Kurth et al., 2021]. In hypersaline microbial mats, phototrophy is possible because light can penetrate salt crusts and therefore the only limit-

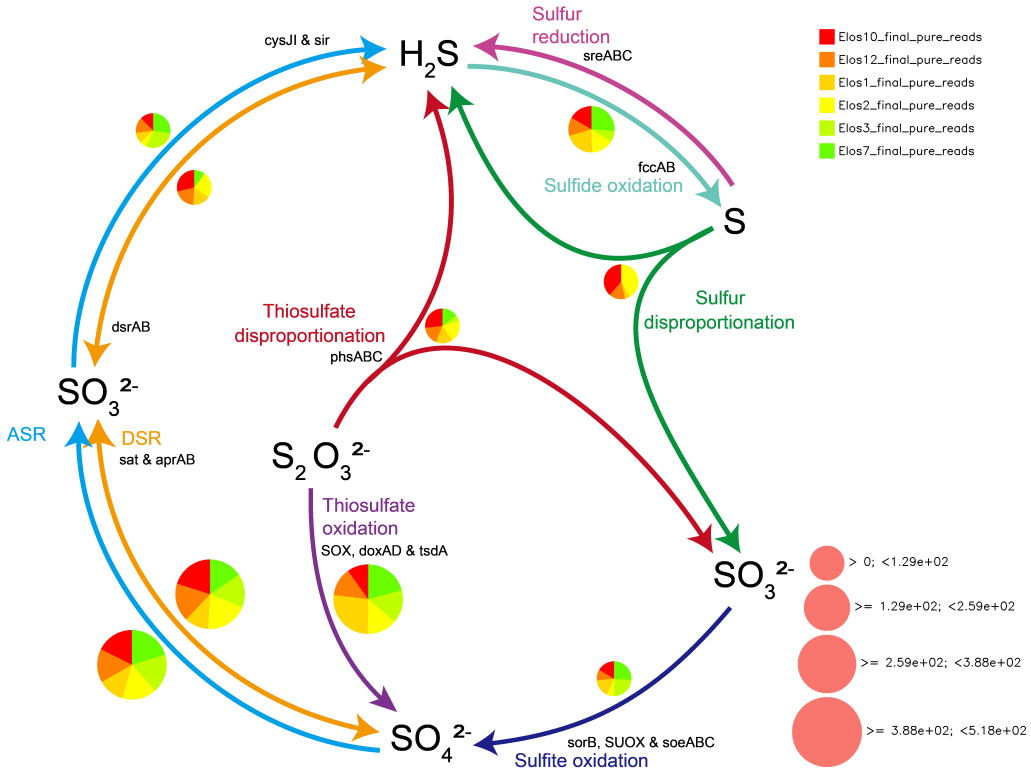


FIGURE 5.5: Relative abundances of the pathways involved in the sulphur cycle. The pie chart indicates the relative abundance of each pathway in each sample. The size of pie charts represent the total relative abundance of each pathway. ASR: assimilatory sulphate reduction; DSR: dissimilatory sulphate reduction.

ing factor is the capability of the microbial communities to actually perform phototrophy under salt-saturated conditions [Meier et al., 2021]. Out of the phyla that have been reported to perform (bacterio)chlorophyll-based phototrophy, i.e. Cyanobacteria, Proteobacteria, Chlorobi/Chlorobia, Chloroflexi/Chloroflexota, Firmicutes, Acidobacteria/Acidobacteriota, Eremiobacterota and Gemmatimonadetes/Gemmatimonadota [Zeng and Koblizek, 2017, Zheng et al., 2022], only Proteobacteria and Chloroflexi/Chloroflexota were found in high abundances. Cyanobacteria were very rare in our samples, which can be attributed to the high salinity of the marsh [DiLoreto et al., 2019] which can lead to osmotic stress and inhibition of photosynthesis [Sudhir and Murthy, 2004]. So, Cyanobacteria were not the foundation of the microbial mats in our study, as has been shown in other examples of hypersaline microbial mats [Bolhuis et al., 2014, Wong et al., 2016], which was expected as oxygenic photosynthesis is completely inhibited at saturation-level salinities (40%) [Meier et al., 2021]. However, cyanobacterial genera that are characterised by strategies and survival mechanisms that allow them to grow in such high salinities [Oren, 2015], such as *Euhalothece* and *Halothece*, were present in our samples. Chloroflexi/Chloroflexota were most present in the bottom sediment layer and in the combined sediment samples. Representatives that were identified from our samples are

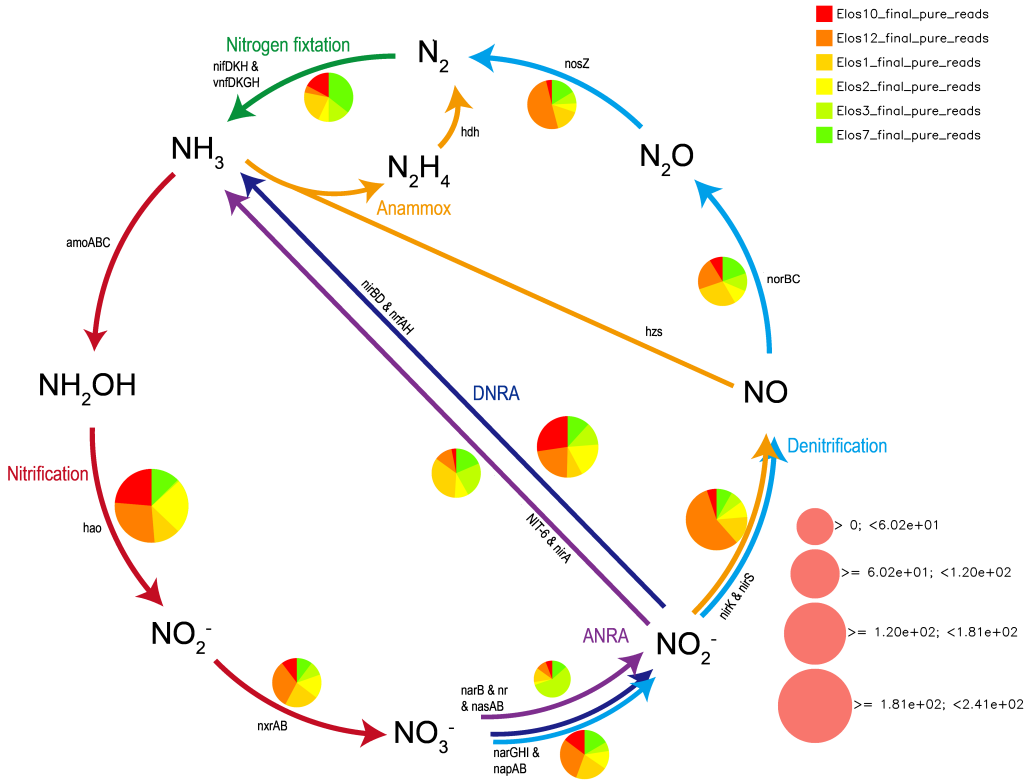


FIGURE 5.6: Relative abundances of the pathways involved in the nitrogen cycle. The pie chart indicates the relative abundance of each pathway in each sample. The size of pie charts represent the total relative abundance of each pathway. ANRA: assimilatory nitrate reduction to ammonium; DNRA: dissimilatory nitrate reduction to ammonium; Anammox: anaerobic ammonium oxidation.

Chloroflexus aurantiacus [Pierson and Castenholz, 1974] and *C. aggregans* [Hanada et al., 1995]; they can grow photoautotrophically and photoheterotrophically under anaerobic conditions but also chemotrophically under aerobic conditions, using sulphide or hydrogen as an electron donor [Tang et al., 2011, Kawai et al., 2019, 2021]. Given that it is unclear if they can harvest light deeper in the sediment and taking into account the low abundance of Cyanobacteria, it is most probable that Chloroflexi/Chloroflexota grow chemotrophically in our samples. According to our data, Proteobacteria seem to be taking over the photosynthetic pathways in our samples. More specifically, some of the species that have been found to perform photosynthesis and are present in our samples are (from the Alphaproteobacteria class) *Citromicrobium* sp. [Jiao et al., 2010], *Rhodopseudomonas palustris*, *Rhodobacter* sp., *Rhodospirillum rubrum*, *Roseobacter denitrificans*, *Bradyrhizobium* sp., *Roseobacter* sp. [Larimer et al., 2004, Bryant and Frigaard, 2006] and (from the Gammaproteobacteria class) *Marichromatium purpuratum* [Shiung et al., 2018], *Congregibacter litoralis* [Fuchs et al., 2007, Spring et al., 2009], *Allochromatium* spp. [Kyndt and Meyer, 2020]. Overall, based on the taxonomic composition of the studied microbial mats, it seems that they resemble microbial mats from an irregularly inundated

tidal flat in Oman [Meier et al., 2021].

Our second hypothesis was that samples from the deeper sediment layers collected in the summer would be more similar to the combined sediment samples collected in the winter. This was also confirmed by our results, as these samples were clustered together. However, despite the similarities between the samples, there were also certain differences; as expected, there is a high degree of spatial variation in the marsh under study [Dillon et al., 2009]. Overall, in the deeper sediment layers or where microbial mat was not formed, acetogens such as *Moorella thermoacetica*, *Clostridium aceticum* and *Acetobacterium woodii* [Schuchmann and Müller, 2016], were utilising the Wood–Ljungdahl pathway, i.e. using H2 hydrogen as an electron donor and CO2 as an electron acceptor. Acetogens are either competing directly with hydrogenotrophic methanogenic archaea or interacting syntrophically with acetotrophic methanogens [Ragsdale and Pierce, 2008]. Utilisation of acetate for methanogenesis is present in the genera *Methanosarcina* and *Methanothrix* (Ferry 1992) which are both found in our samples. Although while abundant in our samples, methanogens might have been contributing little to anaerobic mineralization, since in salinities of 180‰ or less they are inhibited by the increased activity of sulphate-reducing bacteria [Sørensen et al., 2004]. In addition, sulphate-reducing bacteria might have also been using the Wood-Ljungdahl pathway in reverse [Ragsdale and Pierce, 2008]. Sulphate-reducing microorganisms (SRM) from the Euryarchaeota lineage [Muyzer and Stams, 2008] were very abundant in the microbial mat and in the aggregates, while SRM belonging to Deltaproteobacteria [Muyzer and Stams, 2008] are present in the deeper and the combined sediment samples. However, the occurrence of SRMs is not synonymous to the occurrence of sulphate reduction in the given habitat [Muyzer and Stams, 2008].

It has been proposed that arsenic and sulphur cycling can sustain high microbial metabolic rates in permanently anoxic mats [Visscher et al., 2020]. Bacteria capable of performing anoxygenic photosynthesis using arsenite (As(III)) as an electron donor, such as *Ectothiorhodospira* sp. and *Halorhodospira halophila* [Hoeft McCann et al., 2017], are present in our samples. It is suggested that they are performing oxidation of As(III) to arsenate (As(V)), which is afterwards reduced back to As(III) (Hoeft et al. 2004), thus explaining the high occurrence of dissimilatory arsenic reduction in our samples. Arsenate can be an important electron acceptor in the biogeochemical cycling of carbon [Oremland et al., 2000]; thus, arsenate reduction has a great potential to precipitate carbonates and it is energetically better than sulphate reduction [Visscher et al., 2020], although the latter is also very abundant in our samples. Thaumarchaeota, which are involved in nitrification in marine ecosystems [Veuger et al., 2013], were also present in our samples but in very low abundances, in contrast to other hypersaline microbial mats [Ruvindy et al., 2016]. On the other hand, common nitrifying bacteria such as *Nitrobacter* where abundant in our samples; however, since nitrification was mostly present in the combined sediments and in the deeper layer, there seems to be a certain degree of hypersalinity limitation on the growth of nitrifying bacteria, as has been previously suggested [Jeffries et al., 2012]. Denitrification was also present in our samples, although mostly in the combined sediment, and it was not limited by the increased salinity [Laverman et al., 2007]. Regarding functional annotation of MAGs and their clustering according to taxonomy, it seems that it is driven by modules that are present in Archaea and absent in Bacteria, and vice versa. For example, regarding cysteine biosynthesis (M00021), this pathway is still unexplored in

Archaea and although it has been found in certain species, e.g. *Methanosarcina barkeri*, it is suggested that there might be a different cysteine biosynthesis pathway in Archaea [Kitabatake, So, Tumbula, and Söll, 2000]. Likewise, there are no archaeal homologs for the bacterial pantothenate biosynthetic genes [Ronconi et al., 2008], therefore pantothenate biosynthesis (M00913) is one of the pathways that contributes to the differentiation between bacterial and archaeal MAGs. In addition, acetogen (M00618) is only found in Bacteria.

5.6 Conclusions

It has been debated if hypersaline environments are thermodynamic limiting the occurrence of self-sustaining microbial communities [Oren, 2011] or if they are biologically permissive [Lee et al., 2018]. Cells need to implement strategies to counteract the osmotic stress [Gunde-Cimerman et al., 2018] but these strategies come with an energetic cost [Meier et al., 2021]. It has been suggested that hypersaline microbial mats and, in particular, communities below salt crusts, cannot rely solely on primary production from anoxygenic phototrophy and mineralization from sulphate reduction [Meier et al., 2021]. Instead, import of reduced substances and periods of reduced salinity are required, to allow the occurrence of oxygenic photosynthesis [Meier et al., 2021]; in our study site this occurs in winter where evaporation is not as strong as in the summer, which combined with the increased precipitation, lowers the salinity of the marsh. During winter months, both the salt crust and the layering of the microbial mat disappears, as in Cardoso et al. [2019], and it seems that this temporal change and seasonal development of the microbial mats under study is the necessary element for the survival of the microorganisms. In addition, anaplerotic reactions, that are abundant in our samples, may play an important role in replenishing the intermediates of the TCA cycle, which is quite abundant in out samples, and thus allowing microbial growth with a carbohydrate as the sole carbon source [Tong, 2013, Choi et al., 2016]. Although it seems that hypersaline environments are "thermodynamically moderate", DNA based studies can only identify the members of a community and not their metabolic activities. Therefore future studies on hypersaline microbial mats should focus on the combination of metabolomics, metatranscriptomics and metagenomics, in order to elucidate the functional repertoire of microbial communities, their metabolic potential and their metabolic and ecological interactions. Metabolic modelling of the microbial assemblages can shed further light on the effects of the environmental challenges on the mat construction as well as on which processes are taking place within each mat layer and among its different layers.

Supplementary Material

The Supplementary Material for this study will be publically available once it is published. For the time being, they can be found under this Google Drive folder.

Supplementary Figure 1: nMDS of the similarity matrix of the samples based on the microbial OTUs relative abundances, as retrieved from the amplicon sequencing.

Supplementary Figure 2: Bar chart showing the abundances of the main microbial taxa, at the phylum level, at each sample, as retrieved from Kraken2.

Supplementary Figure 3: Blobology scatterplot showing the contigs that were binned (blue) and the contigs that were not binned (grey).

Supplementary Figure 4: Blobology scatterplot showing the contigs that were binned and their taxonomic annotation. blue: Bacteria; yellow: Archaea; red: Eukaryota; grey: not assigned.

Supplementary Figure 5: Bar chart showing the abundances of the main microbial taxa, at the phylum level, at each sample, as retrieved from the classification of the retrieved MAGs.

Supplementary Figure 6: MetaWRAP heatmap showing the similarity and clustering of the samples based on the identified bins.

Supplementary Figure 7: Distribution of the number of KO terms annotated per MAG.

Supplementary Figure 8: Distribution of the number of the complete KEGG modules present per MAG.

Supplementary Figure 9: nMDS of the similarity matrix of the MAGs based on the presence/absence of their modules.

Supplementary Figure 10: Relative abundances of the pathways involved in the carbon cycle. The pie chart indicates the relative abundance of each pathway in each sample. The size of pie charts represent the total relative abundance of each pathway. CBB: Calvin-Benson-Bassham cycle; rTCA: reductive citric acid cycle; WL: Wood-Ljungdahl pathway; 3HB: 3-hydroxypropionate bicycle; DHC: Dicarboxylate-hydroxybutyrate cycle.

Supplementary Figure 11: Relative abundances of the pathways involved in the DMSP cycle. The pie chart indicates the relative abundance of each pathway in each sample. The size of pie charts represent the total relative abundance of each pathway.

Supplementary File 1: The parameters for the PEMA processing of the 16S rRNA amplicon data.

Supplementary File 2: Details of the MAGs' processing.

MAGs description

In Appendix C the methodology for the description of the reconstructed MAGs from the Tristomo marsh data can be found. The novel taxa found will be thoroughly investigated in the near future.

Chapter 6

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

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[1]

6.1 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 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.

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

6.2 Introduction

The ubiquitous marine environments (more than 70% of the global surface [NOAA, 2021]) mold Earth's conditions to a great extent. The interconnected abiotic [Falkowski et al., 2008] and biotic factors (from bacteria [Falkowski et al., 2008] to megafauna [Estes et al., 2016]), shape biogeochemical cycles [Arrigo, 2005] and climates [Boero and Bonsdorff, 2007, Beal et al., 2011] from local to global scales. In addition, marine systems have high socio-economic value [Remoundou et al., 2009] as an essential source of food and by supporting renewable energy and transport, among other services [Pörtner et al., 2019]. The study of marine environments involves a series of disciplines (scientific fields): from Biodiversity [Sala and Knowlton, 2006] and Oceanography to (eco)systems biology [Tonon and Eveillard, 2015] and from Biotechnology [Dionisi et al., 2012] to Aquaculture [Tidwell and Allan, 2001].

To shed light on the evolutionary history of (commercially important) marine species [Carvalho and Hauser, 1995], as well as on how invasive species respond and adapt to novel environments [Sakai et al., 2001], the analysis of their genetic stock structure is fundamental [Begg and Waldman, 1999]. Similarly, biodiversity assessment is essential to elucidate ecosystem functioning [Loreau, 2000] and to identify taxa with potential for bioprospecting applications [Leal et al., 2012]. Furthermore, systems biology approaches provide both theoretical and technical backgrounds in which integrative analyses flourish [Norberg et al., 2001]. 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 [Mardis, 2008, Kulski, 2016]. These technologies allowed access to the genetic, transcript, protein, and metabolite repertoire [Goodwin et al., 2016] of studied taxa or populations, and facilitated the analysis of organism-environment interactions in communities and ecosystems [Bundy et al., 2009]. Whole-genome sequencing and whole-transcriptome sequencing approaches provide valuable information for the study of non-model taxa [Cahais et al., 2012]. This information can be further enriched by genotyping-by-sequencing approaches, such as restriction site-associated DNA sequencing [Baird et al., 2008], or by investigating gene expression dynamics through Differential Expression (DE) analyses [Tarazona et al., 2011]. 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 [Goldford et al., 2018]. 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 [Merelli et al., 2014]. 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) [Sohn and Nam, 2018]. Hence, the challenges of how to exploit all these

data and how to transform data into knowledge set the present framework in biological research [Greene et al., 2014, Pal et al., 2020].

To address these computational challenges, the use of high-performance computing (HPC) systems has become essential in life sciences and systems biology [Lampa et al., 2013]. 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" [Sterling et al., 2017]. 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] [Wikipedia contributors, 2021]. 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 [Lindahl, 2018]. 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 [Lindahl, 2018]. Virtual Research Environments are web-based e-service platforms that are particularly useful for researchers lacking expertise and/or computing resources [Candela et al., 2013]. 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 [Rad et al., 2017] and Singularity [Kurtzer et al., 2017]—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.

6.3 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.

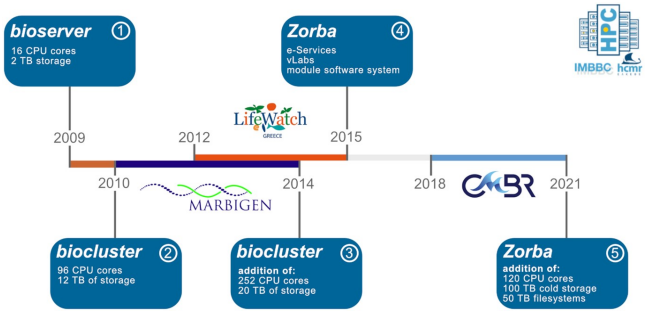


FIGURE 6.1: Evolution of the IMBBC HPC facility during the past 12 years, with hardware upgrades (blue boxes) and funding milestones (logos of RIs) highlighted. A single server that launched the bioinformatics era in 2009 evolved to the current Tier 2 system Zorba (Box 4), which allows processing of a wide variety of information from DNA sequences to biodiversity data. Different names of the facility denote distinct system architectures.

6.4 Methods

6.4.1 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. [Zafeiropoulos et al., 2021b]). 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 [Dongarra et al., 2003]. 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. [Zafeiropoulos et al., 2021b].

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

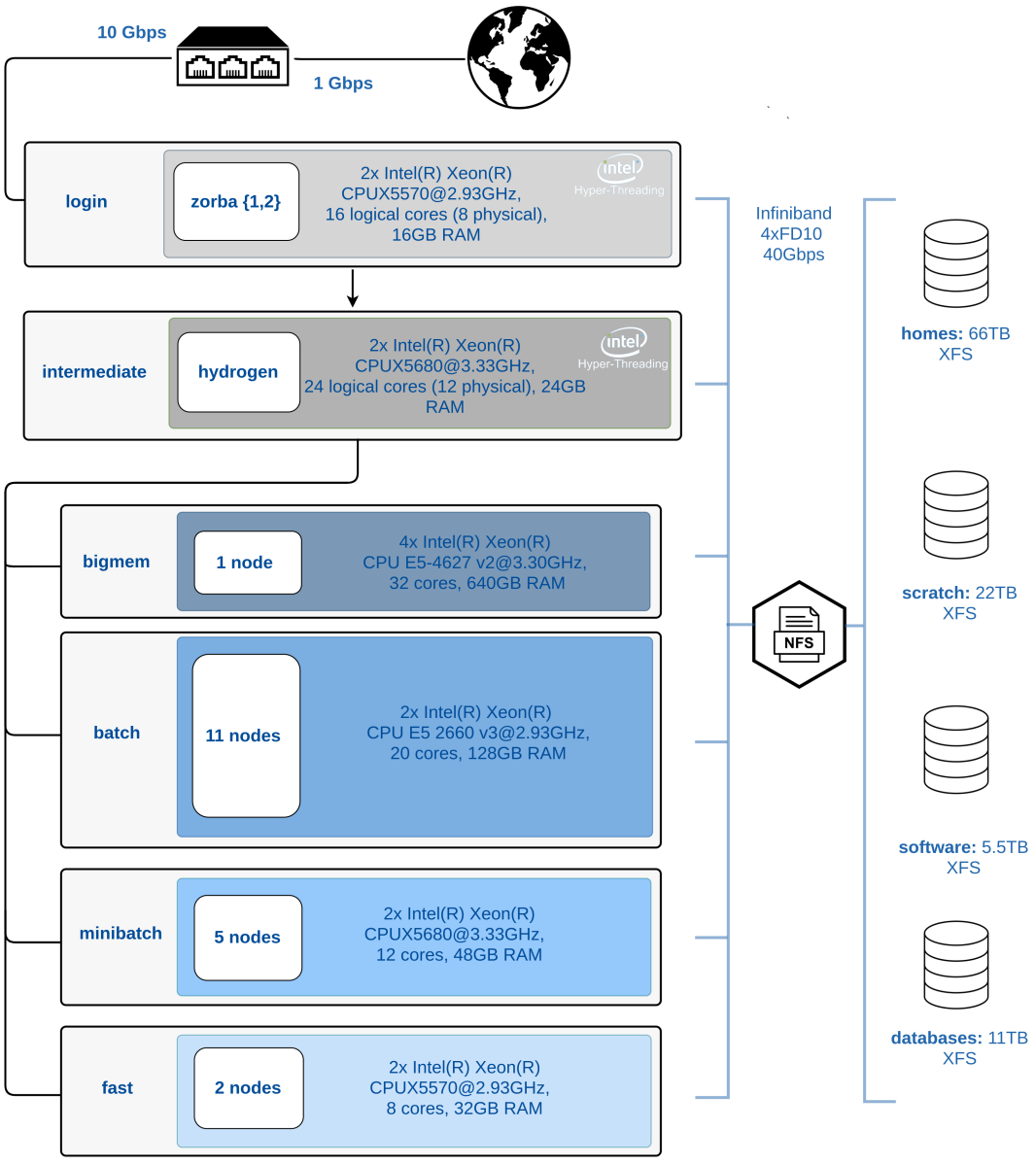


FIGURE 6.2: Block diagram of the *Zorba* architecture. This is the IMBBC HPC facility architecture in its current setup, after 12 years of development. There are 2 login nodes and 1 intermediate where users may develop their analyses. Computational nodes are split into 4 partitions with different specs 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

modeling, to name a few. Access to these packages is provided through Environment Modules, a broadly used means of accessing software in HPC systems [Castrignanò et al., 2020].

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) [Zafeiropoulos et al., 2020], the chained and automated use of STACKS, software for population genetics analysis from short-length sequences [Catchen et al., 2013] (latest version), a set of statistical functions in R for the computation of biodiversity indices, and analyses in cases of high computational demands [Varsos et al., 2016], 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. [Zafeiropoulos et al., 2021b].

Zorba's daily functioning is ensured by a core team of 4 full-time, experienced staff: 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. [Zafeiropoulos et al., 2021b]). An extended description of user and job administrations and orchestration can be found in Section C1 of Zafeiropoulos et al. [Zafeiropoulos et al., 2021b]).

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. [Zafeiropoulos et al., 2021b]). 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. [Zafeiropoulos et al., 2021b].

6.5 Results

6.5.1 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 (see Section D2 in Zafeiropoulos et al. [Zafeiropoulos et al., 2021b]). 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. [Zafeiropoulos et al., 2021b]).

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 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 [Zafeiropoulos et al., 2021b]).

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

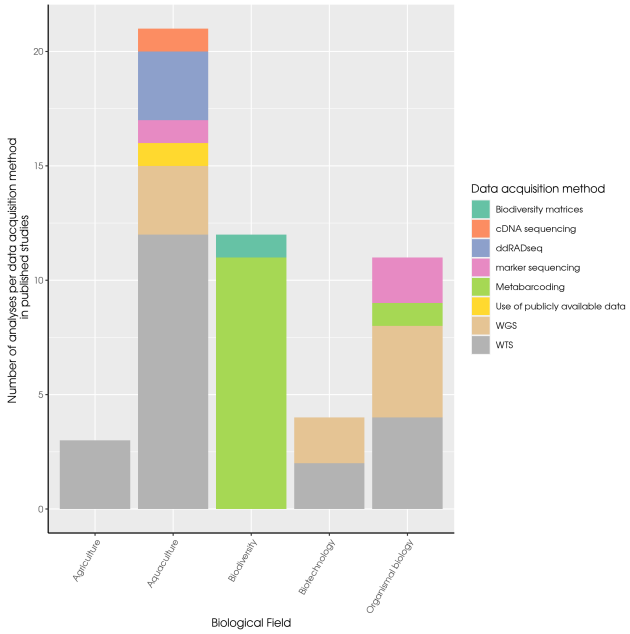


FIGURE 6.3: Bar chart with the number of publications that have used IMBBC HPC facility resources, grouped by scientific field. The different methods for data acquisition are also presented. WGS, whole-genome sequencing; WTS, whole-transcriptome sequencing.

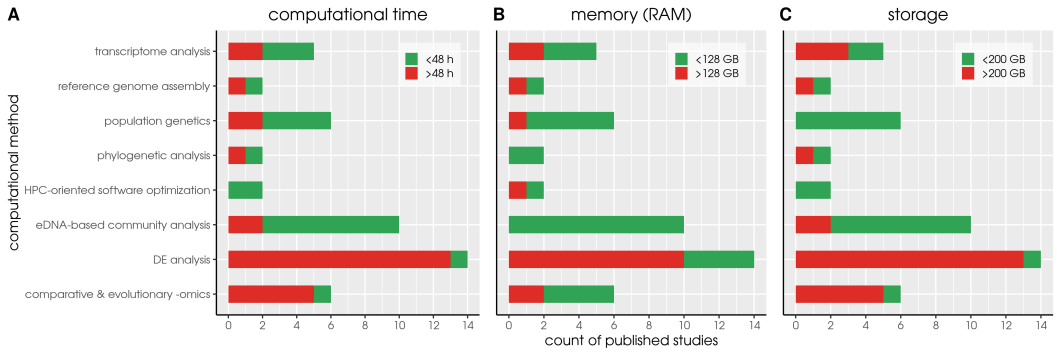


FIGURE 6.4: Red bars denote published research with high resource requirements of the various computational methods employed at the IMBBC HPC facility due to (a) long computational times (> 48 h), (b) high memory requirements (> 128 GB), or (c) high storage requirements (> 200 GB). For instance, no eDNA-based community analyses performed at *Zorba* thus far have required a large amounts of memory.

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. [Zafeiropoulos et al., 2021b].

6.6 Discussion

6.6.1 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 [Klymus et al., 2017], 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 [Bariche et al., 2020]. 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 Observation Network). PEMA software [Zafeiropoulos et al., 2020] was used in this study, as well as in the 30-month pilot phase of ARMS-MBON [Katsanevakis et al., 2014].

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* [Pauletto et al., 2018] or the greater amberjack *Seriola dumerili* [Sarropoulou et al., 2017]. 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 [Tsakogiannis et al., 2020]. 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 [Danis et al., 2020]. In the context of this study, an automated containerized pipeline allowing high-quality genome assemblies from Oxford Nanopore and Illumina data was developed (SnakeCube [Angelova et al., 2021]). 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. [Natsidis et al., 2019]).

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 [Papadaki et al., 2020].

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 [Warwick and Somerfield, 2008]. In Arvanitidis et al. [Arvanitidis et al., 2018], the average taxonomic distinctness index ($\Delta+$) and its variation ($Lambda+$) were calculated on a matrix deriving from the complete World Register of Marine Species [Vandepitte et al., 2018], 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 [Gioti et al., 2020], transcriptome assembly and a phylogenetic analysis revealed the full diversity of the symbiont's multicopper oxidases, enzymes commonly involved in oxidative degradation [Nikolaivits et al., 2021]. 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.

6.6.2 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 [Zafeiropoulos et al., 2021b]). 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 [Lampa et al., 2013]. 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 [Zafeiropoulos et al., 2021b]).

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 [Lampa et al., 2013]). 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 [Dagum and Menon, 1998]. 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 [Vouzis and Sahinidis, 2011, Nobile et al., 2017], 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 [Mell et al., 2011]. 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 [Langmead and Nellore, 2018]). 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 [Zafeiropoulos et al., 2021b]), similarly to other HTS-oriented HPC systems [Lampa et al., 2013]. 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 [Dahlö et al., 2018], in order to allow even more efficient use of computational resources.

6.7 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 [Lampa et al., 2013], 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

7.1 Bioinformatics approaches enhance microbial diversity assesment based on HTS data

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

The 16S rRNA gene has been used for decades as the golden standard for the study of microbial communities. It has been shown that the full-length 16S sequence combined with appropriate treatment of the intragenomic copy variants has the potential to provide taxonomic resolution of bacterial communities even at the strain level [Johnson et al., 2019]. However, when the region is chosen carefully and a thorough alignment procedure is applied, even short short reads may return phylogenetic information comparable with the one fron full-length 16S rRNA reads [Jeraldo et al., 2011]. This was also shown in Chapter 5 as the 16S rRNA amplicon analysis was in line with the taxonomy assignment of the shotgun reads.

Even if amplicon studies have proven themselves essential for the assessment of microbial diversity, the bioinformatics analaysis in such studies, usually comes with several issues; with the lack of parameter tuning being among the most crucial ones. As shown in Chapter 2.1 where mock communities were used to validate the PEMA results, it is parameter tuning that determines the precision and recall scores in such analyses. Sequencing mock communities along with the rest of the samples allows the tuning of the bioinformatics analysis based on a known assemblage and thus, it enables parameter tuning based on the idiosyncracy of each particular experiment/study [Bokulich et al., 2020].

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

invoked with any local database is possible.

Moreover, it is also commonly known that pseudogenes as well as nuclear mitochondrial pseudogenes (numts) can lead to several biases in such studies [Song et al., 2008]. To address this challenge multiple computational efforts have been implemented [Porter and Hajibabaei, 2021] This issue also applies for the case of Bacteria and Archaea and the 16S rRNA gene [Pei et al., 2010] even if it has been shown that bacterial pseudogenes have a great chance of being removed almost directly after their formation; so fast that to be governed by a strictly neutral model of stochastic loss [Kuo and Ochman, 2010]. As shown in Chapter 2.2, a great part of the OTUs/ASVs retrieved from COI amplicon data may actually come from bacterial and/or archaeal taxa. Such approaches need to be merged in amplicon studies as an extra quality control stop but also to enable further investigation of the unassigned OTUs/ASVs. In Chapter 2.2 is also shown the need for reference databases to also include non-target sequences so they can distinguish actual hits.

However, there is still a major question regarding the microbial diversity assessment; how could HTS methods be used to recognise novel taxa and their metabolic potential? As shown in Chapter 5, the reconstruction of MAGs from shotgun metagenomics data may play a great role in the description of unknown and currently uncultivated taxa. Such studies and their corresponding MAGs have enriched our knowledge on the tree of life to a great extent over the last few years, uncovering several prokaryotic phyla, leading to radical challenges on their taxonomy and the taxonomy scheme [Parks et al., 2022]. Long-read sequencing technologies such as Nanopore and PacBio, have improved their accuracy to a great extent, offering high-quality, cutting-edge alternatives for testing hypotheses about microbiome structure and functioning as well as assembly of eukaryote genomes from complex environmental DNA samples [Tedersoo et al., 2021].

7.2 Containerization technologies and e-infrastructures provide the means for computational capacity and reproducibility

As shown in Chapter 6 the computing resources required for the analysis of several microbiome studies may range from those covered by a personal computer to overwhelming the capacity of Tier-2 HPC facilities; this also applies for any biological topic using HTS data. On top of that, as also shown in Chapter 6, the maintance of bioinformatics-oriented HPC facilities comes with a great number of challenges.

By encapsulating a software along with its dependencies in an isolated and easy to reinstall environment (container) containerization addresses several of them at once; first, the distribution and the installation becomes now a straight-forward task, requiring only for a containerization technology present on the facility, second, versioning of the various software used is not an issue anymore as a container may either "save" a version from being obsolete in case it is strongly dependent on that, either keep track of the latest version of the encapsulated software, morevover, several versions of the same software may be part of different containers without any conficts. In addition, the creation and management of standardized workflows/pipelines is facilitated to a great extent. Workflow

tools such as Common Workflow Language (CWL) [1], Snakemake [2] and Nextflow [3] have been proven of high value in building such pipelines as they support the connection of multiple independent software. MGnify [Mitchell et al., 2020] is a great example of this case.

However, more often than not, such workflows require major computing resources to analyse real-world microbiome data sets. To this end, HPC facilities and cloud solutions are required. Therefore, efforts such as those discussed in Section 2.1 for containerised tools such as the PEMA workflow [Zafeiropoulos et al., 2020] to be integrated in e-infrustructures, are rather significant. This way, reproducability is secured and analyses that cannot be performed in a personal computer is accessible to researches that have no access to local servers or HPCs. Further insight on how to address computing-oriented challenges on the analysis of microbiome data are discused in [Hu et al., 2022].

7.3 High quality metadata enable efficient exploitation of sequecning data in a meta-analysis level

It is common knowledge that both shotgun and long-read metagenomics provide high quality data enabling the study of real-world microbial communities even at the strain level [Meyer et al., 2022]. However, these data are not fully exploited unless they come with thorough and standardized metadata; indicatively it has been shown that more than the 20% of the metagenomes published between 2016 and 2019 were not even accessible [Eckert et al., 2020]. To address such challenges, community-driven initiatives can be of great importance [Yilmaz et al., 2011b, Vangay et al., 2021, Hu et al., 2022]. Similar initiatives focusing on specific topics, e.g. protocols in the framework of the Ocean Best Practices System [Samuel et al., 2021], are essential.

FAIR principles have set a new era on how data are stored and ditributed. Data and metadata provenance goes even further by allowing not only the reuse of the data, but also keeping track of where the data came from, potential edits, as well as of the analyses that are linked to those, both regarding their outcome and the analysis *per se*. Ensuring FAIR (meta)data could be considered as the first step of data integration [Freire et al., 2008, Deelman et al., 2010]. Nevertheless, (meta)data provenance and data integration share several challenges [Cheney et al., 2009].

As shown in Section 3, the higher the quality of the accompanying metadata, the higher the confidence for meta-analysis approaches may get. Furthermore, metadata accompanying the analyses, i.e. the workflows and their implementation, can provide further insight on the effect of the various softwares and databases used. In general, community efforts to set best-practice for formal metadata description and their use for packaging research data with their accompanying metada such as the one of RO-Crate may have a great impact. Focusing on the microbiome community, further challenges need to be addressed to this end, with taxonomy and nomenclature being among the most crucial ones. The GTDB [Parks et al., 2022] that provides *a phylogenetically consistent*

[1] https://github.com/common-workflow-language/common-workflow-language
[2] https://snakemake.github.io
[3] https://www.nextflow.io

and rank normalized genome-based taxonomy for prokaryotic genomes sourced from the NCBI Assembly database combined with efforts such as those of Pallen et al. to ensure valid names for novel taxa that will keep being discovered over the next years, could play a great role on addressing this issue. However, moving all the so-far knowledge in a specific format is far from easy. At the same time, as discussed in 3, bringing together data of different types can return great insight by either generate hypotheses or further supporting hypotheses such as the one of Pavloudi et al. on the potential use of various electron acceptors from the different strains present in different environmental types (see 3.4.2). This becomes clearer considering that up to now only a small fraction of the microbial diversity has been described and named (24,000 species [Parte et al., 2020]) with almost 10,000 of them to have done so, over the last 6 years (see LPSN statistics [4]).

Moreover, linking the various ontologies related to a certain domain without loosing the benefits of each of those, is essential for the community. That means that efforts for mapping entities of an ontology or a database to those of others sharing a common field, even if there is not a one-to-one relationship, will increase considerably their impact. Efforts such the one of the Rhea reaction knowledgebase [Bansal et al., 2022]

By addessing these challenges and by ensuring high quality metadata, data integration techniques will be improved considerably. Combined with machine learning techniques, such methods can contribute the most in exploiting the full potential of the HTS data produced and the so-far knowledge for the utmost biological insights that can be drawn [Noor et al., 2019].

7.4 Markov Chain Monte Carlo approaches enable flux sampling at the microbial community level

Metabolic modelling and genome-scale metabolic models in particular, provide a great framework to study the genotype - phenotype relationship [Lewis et al., 2012]. Thus, it enables the investigation on how a species respond under changing environmental conditions too [Herrmann et al., 2019]. Flux sampling on microbial GEMs has been proved the most valuable for the identification of specific reactions that are transcriptionally regulated [Bordel et al., 2010] or required under certain conditions for the species to survive [Herrmann et al., 2019]. But also in the study of the variant by-produts produced by the different strains of a species [Scott et al., 2021].

As shown in Chapter 4 the higher the dimension of the polytope derived from a metabolic model, the more challenging the sampling on its flux space gets. The dimension of a polytope derived from any single microbial GEM is at least one order of magnitude lower than the one from species such as *H. sapiens*. However, in real-world microbial communities it is rare for a species to be on its own. Based on Perez-Garcia et al. [2016] there are various approaches for modelling a community as a whole, each coming with several pros and cons. As the *lumped network* approach neglects microbial diversity dynamics and the dynamics of their corresponding processes, assuming a *super-organism* where all species share the same reactions and exploit their environment

[4] https://lpsn.dsmz.de/statistics/figure/10

in the same way, it cannot be used for the study of microbial interactions. To this end, models that integrate a GEM for every species present, taking into account the relative abundance of each species, and also support the exchange of compounds between each species and the environment are required [Diener et al., 2020]. Dynamic versions of such models also allow the changes in the biomass concentrations of each individual species to be taken into account [Zhuang et al., 2011]. This way, ecological interactions can be inferred; illustratively, Zhuang et al. demonstrated how the concentration of acetate in the environment can define whether *Rhodoferax* or its competitor *Geobacter* will survive. Approaches like COMETS [Dukovski et al., 2021] and MICOM have been essential towards this direction. However, flux sampling has not been merged yet to large-scale approaches mostly because of the computational challenges that arise as the dimension of the polytope that derives from a model increases. Therefore, approaches such as the MMCS algorithm desribed in Chapter 4 may benefit the community to this end.

7.5 Hypersaline mats host a great range of novel taxa & their functioning might be subject to anaplerotic reactions

As discussed in Chapter 5.2 hypersaline mats have been studied to a great extent over the last decades. Several novel taxa have been identified for the first time in such environments. It is well established that in such communities a certain organisation of the taxa present are stratified [Saghaï et al., 2017] and that photosynthesis plays a great part for the mat communities as a total [Oren, 2015].

Thanks to their geomorphological and biogeochemical heterogeneity they provide a profuse number of ecological niches for microorganisms leading to highly diverse communities multiple taxa of which have not yet recorded. This was demonstrated to a great extent in Chapter 5.2 (see also Appendix A.3) where more than 200 novel taxa have been found and about 65 of them have been described, highlight also the potential of shotgun metagenomics.

In their study, Lee et al. demonstrated that NaCl saturated environments can be considered "thermodynamically moderate", meaning they are capable of active and complete element cycling and thus they can be considered as self-sustaining systems. To that end and assuming that sunlight is the sole energy source of the system, primary production of organic matter by oxygenic phototrophs is essential for the community to survive [Meier et al., 2021]. That is further supported by the the fact that descpite the fact that oxygenic photosynthesis is thermodynamically possible at high salt concentrations [Oren, 2011], a kinetic inhibition of oxygenic photosynthesis by decreased oxygen solubility [Abed et al., 2007]. Thus, Cyanobacteria have been found to be fundamental in hypersaline microbial mats; even when their abundance is relatively lower they are assumed to remain the major group of primary producers [Bolhuis et al., 2014]. However, in the Tristomo marsh mats, several phototrophic Proteobacteria were identified suggesting they also play a great part in fueling the system. To investigate what is actually the case, further studies are required (see Conlusion 7.6). Moreover, anaplerotic reactions were identified in high abundances across the samples. It is commonly known that anaplerotic reactions play a key role in replenishing the intermediates of the TCA cycle [Owen et al., 2002], allowing microbial

growth when there is only a certain carbohydrate available [Choi et al., 2016], indicating that sunlight is actually the only carbon source of the system.

7.6 Future perspectives: more holistic approaches are essential to uncover the underlying mechanisms governing microbial communities

Aim of this PhD was to enhance the analysis of microbiome data from a bioinformatics-point of view and to implement such analyses to investigate life in extreme cases. As already discussed, HTS methods come up with several challenges and approaches such as PEMA (see Chapter 2.1 [Zafeiropoulos et al., 2020]) and DARN (see Chapter 2.2 [Zafeiropoulos et al., 2021a]) can benefit the community in terms of fine parameter tuning and easy-to-use analysis and quality checks for non-target taxa correspondingly. Apparently, there is still a great number of issues need to be addressed. In case of amplicon studies, abundance estimation of the identified taxa has been crucial. Gloor et al. in their study in 2017 claimed that it is the nature of the microbiome data that does not allow them to be anything else but compositional. However, several studies have been implemented approaches to retrieve absolute abundances from amplicon data, e.g. [Kim et al., 2021, Zemb et al., 2020]. On top of that, long-read sequencing technologies have been improved radically, allowing quality sequencing of the complete marker and not just of a certain part of it. Therefore, approaches such as the Unique Molecular Identifiers (UMIs) [Hiatt et al., 2010] need to be considered and benchmarked. Yet, as discussed in [Karst et al., 2021] by using paired UMIs, certain issues that characterise amplicon studies such as chimeras and degree GC [Benita et al., 2003] ease; however, long-range PCR issues, lead to further challenges. Further improvement of the bioinformatics analysis of the data provided from both these technologies can provide more quantitave microbiome data; a milestone for microbial ecology.

As already discussed, data integration is essential in exploiting both the data and the so-far knowledge. Approaches such as PREGO (see Chapter 3 [Zafeiropoulos et al., 2022]) will not reveal their true potential unless the community embrace a series of standards and metadata protocols. Thus, it is essential both to develop such checklists but also to convince and train the community to use them. PREGO-like approaches could benefit a lot from multiomics datasets as well as by integrating information included in GEMs. We are now in a position when we can imagine (and why not, start implementing) of systems where a species will not be a node in a network, but a complete network within the one of the community it belongs to. Moreover, to infer microbe - microbe associations and/or interactions, the incorporation of phenotypical data such as pH values, optimal temperatures etc. to such networks, is essential. Resources such as FAPROTAX [Louca, Parfrey, and Doebeli, 2016], PhenDB [Feldbauer et al., 2015] and BugBase [Ward et al., 2017] provide great input for such a task.

Metagenomics and the rest of the 'omics technologies have turned the page on microbial ecology studies. However, to address questions such as the seasonality effects on the community structure and its functioning (see Section 5.5 and Conclusion 7.5) requires the thorough investigation of the dynamics of the community and of the occurying processes

within each of them as the environmental conditions change, as well as the interactions among the species and their environment. As discussed by Bajic and Sanchez [2020] *"a combination of quantitative high - throughput experiments and predictive metabolic models can help us map the genotype - phenotype map of microbial metabolic strategies.* Further technologies, such as Raman micro-spectroscopy [Jing et al., 2018], can also be of great use to this end. According to Bajic and Sanchez the prediction of such strategies will also provide great insight on the evolvability of metabolic decisions and will shed light on how these decisions affect microbial coexistence in the communities. Therefore, HTS approaches and metabolic modeling at the community level build a promising framework to this end. Sampling (see 4 [Chalkis et al., 2021]) the flux space of such models will benefit microbial interactions inference to a great extent and will provide essential insight in the study of the mechanisms that govern real-world microbial assemblages. As the steady-state assumption does not consider kinetics or regulatory events, the integration of machine learning approaches can enhance this framework even further [Sahu et al., 2021].

Acknowledgments

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

Appendices

Appendix A

PREGO

A.1 Mappings

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

A.2 Daemons

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

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

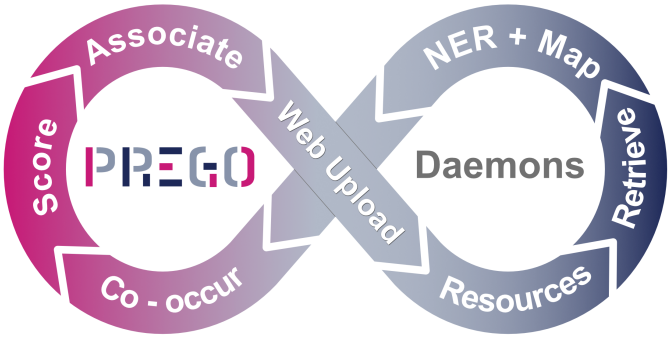


FIGURE A.1: Software daemons perform all steps of the PREGO methodology in a continuous manner similar to the Continuous Development and Continuous Integration method.

A.3 Scoring

Scoring in PREGO is used to answer the questions:

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

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

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

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

Literature Channel

Scoring in the Literature channel is implemented as in STRING 9.1 [Franceschini et al., 2012] and COMPARTMENTS [Binder et al., 2014], where the text mining method uses a three-step scoring scheme. First, for each co-mention/co-occurrence between entities (e.g., Methanosarcina mazei with Sulfur carrier activity), a weighted count is calculated because of the complexity of the text.

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

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

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

where $a = 0.6$ is a weighting factor, and the $c_{x,.}$, $c_{.,.}$, $c_{.,y}$ are the weighted counts as shown in Table A.1 estimated using the same Equation A.2. This value of the weighting factor has been chosen because it has been optimized and benchmarked in various applications of text mining [Franceschini et al., 2012, Binder et al., 2014, Pletscher-Frankild et al., 2015]. The value of Equation A.2 is sensitive to the increasing size of the number of documents (MEDLINE PubMed—PMC OA). Therefore, to obtain a more robust measure, the value of the score is transformed to z-score. This transformation is elaborated in detail in the COMPARTMENTS resource [Binder et al., 2014]. Finally, the confidence score is the z-score divided by two. Cases in which the scores exceed the (0,4] interval are capped to a maximum of 4 to reflect the uncertainty of the text mining pipeline.

Environmental Samples Channel

Data from environmental samples are OTU tables and metagenomes. Thus, for each entity x, the number of samples is calculated as the background and a number of samples of the associated entity (metadata background) $c_{.,y}$ (see Table A.1). Each association between entities x,y has a number of samples, $c_{x,y}$ that they co-occur. Note that each resource is independent and the scoring scheme is applied to its entities. This means that the same association can appear in multiple resources with different scores. The score is calculated with the following formula:

$$score_{x,y} = 2.0 * \frac{\sqrt{c_{x,y}}}{c_{.,y}^{0.1}} \tag{A.3}$$

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

A.4 Bulk download

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

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

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

Appendix B

Computational Geometry

B.1 Moving from the concentration to the flux vector

The dynamic mass balance on a chemical compound is the difference between the sum of the fluxes of all the reactions that form it and the sum of all the reactions that degrade it.

In general, the following ordinary differential equation expresses such a mass balance:

$$\frac{d\omega_i}{dt} = \sum_k s_{ik} v_k = \langle s_i \cdot v \rangle, \tag{B.1}$$

where ω_i is the total mass of the i-th metabolite, s_{ik} is the stoichiometric coefficient for this metabolite in the k-th reaction, and v_k is the flux of the k-th reaction. By considering all the differential equations expressing the dynamic mass balance of all the compounds present in a metabolic network, we have

$$\frac{d\omega}{dt} = Sv, \tag{B.2}$$

where S is the stoichiometric matrix having as rows the vectors s_i. In this setting, S is the map of the linear transformation that sends the flux vector to a vector of time derivatives of the concentration vector Palsson [2015].

B.2 Definitions & concepts

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

A *hyperplane* is a set of the form

$$H = \{x \in \mathbb{R}^n : p^T \cdot x = t\} \tag{B.3}$$

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

$$\begin{aligned} H_- &= \{x \in \mathbb{R}^n : p^T \cdot x \leq t\} \\ H_+ &= \{x \in \mathbb{R}^n : p^T \cdot x \geq t\} \end{aligned} \tag{B.4}$$

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

$$a_i^t \cdot x \leq b_i, i \in \{1,..,m\} \tag{B.5}$$

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

$$P = \{x \in \mathbb{R}^n : A \cdot x \leq b\} \tag{B.6}$$

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

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

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

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

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

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

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

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

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

$$P = \{x \in \mathbb{R}^n : a_i^T \cdot x \leq b_i, i = \{1,..m\}\} \tag{B.7}$$

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

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

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

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

Appendix C

Metagenome assembled genomes of novel prokaryotic taxa from a hypersaline marsh microbial mat

C.1 MAGs description

Using Barrnap [Seemann, 2014a] the 16S rRNA gene was extracted from the retrieved MAGs. For those where a 16S rRNA gene was found, Protologger (version 1.0) [Hitch et al., 2021] was used for a thorough description of their properties. On the Protologger framework, MAG's closest relatives based on 16S rRNA gene sequence similarity were retrieved using blastn (version 2.12.0+) and The All-Species Living Tree database [Ludwig et al., 2021]. Each MAG was placed on the GTDB phylogeny tree using the GTDB-Tk and average nucleotide identity (ANI) values were calculated to check whether the MAG is a representative of an already known species; no ANI value is reported for a genome pair if ANI value is much below 80%.

MAGs were then annotated with Prokka (Seemann 2014b) [Seemann, 2014b] and percentage of conserved proteins (POCP) values [Qin et al., 2014] was calculated between MAGs and the genomes that are close to it based on both the 16S rRNA and the genome-based assignment modules. This was the case only for genomes with validly published names according to the DSMZ nomenclature list. POCP analysis has been used to distinguish prokaryotic genera since a prokaryotic genus can be defined as a group of species with all pairwise POCP values higher than 50% [Qin et al., 2014]. The outcome of the Protologger tool for each archaeal MAG can be found on GitHub and for each bacterial too.

For MAG cases suggesting multiple novel entries in novel taxonomic groups higher than the species level, e.g. multiple novel genera within the same family, further POCP values were calculated between each of the MAGs and all of its closest relatives having a genome on GTDB using in-house scripts. For these cases, a phylogenetic tree using the MAG's alignment by the GTDB-Tk and the genomes' entries in the GTDB Multiple Sequence Alignment (MSA) was built. Both the phylogenetic trees and the POCP analyses for these cases are available here. In the "Etymology" section the names given to the new

taxa are described. For a thorough investigation of the so-far described characteristics of the reconstructed MAGs' higher taxonomic levels (e.g., genus, family etc.) the PREGO knowledge base [Zafeiropoulos et al., 2022] was exploited. All bioinformatics analyses were supported by the IMBBC High Performance Computing system [Zafeiropoulos et al., 2021c].

C.2 Results

The vast majority of the taxonomic assignments returned by the GTDB-Tk were at higher than the species level, suggesting that most of the MAGs are representatives of novel taxa; only 4 MAGs were assigned at the species level. For all non low quality MAGs for which the 16S rRNA gene was retrieved. Protologger was used to investigate further their uniqueness. Thus, protologues [Tindall, 1999] accompany 25 archaeal and 100 bacterial MAGs. For the cases where multiple MAGs were supposed to be representatives of a novel higher taxonomic group, e.g. several MAGs suggesting novel genus or genera within a certain family, the extra POCP analyses that were performed along with the corresponding phylogenetic tree are available on GitHub (archaeal and bacterial taxa) [1]. In total, our MAGs correspond to the novel taxa present in Table C.1.

Level of novel taxa	# of novel Archaea	# of novel Bacteria
species	5	11
genera	13	22
families	4	6
orders	2	2
phyla	-	1

TABLE C.1: Number of novel taxa described with a protologue based on the MAGs retrieved

GTDB R07-RS207 was recently released (2022, April); the reconstructed MAGs will be classified against this new version of GTDB before publishing to take into account genomes and taxa added in this updated version.

[1] Links to the protologues will be publically accessible once the MAGs will be published

Bibliography

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

IUPAC-IUBMB joint commission on biochemical nomenclature (JCBN) and nomenclature committee of IUBMB (NC-IUBMB), newsletter 1999. *Eur. J. Biochem.*, 264(2):607–609, Sept. 1999.

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

The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1): D325–D334, 2021.

R. M. Abed, K. Kohls, and D. De Beer. Effect of salinity changes on the bacterial diversity, photosynthesis and oxygen consumption of cyanobacterial mats from an intertidal flat of the arabian gulf. *Environmental Microbiology*, 9(6):1384–1392, 2007.

R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010. ISSN 0894-0347, 1088-6834. doi: 10.1090/S0894-0347-09-00650-X.

C. Akpolat, A. B. Fernández, P. Caglayan, B. Calli, M. Birbir, and A. Ventosa. Prokaryotic Communities in the Thalassohaline Tuz Lake, Deep Zone, and Kayacik, Kaldirim and Yavsan Salterns (Turkey) Assessed by 16S rRNA Amplicon Sequencing. *Microorganisms*, 9(7):1525, July 2021. ISSN 2076-2607. doi: 10.3390/microorganisms9071525. URL https://www.mdpi.com/2076-2607/9/7/1525. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

R. Al-Thani, M. A. A. Al-Najjar, A. M. Al-Raei, T. Ferdelman, N. M. Thang, I. A. Shaikh, M. Al-Ansi, and D. d. Beer. Community Structure and Activity of a Highly Dynamic and Nutrient-Limited Hypersaline Microbial Mat in Um Alhool Sabkha, Qatar. *PLOS ONE*, 9 (3):e92405, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0092405. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0092405. Publisher: Public Library of Science.

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.

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. URL https://doi.org/10.5281/zenodo.4670966.

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*, 22(1):1–24, 2021.

T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, and H. Ogata. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, Apr. 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz859. URL https://doi.org/10.1093/bioinformatics/btz859.

K. R. Arrigo. Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349–355, 2005.

S. Artstein-Avidan, H. Kaplan, and M. Sharir. On radial isotropic position: Theory and algorithms, 2020.

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? Technical report, PeerJ Preprints, 2018.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

F. Asnicar, G. Weingart, T. L. Tickle, C. Huttenhower, and N. Segata. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, 3:e1029, June 2015. ISSN 2167-8359. doi: 10.7717/peerj.1029. URL https://peerj.com/articles/1029. Publisher: PeerJ Inc.

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*, 8(4):giz029, 2019.

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

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*, 3(10):e3376, 2008.

E. Bairey, E. D. Kelsic, and R. Kishony. High-order species interactions shape ecosystem diversity. *Nature communications*, 7(1):1–7, 2016.

D. Bajic and A. Sanchez. The ecology and evolution of microbial metabolic strategies. *Current opinion in biotechnology*, 62:123–128, 2020.

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

M. Bálint, M. Bahram, A. M. Eren, K. Faust, J. A. Fuhrman, B. Lindahl, R. B. O'Hara, M. Öpik, M. L. Sogin, M. Unterseher, et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS microbiology reviews*, 40(5):686–700, 2016.

F. A. Baltoumas, S. Zafeiropoulou, E. Karatzas, M. Koutrouli, F. Thanati, K. Voutsadaki, M. Gkonta, J. Hotova, I. Kasionis, P. Hatzis, et al. Biomolecule and bioentity interaction databases in systems biology: A comprehensive review. *Biomolecules*, 11(8):1245, 2021a.

F. A. Baltoumas, S. Zafeiropoulou, E. Karatzas, S. Paragkamian, F. Thanati, I. Iliopoulos, A. G. Eliopoulos, R. Schneider, L. J. Jensen, E. Pafilis, et al. Onthefly2. 0: a text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis. *bioRxiv*, 2021b.

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*, 19(5): 455–477, 2012.

P. Bansal, A. Morgat, K. B. Axelsen, V. Muthukrishnan, E. Coudert, L. Aimo, N. Hyka-Nouspikel, E. Gasteiger, A. Kerhornou, T. B. Neto, et al. Rhea, the reaction knowledgebase in 2022. *Nucleic acids research*, 50(D1):D693–D700, 2022.

Y. M. Bar-On, R. Phillips, and R. Milo. The biomass distribution on earth. *Proceedings of the National Academy of Sciences*, 115(25):6506–6511, 2018.

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*, 68 (2):365–369, 2019.

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*, 21(1):129–145, 2020.

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*, 472(7344):429–436, 2011.

G. A. Begg and J. R. Waldman. An holistic approach to fish stock identification. *Fisheries research*, 43(1-3):35–44, 1999.

J. Belilla, D. Moreira, L. Jardillier, G. Reboul, K. Benzerara, J. M. López-García, P. Bertolino, A. I. López-Archilla, and P. López-García. Hyperdiverse archaea near life limits at the polyextreme geothermal dallol area. *Nature ecology & evolution*, 3(11):1552–1561, 2019.

K. L. Bell, R. A. Petit III, A. Cutler, E. K. Dobbs, J. M. Macpherson, T. D. Read, K. S. Burgess, and B. J. Brosi. Comparing whole-genome shotgun sequencing and dna metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and evolution*, 11(22):16082–16098, 2021a.

K. L. Bell, R. A. Petit III, A. Cutler, E. K. Dobbs, J. M. Macpherson, T. D. Read, K. S. Burgess, and B. J. Brosi. Comparing whole-genome shotgun sequencing and DNA metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and Evolution*, 11(22):16082–16098, 2021b. ISSN 2045-7758. doi: 10.1002/ece3.8281. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.8281. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.8281.

Y. Benita, R. S. Oosting, M. C. Lok, M. J. Wise, and I. Humphery-Smith. Regionalized gc content of template dna as a predictor of pcr success. *Nucleic acids research*, 31(16): e99–e99, 2003.

D. Bensasson, D.-X. Zhang, D. L. Hartl, and G. M. Hewitt. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in ecology & evolution*, 16(6):314–321, 2001.

D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, and E. W. Sayers. Genbank. *Nucleic acids research*, 46(D1):D41–D47, 2018.

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

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*, 10(3):707–715, 2018.

D. B. Bernstein, F. E. Dewhirst, and D. Segre. Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome. *Elife*, 8:e39733, 2019.

D. Bertsimas and S. Vempala. Solving convex programs by random walks. *J. ACM*, 51 (4):540–556, July 2004. ISSN 0004-5411. doi: 10.1145/1008731.1008733. URL https://doi.org/10.1145/1008731.1008733.

M. B. Biggs, G. L. Medlock, G. L. Kolling, and J. A. Papin. Metabolic network modeling of microbial communities. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(5):317–334, 2015.

J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, and L. J. Jensen. Compartments: unification and visualization of protein subcellular localization evidence. *Database*, 2014, 2014.

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.

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*, 18(5):1020–1034, 2018.

S. J. Blazewicz, R. L. Barnard, R. A. Daly, and M. K. Firestone. Evaluating rrna as an indicator of microbial activity in environmental communities: limitations and uses. *The ISME journal*, 7(11):2061–2068, 2013.

R. A. Bobadilla Fazzini, M. P. Cortés, L. Padilla, D. Maturana, M. Budinich, A. Maass, and P. Parada. Stoichiometric modeling of oxidation of reduced inorganic sulfur compounds (riscs) in acidithiobacillus thiooxidans. *Biotechnology and Bioengineering*, 110(8):2242–2251, 2013.

F. Boero and E. Bonsdorff. A conceptual framework for marine biodiversity and ecosystem functioning. *Marine Ecology*, 28:134–145, 2007.

N. A. Bokulich, M. Ziemski, M. S. Robeson II, and B. D. Kaehler. Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. *Computational and Structural Biotechnology Journal*, 18:4048–4062, 2020.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

H. Bolhuis, M. S. Cretoiu, and L. J. Stal. Molecular ecology of microbial mats. *FEMS Microbiology Ecology*, 90(2):335–350, Nov. 2014. ISSN 0168-6496. doi: 10.1111/1574-6941.12408. URL https://doi.org/10.1111/1574-6941.12408.

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. Technical report, PeerJ Preprints, 2018.

S. Bordel, R. Agren, and J. Nielsen. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLOS Computational Biology*, 6(7):1–13, 07 2010. doi: 10.1371/journal.pcbi.1000859. URL https://doi.org/10.1371/journal.pcbi.1000859.

R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. Murat Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.

Nature Biotechnology, 35(8):725–731, Aug. 2017. ISSN 1546-1696. doi: 10.1038/nbt.3893. URL https://www.nature.com/articles/nbt.3893. Number: 8 Publisher: Nature Publishing Group.

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*, 16(1): 176–182, 2016.

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*, 82(19): 5878–5891, 2016. doi: 10.1128/AEM.01630-16.

R. M. Braga, M. N. Dourado, and W. L. Araújo. Microbial interactions: ecology in a molecular perspective. *Brazilian Journal of Microbiology*, 47:86–98, 2016.

F. P. Breitwieser, J. Lu, and S. L. Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, 20(4):1125–1136, 2019.

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*, 36(3):272, 2018.

D. A. Bryant and N.-U. Frigaard. Prokaryotic photosynthesis and phototrophy illuminated. *Trends in Microbiology*, 14(11):488–496, Nov. 2006. ISSN 0966-842X. doi: 10.1016/j.tim.2006.09.001. URL https://www.sciencedirect.com/science/article/pii/S0966842X06002265.

J. P. Buchmann and E. C. Holmes. Collecting and managing taxonomic data with ncbitaxonomist. *Bioinformatics*, 36(22-23):5548–5550, 2020.

J. G. Bundy, M. P. Davey, and M. R. Viant. Environmental metabolomics: a critical review and future perspectives. *Metabolomics*, 5(1):3–21, 2009.

P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, and C. J. Mungall. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of biomedical semantics*, 7(1):1–12, 2016.

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*, 12 (5):834–845, 2012.

A. Cakmak, X. Qi, A. E. Cicek, I. Bederman, L. Henderson, M. Drumm, and G. Ozsoyoglu. A new metabolomics analysis technique: steady-state metabolic network dynamics analysis. *Journal of bioinformatics and computational biology*, 10(01):1240003, 2012.

L. Calès, A. Chalkis, I. Z. Emiris, and V. Fisikopoulos. Practical Volume Computation of Structured Convex Bodies, and an Application to Modeling Portfolio Dependencies and Financial Crises. In B. Speckmann and C. D. Tóth, editors, *34th International Symposium on Computational Geometry (SoCG 2018)*, volume 99 of *LIPIcs*, pages 19:1–19:15, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi: 10.4230/LIPIcs.SoCG.2018.19.

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*, 13(7):581–583, 2016.

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*, 11(12): 2639–2643, 2017.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):1–9, 2009.

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

Y. Cao, Y. Wang, X. Zheng, F. Li, and X. Bo. Revecor: an r package for the reverse ecology analysis of microbiomes. *BMC bioinformatics*, 17(1):1–6, 2016.

D. C. Cardoso, M. S. Cretoiu, L. J. Stal, and H. Bolhuis. Seasonal development of a coastal microbial mat. *Scientific Reports*, 9(1):9035, June 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-45490-8. URL https://www.nature.com/articles/s41598-019-45490-8. Number: 1 Publisher: Nature Publishing Group.

G. Carvalho and L. Hauser. Molecular genetics and the stock concept in fisheries. In *Molecular genetics in fisheries*, pages 55–79. Springer, 1995.

A. Casanueva, N. Galada, G. C. Baker, W. D. Grant, S. Heaphy, B. Jones, M. Yanhe, A. Ventosa, J. Blamey, and D. A. Cowan. Nanoarchaeal 16S rRNA gene sequences are widely dispersed in hyperthermophilic and mesophilic halophilic environments. *Extremophiles*, 12(5):651–656, Sept. 2008. ISSN 1433-4909. doi: 10.1007/s00792-008-0170-x. URL https://doi.org/10.1007/s00792-008-0170-x.

R. Caspi, R. Billington, I. M. Keseler, A. Kothari, M. Krummenacker, P. E. Midford, W. K. Ong, S. Paley, P. Subhraveti, and P. D. Karp. The metacyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic acids research*, 48(D1):D445–D453, 2020.

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*, 21(10):1–17, 2020.

J. Catchen, P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. Stacks: an analysis tool set for population genomics. *Molecular ecology*, 22(11):3124–3140, 2013.

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*, 17(9):569–586, 2019.

A. Chalkis and V. Fisikopoulos. volesti: Volume approximation and sampling for convex polytopes in R, 2020. https://github.com/GeomScale/volume_approximation.

A. Chalkis, I. Z. Emiris, and V. Fisikopoulos. Practical volume estimation of zonotopes by a new annealing schedule for cooling convex bodies. In A. M. Bigatti, J. Carette, J. H. Davenport, M. Joswig, and T. de Wolff, editors, *Mathematical Software – ICMS 2020*, pages 212–221, Cham, 2020. Springer International Publishing. ISBN 978-3-030-52200-1.

A. Chalkis, V. Fisikopoulos, E. Tsigaridas, and H. Zafeiropoulos. Geometric Algorithms for Sampling the Flux Space of Metabolic Networks. In K. Buchin and E. Colin de Verdière, editors, *37th International Symposium on Computational Geometry (SoCG 2021)*, volume 189 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 21:1–21:16, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-184-9. doi: 10.4230/LIPIcs.SoCG.2021.21. URL https://drops.dagstuhl.de/opus/volltexte/2021/13820.

C. S. Chan, K.-G. Chan, Y.-L. Tay, Y.-H. Chua, and K. M. Goh. Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. *Frontiers in Microbiology*, 6, 2015. ISSN 1664-302X. URL https://www.frontiersin.org/article/10.3389/fmicb.2015.00177.

S. P. Chapman, C. M. Paget, G. N. Johnson, and J.-M. Schwartz. Flux balance analysis reveals acetate metabolism modulates cyclic electron flow and alternative glycolytic pathways in chlamydomonas reinhardtii. *Frontiers in plant science*, 6:474, 2015.

P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 36(6):1925–1927, Mar. 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz848. URL https://doi.org/10.1093/bioinformatics/btz848.

I.-M. A. Chen, K. Chu, K. Palaniappan, A. Ratner, J. Huang, M. Huntemann, P. Hajek, S. Ritter, N. Varghese, R. Seshadri, et al. The img/m data management and analysis system v. 6.0: new tools and advanced capabilities. *Nucleic acids research*, 49(D1): D751–D763, 2021.

Q.-L. Chen, J. Ding, D. Zhu, H.-W. Hu, M. Delgado-Baquerizo, Y.-B. Ma, J.-Z. He, and Y.-G. Zhu. Rare microbial taxa as the major drivers of ecosystem multifunctionality in long-term fertilized soils. *Soil Biology and Biochemistry*, 141:107686, 2020a.

R. Chen, H. L. Wong, G. S. Kindler, F. I. MacLeod, N. Benaud, B. C. Ferrari, and B. P. Burns. Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats. *Frontiers in Microbiology*, 11, 2020b. ISSN 1664-302X. URL https://www.frontiersin.org/article/10.3389/fmicb.2020.01950.

Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mcmc sampling algorithms on polytopes. *Journal of Machine Learning Research*, 19(55):1–86, 2018. URL http://jmlr.org/papers/v19/18-158.html.

J. Cheney, S. Chong, N. Foster, M. Seltzer, and S. Vansummeren. Provenance: a future history. In *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*, pages 957–964, 2009.

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. URL https://hal.archives-ouvertes.fr/hal-01919855.

P. H. Choi, J. Jo, Y.-C. Lin, M.-H. Lin, C.-Y. Chou, L. E. P. Dietrich, and L. Tong. A distinct holoenzyme organization for two-subunit pyruvate carboxylase. *Nature Communications*, 7(1):12713, Oct. 2016. ISSN 2041-1723. doi: 10.1038/ncomms12713. URL https://www.nature.com/articles/ncomms12713. Number: 1 Publisher: Nature Publishing Group.

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*, 19(1):27–46, 2019.

S. Cinar and M. B. Mutlu. Prokaryotic Community Compositions of the Hypersaline Sediments of Tuz Lake Demonstrated by Cloning and High-Throughput Sequencing. *Microbiology*, 89(6):756–768, Nov. 2020. ISSN 1608-3237. doi: 10.1134/S0026261720060028. URL https://doi.org/10.1134/S0026261720060028.

P. Cingolani, R. Sladek, and M. Blanchette. Bigdatascript: a scripting language for data pipelines. *Bioinformatics*, 31(1):10–16, 2015.

A. G. Clooney, F. Fouhy, R. D. Sleator, A. O. Driscoll, C. Stanton, P. D. Cotter, and M. J. Claesson. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLOS ONE*, 11(2):e0148028, Feb. 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0148028. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0148028. Publisher: Public Library of Science.

E. Coissac, T. Riaz, and N. Puillandre. Bioinformatic challenges for dna metabarcoding of plants and animals. *Molecular ecology*, 21(8):1834–1847, 2012.

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*, 10(11):1985–2001, 2019.

N. Conde-Pueyo, B. Vidiella, J. Sardanyés, M. Berdugo, F. T. Maestre, V. De Lorenzo, and R. Solé. Synthetic biology for terraformation lessons from mars, earth, and the microbiome. *Life*, 10(2):14, 2020.

B. Cousins. *Efficient high-dimensional sampling and integration*. PhD thesis, Georgia Institute of Technology, Georgia, U.S.A., 2017.

B. Cousins and S. Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016.

C. Cummins, A. Ahamed, R. Aslam, J. Burgin, R. Devraj, O. Edbali, D. Gupta, P. W. Harrison, M. Haseeb, S. Holt, T. Ibrahim, E. Ivanov, S. Jayathilaka, V. Kadhirvelu, S. Kay, M. Kumar, A. Lathi, R. Leinonen, F. Madeira, N. Madhusoodanan, M. Mansurova, C. O'Cathail, M. Pearce, S. Pesant, N. Rahman, J. Rajan, G. Rinck, S. Selvakumar, A. Sokolov, S. Suman, R. Thorne, P. Totoo, S. Vijayaraja, Z. Waheed, A. Zyoud, R. Lopez, T. Burdett, and G. Cochrane. The European Nucleotide Archive in 2021. *Nucleic Acids Research*, 50(D1):D106–D110, Jan. 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1051. URL https://doi.org/10.1093/nar/gkab1051.

L. Czech, P. Barbera, and A. Stamatakis. Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics*, 35(7):1151–1158, 2019.

L. Czech, P. Barbera, and A. Stamatakis. Genesis and gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics*, 36(10):3263–3265, 2020.

L. Dagum and R. Menon. Openmp: an industry standard api for shared-memory programming. *IEEE computational science and engineering*, 5(1):46–55, 1998.

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*, 7 (5):giy028, 2018.

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.

S. Dávila-Ramos, H. G. Castelán-Sánchez, L. Martínez-Ávila, M. d. R. Sánchez-Carbente, R. Peralta, A. Hernández-Mendoza, A. D. Dobson, R. A. Gonzalez, N. Pastor, and R. A. Batista-García. A review on viral metagenomics in extreme environments. *Frontiers in microbiology*, page 2403, 2019.

J. de la Cuesta-Zuluaga, R. E. Ley, and N. D. Youngblut. Struo: a pipeline for building custom databases for common metagenome profilers. *Bioinformatics*, 36(7):2314–2315, 2020.

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*, 72(4):698–705, 2020.

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*, 10 (9):20140562, 2014.

E. Deelman, B. Berriman, A. Chervenak, O. Corcho, P. Groth, and L. Moreau. Metadata and provenance management. *arXiv preprint arXiv:1005.2643*, 2010.

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*, 26(21):5872–5895, 2017.

J. Del Campo, M. Kolisko, V. Boscaro, L. F. Santoferrara, S. Nenarokov, R. Massana, L. Guillou, A. Simpson, C. Berney, C. de Vargas, et al. Eukref: phylogenetic curation of ribosomal rna to enhance understanding of eukaryotic diversity and distribution. *PLoS biology*, 16(9):e2005849, 2018.

M. Delgado-Baquerizo, F. T. Maestre, P. B. Reich, T. C. Jeffries, J. J. Gaitan, D. Encinar, M. Berdugo, C. D. Campbell, and B. K. Singh. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nature communications*, 7(1):1–8, 2016.

T. O. Delmont, C. Malandain, E. Prestat, C. Larose, J.-M. Monier, P. Simonet, and T. M. Vogel. Metagenomic mining for microbiologists. *The ISME Journal*, 5(12):1837–1843, 2011.

A. B. Dieker and S. S. Vempala. Stochastic billiards for sampling from the boundary of a convex set. *Mathematics of Operations Research*, 40(4):888–901, 2015. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/24540983.

C. Diener, S. M. Gibbons, and O. Resendis-Antonio. Micom: metagenome-scale modeling to infer metabolic interactions in the gut microbiota. *MSystems*, 5(1):e00606–19, 2020.

J. G. Dillon, S. Miller, B. Bebout, M. Hullar, N. Pinel, and D. A. Stahl. Spatial and temporal variability in a stratified hypersaline microbial mat community. *FEMS Microbiology Ecology*, 68(1):46–58, Apr. 2009. ISSN 0168-6496. doi: 10.1111/j.1574-6941.2009.00647.x. URL https://doi.org/10.1111/j.1574-6941.2009.00647.x.

Z. A. DiLoreto, T. R. R. Bontognali, Z. A. Al Disi, H. A. S. Al-Kuwari, K. H. Williford, C. J. Strohmenger, F. Sadooni, C. Palermo, J. M. Rivers, J. A. McKenzie, M. Tuite, and M. Dittrich. Microbial community composition and dolomite formation in the hypersaline microbial mats of the Khor Al-Adaid sabkhas, Qatar. *Extremophiles*, 23 (2):201–218, Mar. 2019. ISSN 1433-4909. doi: 10.1007/s00792-018-01074-4. URL https://doi.org/10.1007/s00792-018-01074-4.

H. M. Dionisi, M. Lozada, and N. L. Olivera. Bioprospection of marine microorganisms: biotechnological applications and methods. *Revista argentina de microbiología*, 44(1): 49–60, 2012.

S. M. Dittami and E. Corre. Detection of bacterial contaminants and hybrid sequences in the genome of the kelp saccharina japonica using taxoblast. *PeerJ*, 5:e4073, 2017.

J. J. Dongarra, P. Luszczek, and A. Petitet. The linpack benchmark: past, present and future. *Concurrency and Computation: practice and experience*, 15(9):803–820, 2003.

I. Dukovski, D. Bajić, J. M. Chacón, M. Quintin, J. C. Vila, S. Sulheim, A. R. Pacheco, D. B. Bernstein, W. J. Riehl, K. S. Korolev, et al. A metabolic modeling platform for the computation of microbial ecosystems in time and space (comets). *Nature protocols*, 16 (11):5030–5082, 2021.

P. V. Dunlap. Microbial diversity. 2001.

S. Dávila-Ramos, H. G. Castelán-Sánchez, L. Martínez-Ávila, M. d. R. Sánchez-Carbente, R. Peralta, A. Hernández-Mendoza, A. D. W. Dobson, R. A. Gonzalez, N. Pastor, and R. A. Batista-García. A Review on Viral Metagenomics in Extreme Environments. *Frontiers in Microbiology*, 10, 2019. ISSN 1664-302X. URL https://www.frontiersin.org/article/10.3389/fmicb.2019.02403.

K. D'Hondt, T. Kostic, R. McDowell, F. Eudes, B. K. Singh, S. Sarkar, M. Markakis, B. Schelkle, E. Maguin, and A. Sessitsch. Microbiome innovations for a sustainable future. *Nature Microbiology*, 6(2):138–142, 2021.

E. M. Eckert, A. Di Cesare, D. Fontaneto, T. U. Berendonk, H. Bürgmann, E. Cytryn, D. Fatta-Kassinos, A. Franzetti, D. J. Larsson, C. M. Manaia, et al. Every fifth published metagenome is not available to science. *PLoS biology*, 18(4):e3000698, 2020.

R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, Mar. 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh340. URL https://doi.org/10.1093/nar/gkh340.

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

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*, 1817(6):898–910, 2012.

V. Elbrecht and F. Leese. Validation and development of coi metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers in Environmental Science*, 5:11, 2017.

A. Eng and E. Borenstein. Taxa-function robustness in microbial communities. *Microbiome*, 6(1):1–19, 2018.

P. L. B. EnvironmentOntology. Using envo with mixs · environmentontology/envo wiki, Apr 2021. URL https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS.

A. M. Eren, O. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, Oct. 2015. ISSN 2167-8359. doi: 10.7717/peerj.1319. URL https://peerj.com/articles/1319. Publisher: PeerJ Inc.

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*, 41:83–116, 2016.

S. Estrela, J. C. Vila, N. Lu, D. Bajić, M. Rebolleda-Gómez, C.-Y. Chang, J. E. Goldford, A. Sanchez-Gorostiaga, and Á. Sánchez. Functional attractors in microbial community assembly. *Cell Systems*, 13(1):29–42, 2022.

P. G. Falkowski, T. Fenchel, and E. F. Delong. The microbial engines that drive earth's biogeochemical cycles. *science*, 320(5879):1034–1039, 2008.

S. Fallahi, H. J. Skaug, and G. Alendal. A comparison of Monte Carlo sampling methods for metabolic network models. *PLOS One*, 15(7):e0235393, 2020.

K. Faust and J. Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012.

A. M. Feist and B. O. Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–349, 2010.

A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Ø. Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2009.

R. Feldbauer, F. Schulz, M. Horn, and T. Rattei. Prediction of microbial phenotypes based on comparative genomics. *BMC bioinformatics*, 16(14):1–8, 2015.

C. Ferguson, D. Araújo, L. Faulk, Y. Gou, A. Hamelers, Z. Huang, M. Ide-Smith, M. Levchenko, N. Marinos, R. Nambiar, et al. Europe pmc in 2020. *Nucleic acids research*, 49(D1):D1507–D1514, 2021.

V. G. Fonseca. Pitfalls in relative abundance estimation using edna metabarcoding, 2018.

A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1): D808–D815, 2012.

S. Freilich, A. Kreimer, I. Meilijson, U. Gophna, R. Sharan, and E. Ruppin. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic acids research*, 38(12):3857–3868, 2010.

J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21, 2008.

B. M. Fuchs, S. Spring, H. Teeling, C. Quast, J. Wulf, M. Schattenhofer, S. Yan, S. Ferriera, J. Johnson, F. O. Glöckner, and R. Amann. Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. *Proceedings of the National Academy of Sciences*, 104(8):2891–2896, Feb. 2007. doi: 10.1073/pnas.0608046104.

URL https://www.pnas.org/doi/abs/10.1073/pnas.0608046104. Publisher: Proceedings of the National Academy of Sciences.

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

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. ISSN 0883-4237. URL https://www.jstor.org/stable/2246093. Publisher: Institute of Mathematical Statistics.

C. J. Geyer. Practical Markov chain Monte Carlo. *Statist. Sci.*, 7(4):473–483, 11 1992. doi: 10.1214/ss/1177011137. URL https://doi.org/10.1214/ss/1177011137.

J. S. Ghurye, V. Cepeda-Espinoza, and M. Pop. Focus: microbiome: metagenomic assembly: overview, challenges and applications. *The Yale journal of biology and medicine*, 89 (3):353, 2016.

J. A. Gilbert, J. K. Jansson, and R. Knight. The earth microbiome project: successes and aspirations. *BMC biology*, 12(1):1–4, 2014.

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

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*, 9(18):e00311–20, 2020.

S. Giri, S. Shitut, and C. Kost. Harnessing ecological and evolutionary principles to guide the design of microbial production consortia. *Current Opinion in Biotechnology*, 62: 228–238, 2020.

S. Giri, L. Oña, S. Waschina, S. Shitut, G. Yousif, C. Kaleta, and C. Kost. Metabolic dissimilarity determines the establishment of cross-feeding interactions in bacteria. *Current Biology*, 31(24):5547–5557, 2021.

S. I. Glassman and J. B. Martiny. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *MSphere*, 3(4):e00148–18, 2018.

G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

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*, 34(9): 942–949, 2016.

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*, 361(6401):469–474, 2018.

D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8(2):1–10, 2014.

D. Gonze, K. Z. Coyte, L. Lahti, and K. Faust. Microbial communities as dynamical systems. *Current opinion in microbiology*, 44:41–49, 2018.

S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

W. D. Grant. Halobacterium. In *Bergey's Manual of Systematics of Archaea and Bacteria*, pages 1–11. John Wiley & Sons, Ltd, 2015. ISBN 978-1-118-96060-8. doi: 10.1002/9781118960608.gbm00482. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118960608.gbm00482. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118960608.gbm00482.

W. Greece. Inventory report: KAR001 - Tristomo marsh, 2022. URL http://www.oikoskopio.gr/ygrotopio/general/report.php?id=171&param=themeleiwdn&wetland_lang=en_US.

C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng. Big data bioinformatics. *Journal of cellular physiology*, 229(12):1896–1900, 2014.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

E. Gryazina and B. Polyak. Random sampling: Billiard walk algorithm. *European Journal of Operational Research*, 238(2):497 – 504, 2014. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2014.03.041.

S. Gudmundsson and I. Thiele. Computationally efficient flux variability analysis. *BMC bioinformatics*, 11(1):1–3, 2010.

G. Guennebaud, B. Jacob, et al. *Eigen v3*, 2010. URL http://eigen.tuxfamily.org.

L. Guillou, D. Bachar, S. Audic, D. Bass, C. Berney, L. Bittner, C. Boutte, G. Burgaud, C. De Vargas, J. Decelle, et al. The protist ribosomal reference database (pr2): a catalog of unicellular eukaryote small sub-unit rrna sequences with curated taxonomy. *Nucleic acids research*, 41(D1):D597–D604, 2012.

N. Gunde-Cimerman, A. Plemenitas, and A. Oren. Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS Microbiology Reviews*, 42(3):353–375, May 2018. ISSN 0168-6445. doi: 10.1093/femsre/fuy009. URL https://doi.org/10.1093/femsre/fuy009.

A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, Apr. 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt086. URL https://doi.org/10.1093/bioinformatics/btt086.

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*, 6 (8):973–980, 2015.

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*, (5), 2017.

E. K. Hall, E. S. Bernhardt, R. L. Bier, M. A. Bradford, C. M. Boot, J. B. Cotner, P. A. Del Giorgio, S. E. Evans, E. B. Graham, S. E. Jones, et al. Understanding how microbiomes influence the systems they inhabit. *Nature Microbiology*, 3(9):977–982, 2018.

S. Hanada, A. Hiraishi, K. Shimada, and K. . Matsuura. Chloroflexus aggregans sp. nov., a Filamentous Phototrophic Bacterium Which Forms Dense Cell Aggregates by Active Gliding Movement. *International Journal of Systematic and Evolutionary Microbiology*, 45(4):676–681, 1995. ISSN 1466-5034,. doi: 10.1099/00207713-45-4-676. URL https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-45-4-676. Publisher: Microbiology Society,.

X. Hao, R. Jiang, and T. Chen. Clustering 16s rrna for otu prediction: a method of unsupervised bayesian clustering. *Bioinformatics*, 27(5):611–618, 2011.

H. S. Haraldsdóttir, B. Cousins, I. Thiele, R. M. Fleming, and S. Vempala. CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743, 01 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx052.

N. Harmston, W. Filsell, and M. P. Stumpf. What the papers say: Text mining for genomics and systems biology. *Human genomics*, 5(1):1–13, 2010.

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*, 47(D1):D84–D88, 2019.

P. W. Harrison, A. Ahamed, R. Aslam, B. T. Alako, J. Burgin, N. Buso, M. Courtot, J. Fan, D. Gupta, M. Haseeb, et al. The european nucleotide archive in 2020. *Nucleic acids research*, 49(D1):D82–D85, 2021.

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*, 270(suppl_1):S96–S99, 2003.

J. I. Hedges and J. H. Stern. Carbon and nitrogen determinations of carbonate-containing solids1. *Limnology and Oceanography*, 29(3):657–663, 1984. ISSN 1939-5590. doi: 10.4319/lo.1984.29.3.0657. URL https://onlinelibrary.wiley.com/doi/abs/10.4319/lo.1984.29.3.0657. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.4319/lo.1984.29.3.0657.

B. P. Hedlund, J. A. Dodsworth, S. K. Murugapiran, C. Rinke, and T. Woyke. Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles*, 18(5):865–875, Sept. 2014. ISSN 1433-4909. doi: 10.1007/s00792-014-0664-7. URL https://doi.org/10.1007/s00792-014-0664-7.

A. Heinken, A. Basile, and I. Thiele. Advances in constraint-based modelling of microbial communities. *Current Opinion in Systems Biology*, 27:100346, 2021.

M. Heinonen, M. Osmala, H. Mannerström, J. Wallenius, S. Kaski, J. Rousu, and H. Lähdesmäki. Bayesian metabolic flux analysis reveals intracellular flux couplings. *Bioinformatics*, 35(14):i548–i557, 2019.

L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdóttir, J. Wachowiak, S. M. Keating, V. Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019.

T. Henckel, M. Friedrich, and R. Conrad. Molecular Analyses of the Methane-Oxidizing Microbial Community in Rice Field Soil by Targeting the Genes of the 16S rRNA, Particulate Methane Monooxygenase, and Methanol Dehydrogenase. *Applied and Environmental Microbiology*, 65(5):1980–1990, May 1999. doi: 10.1128/AEM.65.5.1980-1990.1999. URL https://journals.asm.org/doi/full/10.1128/AEM.65.5.1980-1990.1999. Publisher: American Society for Microbiology.

H. A. Herrmann, B. C. Dyson, L. Vass, G. N. Johnson, and J.-M. Schwartz. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *NPJ systems biology and applications*, 5(1):1–8, 2019.

J. B. Hiatt, R. P. Patwardhan, E. H. Turner, C. Lee, and J. Shendure. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature methods*, 7(2):119–122, 2010.

F. Hildebrand, R. Tadeo, A. Y. Voigt, P. Bork, and J. Raes. Lotus: an efficient and user-friendly otu processing pipeline. *Microbiome*, 2(1):1–7, 2014.

T. C. A. Hitch, T. Riedel, A. Oren, J. Overmann, T. D. Lawley, and T. Clavel. Automated analysis of genomic sequences facilitates high-throughput and comprehensive description of bacteria. *ISME Communications*, 1(1):1–16, May 2021. ISSN 2730-6151. doi: 10.1038/s43705-021-00017-z. URL https://www.nature.com/articles/s43705-021-00017-z. Number: 1 Publisher: Nature Publishing Group.

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*, 35(2):518–522, 2018a.

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*, 35(2):518–522, Feb. 2018b. ISSN 0737-4038. doi: 10.1093/molbev/msx281. URL https://doi.org/10.1093/molbev/msx281.

S. Hoeft McCann, A. Boren, J. Hernandez-Maldonado, B. Stoneburner, C. W. Saltikov, J. F. Stolz, and R. S. Oremland. Arsenite as an Electron Donor for Anoxygenic Photosynthesis: Description of Three Strains of Ectothiorhodospira from Mono Lake, California and Big Soda Lake, Nevada. *Life*, 7(1):1, Mar. 2017. ISSN 2075-1729. doi: 10.3390/life7010001. URL https://www.mdpi.com/2075-1729/7/1/1. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

T. M. Hoehler, B. M. Bebout, and D. J. Des Marais. The role of microbial mats in the production of reduced gases on the early Earth. *Nature*, 412(6844):324–327, July 2001. ISSN 1476-4687. doi: 10.1038/35085554. URL https://www.nature.com/articles/35085554. Number: 6844 Publisher: Nature Publishing Group.

B. Hu, S. Canon, E. A. Eloe-Fadrosh, M. Babinski, Y. Corilo, K. Davenport, W. D. Duncan, K. Fagnan, M. Flynn, B. Foster, et al. Challenges in bioinformatics workflows for processing microbiome omics data at scale. *Frontiers in Bioinformatics*, 1, 2022.

H. Huber, M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, 417(6884): 63–67, May 2002. ISSN 1476-4687. doi: 10.1038/417063a. URL https://www.nature.com/articles/417063a. Number: 6884 Publisher: Nature Publishing Group.

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*, 56(6):553–557, 2008.

D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, Mar. 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119. URL https://doi.org/10.1186/1471-2105-11-119.

T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2(1):343–372, 2001.

J. F. Jadebeck, A. Theorell, S. Leweke, and K. Noh. Hops: high-performance library for non uniform sampling of convex constrained models. *Bioinformatics*, 2020.

C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):5114, Nov. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07641-9. URL https://www.nature.com/articles/s41467-018-07641-9. Number: 1 Publisher: Nature Publishing Group.

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*, 20(2):429–443, 2020.

J. K. Jarett, S. Nayfach, M. Podar, W. Inskeep, N. N. Ivanova, J. Munson-McGee, F. Schulz, M. Young, Z. J. Jay, J. P. Beam, N. C. Kyrpides, R. R. Malmstrom, R. Stepanauskas, and T. Woyke. Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. *Microbiome*, 6(1):161, Sept. 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0539-8. URL https://doi.org/10.1186/s40168-018-0539-8.

T. C. Jeffries, J. R. Seymour, K. Newton, R. J. Smith, L. Seuront, and J. G. Mitchell. Increases in the abundance of microbial genes encoding halotolerance and photosynthesis along a sediment salinity gradient. *Biogeosciences*, 9(2):815–825, Feb. 2012. ISSN 1726-4170. doi: 10.5194/bg-9-815-2012. URL https://bg.copernicus.org/articles/9/815/2012/. Publisher: Copernicus GmbH.

L. J. Jensen. One tagger, many uses: Illustrating the power of ontologies in dictionary-based named entity recognition. *bioRxiv*, page 067132, 2016.

L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7(2):119–129, 2006.

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

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*, 16(10):1245–1257, 2013.

N. Jiao, R. Zhang, and Q. Zheng. Coexistence of Two Different Photosynthetic Operons in Citromicrobium bathyomarinum JL354 As Revealed by Whole-Genome Sequencing. *Journal of Bacteriology*, 192(4):1169–1170, Feb. 2010. doi: 10.1128/JB.01504-09. URL https://journals.asm.org/doi/full/10.1128/JB.01504-09. Publisher: American Society for Microbiology.

Q. Jin and M. F. Kirk. ph as a primary control in environmental microbiology: 1. thermodynamic perspective. *Frontiers in Environmental Science*, 6:21, 2018.

X. Jing, H. Gou, Y. Gong, X. Su, L. Xu, Y. Ji, Y. Song, I. P. Thompson, J. Xu, and W. E. Huang. Raman-activated cell sorting and metagenomic sequencing revealing carbon-fixing bacteria in the ocean. *Environmental microbiology*, 20(6):2241–2255, 2018.

F. John. Extremum Problems with Inequalities as Subsidiary Conditions. In G. Giorgi and T. H. Kjeldsen, editors, *Traces and Emergence of Nonlinear Programming*, pages 197–215. Springer, Basel, 2014. ISBN 978-3-0348-0439-4. doi: 10.1007/978-3-0348-0439-4_9. URL https://doi.org/10.1007/978-3-0348-0439-4_9.

J. S. Johnson, D. J. Spakowicz, B.-Y. Hong, L. M. Petersen, P. Demkowicz, L. Chen, S. R. Leopold, B. M. Hanson, H. O. Agresta, M. Gerstein, E. Sodergren, and G. M. Weinstock. Evaluation of 16S rRNA gene sequencing for species and strain-level

microbiome analysis. *Nature Communications*, 10(1):5029, Nov. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13036-1. URL https://www.nature.com/articles/s41467-019-13036-1. Number: 1 Publisher: Nature Publishing Group.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291420.

A. Jousset, C. Bienhold, A. Chatzinotas, L. Gallien, A. Gobet, V. Kurm, K. Küsel, M. C. Rillig, D. W. Rivett, J. F. Salles, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. *The ISME journal*, 11(4):853–862, 2017.

B. B. Jørgensen. Diffusion processes and boundary layers in microbial mats. In L. J. Stal and P. Caumette, editors, *Microbial Mats*, NATO ASI Series, pages 243–253, Berlin, Heidelberg, 1994. Springer. ISBN 978-3-642-78991-5. doi: 10.1007/978-3-642-78991-5_25.

A. T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/25151723.

S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589, June 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4285. URL https://www.nature.com/articles/nmeth.4285. Number: 6 Publisher: Nature Publishing Group.

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

M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, Jan. 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr988. URL https://doi.org/10.1093/nar/gkr988.

D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, July 2019. ISSN 2167-8359. doi: 10.7717/peerj.7359. URL https://peerj.com/articles/7359. Publisher: PeerJ Inc.

E. Karatzas, F. A. Baltoumas, N. A. Panayiotou, R. Schneider, and G. A. Pavlopoulos. Arena3dweb: Interactive 3d visualization of multilayered networks. *Nucleic Acids Research*, 2021.

S. M. Karst, R. M. Ziels, R. H. Kirkegaard, E. A. Sørensen, D. McDonald, Q. Zhu, R. Knight, and M. Albertsen. High-accuracy long-read amplicon sequences using unique molecular identifiers with nanopore or pacbio sequencing. *Nature methods*, 18(2):165–169, 2021.

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*, 30(14): 3059–3066, 2002.

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*, 1:32, 2014.

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

S. Kawai, A. Nishihara, K. Matsuura, and S. Haruta. Hydrogen-dependent autotrophic growth in phototrophic and chemolithotrophic cultures of thermophilic bacteria, Chloroflexus aggregans and Chloroflexus aurantiacus, isolated from Nakabusa hot springs. *FEMS Microbiology Letters*, 366(10):fnz122, May 2019. ISSN 0378-1097. doi: 10.1093/femsle/fnz122. URL https://doi.org/10.1093/femsle/fnz122.

S. Kawai, J. N. Martinez, M. Lichtenberg, E. Trampe, M. Kühl, M. Tank, S. Haruta, A. Nishihara, S. Hanada, and V. Thiel. In-Situ Metatranscriptomic Analyses Reveal the Metabolic Flexibility of the Thermophilic Anoxygenic Photosynthetic Bacterium Chloroflexus aggregans in a Hot Spring Cyanobacteria-Dominated Microbial Mat. *Microorganisms*, 9(3):652, Mar. 2021. ISSN 2076-2607. doi: 10.3390/microorganisms9030652. URL https://www.mdpi.com/2076-2607/9/3/652. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

J. Y. Kim, M.-h. Yi, M. Kim, S. Lee, H. S. Moon, D. Yong, and T.-S. Yong. Measuring the absolute abundance of the microbiome by adding yeast containing 16s rrna gene from a hyperthermophile. *MicrobiologyOpen*, 10(4):e1220, 2021.

G. S. Kindler, H. L. Wong, A. W. D. Larkum, M. Johnson, F. I. MacLeod, and B. P. Burns. Genome-resolved metagenomics provides insights into the functional complexity of microbial mats in Blue Holes, Shark Bay. *FEMS Microbiology Ecology*, 98(1):fiab158, Jan. 2022. ISSN 0168-6496. doi: 10.1093/femsec/fiab158. URL https://doi.org/10.1093/femsec/fiab158.

Z. A. King, J. Lu, A. Drager, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, and N. E. Lewis. Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2016.

M. Kitabatake, M. W. So, D. L. Tumbula, and D. Söll. Cysteine Biosynthesis Pathway in the ArchaeonMethanosarcina barkeri Encoded by Acquired Bacterial Genes? *Journal of Bacteriology*, 182(1):143–145, Jan. 2000. doi: 10.1128/JB.182.1.143-145.2000. URL https://journals.asm.org/doi/10.1128/JB.182.1.143-145.2000. Publisher: American Society for Microbiology.

E. Klipp, W. Liebermeister, C. Wierling, and A. Kowald. *Systems biology: a textbook*. John Wiley & Sons, 2016.

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*, 12(5):e0177643, 2017.

P. Kohl, E. J. Crampin, T. Quinn, and D. Noble. Systems biology: an approach. *Clinical Pharmacology & Therapeutics*, 88(1):25–33, 2010.

W. Kong, D. R. Meldgin, J. J. Collins, and T. Lu. Designing microbial consortia with defined social interactions. *Nature Chemical Biology*, 14(8):821–829, 2018.

S. M. Kosina, A. M. Greiner, R. K. Lau, S. Jenkins, R. Baran, B. P. Bowen, and T. R. Northen. Web of microbes (wom): a curated microbial exometabolomics database for linking chemistry and microbes. *BMC microbiology*, 18(1):1–10, 2018.

M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos. A guide to conquer the biological network era using graph theory. *Frontiers in bioengineering and biotechnology*, 8:34, 2020.

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*, 35(21):4453–4455, 2019.

W. J. Kress, C. García-Robledo, M. Uriarte, and D. L. Erickson. Dna barcodes for ecology, evolution, and conservation. *Trends in ecology & evolution*, 30(1):25–35, 2015.

F. Krueger. Trim Galore, May 2022. URL https://github.com/FelixKrueger/TrimGalore. original-date: 2016-06-27T08:34:40Z.

W. E. Krumbein, D. M. Paterson, and G. A. Zavarzin. Fossil and recent biofilms: a natural history of life on earth. *Fossil and recent biofilms: a natural history of life on earth.*, 2003. URL https://www.cabdirect.org/cabdirect/abstract/20043078795. Publisher: Kluwer Academic Publishers.

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

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*, 4:237, 2013.

C.-H. Kuo and H. Ochman. The extinction dynamics of bacterial pseudogenes. *PLoS genetics*, 6(8):e1001050, 2010.

D. Kurth, D. Elias, M. C. Rasuk, M. Contreras, and M. E. Farías. Carbon fixation and rhodopsin systems in microbial mats from hypersaline lakes Brava and Tebenquiche, Salar de Atacama, Chile. *PLOS ONE*, 16(2):e0246656, 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0246656. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246656. Publisher: Public Library of Science.

G. M. Kurtzer, V. Sochat, and M. W. Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459, 2017.

J. A. Kyndt and T. E. Meyer. Genome Sequences of Allochromatium palmeri and Allochromatium humboldtianum Expand the Allochromatium Family Tree of Purple Sulfur Photosynthetic Bacteria within the Gammaproteobacteria and Further Refine the Genus. *Microbiology Resource Announcements*, 9(33):e00774–20, Aug. 2020. doi: 10.1128/MRA.00774-20. URL https://journals.asm.org/doi/full/10.1128/MRA.00774-20. Publisher: American Society for Microbiology.

H. J. Laanbroek, M.-J. Baar-Gilissen, and H. L. Hoogveld. Nitrite as a stimulus for ammonia-starved nitrosomonas europaea. *Applied and Environmental Microbiology*, 68(3):1454–1457, 2002.

A. Laddha and S. Vempala. Convergence of Gibbs Sampling: Coordinate Hit-and-Run Mixes Fast, 2020.

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*, 2(1):2047–217X, 2013.

B. Langmead and A. Nellore. Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, 19(4):208–219, 2018.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Apr. 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923. URL https://www.nature.com/articles/nmeth.1923. Number: 4 Publisher: Nature Publishing Group.

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*, 7(11):e49334, 2012.

F. W. Larimer, P. Chain, L. Hauser, J. Lamerdin, S. Malfatti, L. Do, M. L. Land, D. A. Pelletier, J. T. Beatty, A. S. Lang, F. R. Tabita, J. L. Gibson, T. E. Hanson, C. Bobst, J. L. T. y. Torres, C. Peres, F. H. Harrison, J. Gibson, and C. S. Harwood. Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. *Nature Biotechnology*, 22(1):55–61, Jan. 2004. ISSN 1546-1696. doi: 10.1038/nbt923. URL https://www.nature.com/articles/nbt923. Number: 1 Publisher: Nature Publishing Group.

I. Laudadio, V. Fulci, F. Palone, L. Stronati, S. Cucchiara, and C. Carissimi. Quantitative assessment of shotgun metagenomics and 16s rdna amplicon sequencing in the study of human gut microbiome. *Omics: a journal of integrative biology*, 22(4):248–254, 2018.

A. M. Laverman, R. W. Canavan, C. P. Slomp, and P. V. Cappellen. Potential nitrate removal in a coastal freshwater sediment (Haringvliet Lake, The Netherlands) and response to salinization. *Water Research*, 41(14):3061–3068, July 2007. ISSN 0043-1354. doi:

10.1016/j.watres.2007.04.002. URL https://www.sciencedirect.com/science/article/pii/S0043135407002497.

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*, 7(1):e30580, 2012.

C. J. D. Lee, P. E. McMullan, C. J. O Kane, A. Stevenson, I. C. Santos, C. Roy, W. Ghosh, R. L. Mancinelli, M. R. Mormile, G. McMullan, H. L. Banciu, M. A. Fares, K. C. Benison, A. Oren, M. L. Dyall-Smith, and J. E. Hallsworth. NaCl-saturated brines are thermodynamically moderate, rather than extreme, microbial habitats. *FEMS Microbiology Reviews*, 42 (5):672–693, Sept. 2018. ISSN 0168-6445. doi: 10.1093/femsre/fuy026. URL https://doi.org/10.1093/femsre/fuy026.

Y. T. Lee and S. S. Vempala. Convergence rate of riemannian hamiltonian monte carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 1115–1121, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188774. URL https://doi.org/10.1145/3188745.3188774.

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

M. Leray, S.-L. Ho, I.-J. Lin, and R. J. Machida. Midori server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics*, 34(21):3753–3754, 2018.

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

G. E. Leventhal, C. Boix, U. Kuechler, T. N. Enke, E. Sliwerska, C. Holliger, and O. X. Cordero. Strain-level diversity drives alternative community types in millimetre-scale granular biofilms. *Nature microbiology*, 3(11):1295–1303, 2018.

R. Levy and E. Borenstein. Reverse ecology: from systems to environments and back. In *Evolutionary systems biology*, pages 329–345. Springer, 2012.

N. E. Lewis, H. Nagarajan, and B. O. Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305, 2012.

D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, 31(10):1674–1676, May 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv033.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://doi.org/10.1093/bioinformatics/btp324.

K. Li, J. Hu, T. Li, F. Liu, J. Tao, J. Liu, Z. Zhang, X. Luo, L. Li, Y. Deng, et al. Microbial abundance and diversity investigations along rivers: Current knowledge and future directions. *Wiley Interdisciplinary Reviews: Water*, 8(5):e1547, 2021.

C. Lima, H. Muhamadali, and R. Goodacre. The role of raman spectroscopy within quantitative metabolomics. *Annual Review of Analytical Chemistry*, 14:323–345, 2021.

E. Lindahl. The scientific case for computing in europe 2018-2026, 2018.

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

M. Loreau. Biodiversity and ecosystem functioning: recent theoretical advances. *Oikos*, 91(1):3–17, 2000.

S. Louca, L. W. Parfrey, and M. Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016.

S. Louca, M. F. Polz, F. Mazel, M. B. Albright, J. A. Huber, M. I. O'Connor, M. Ackermann, A. S. Hahn, D. S. Srivastava, S. A. Crowe, et al. Function and functional redundancy in microbial systems. *Nature ecology & evolution*, 2(6):936–943, 2018.

L. Lovász, R. Kannan, and M. Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11:1–50, 1997.

L. Lovász and S. Vempala. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithms. *J. Computer & System Sciences*, 72:392–417, 2006.

J. Lu and S. L. Salzberg. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome*, 8(1):124, Aug. 2020. ISSN 2049-2618. doi: 10.1186/s40168-020-00900-2. URL https://doi.org/10.1186/s40168-020-00900-2.

W. Ludwig, T. Viver, R. Westram, J. Francisco Gago, E. Bustos-Caparros, K. Knittel, R. Amann, and R. Rossello-Mora. Release LTP_12_2020, featuring a new ARB alignment and improved 16S rRNA tree for prokaryotic type strains. *Systematic and Applied Microbiology*, 44(4):126218, July 2021. ISSN 0723-2020. doi: 10.1016/j.syapm.2021.126218. URL https://www.sciencedirect.com/science/article/pii/S0723202021000412.

T. Lueders, M. Manefield, and M. W. Friedrich. Enhanced sensitivity of DNA- and rRNA-based stable isotope probing by fractionation and quantitative analysis of isopycnic centrifugation gradients. *Environmental Microbiology*, 6(1):73–78, 2004. ISSN 1462-2920. doi: 10.1046/j.1462-2920.2003.00536.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1462-2920.2003.00536.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1462-2920.2003.00536.x.

M. Lularevic, A. J. Racher, C. Jaques, and A. Kiparissides. Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions. *Biotechnology and bioengineering*, 116(9):2339–2352, 2019.

M. MacGillivray, A. Ko, E. Gruber, M. Sawyer, E. Almaas, and A. Holder. Robust analysis of fluxes in genome-scale metabolic pathways. *Scientific Reports*, 7, 12 2017. doi: 10.1038/s41598-017-00170-3.

D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research*, 46(15):7542–7553, 2018.

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*, 4(1):1–7, 2017.

M. Madigan, K. Bender, D. Buckley, W. Sattley, and D. Stahl. Brock biology of microorganisms. 15th global edition. *Boston, US: Benjamin Cummins*, 2018.

R. Mahadevan, J. S. Edwards, and F. J. Doyle III. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical journal*, 83(3):1331–1340, 2002.

F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593, 2014.

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

K. Mainali, S. Bewick, B. Vecchio-Pagan, D. Karig, and W. F. Fagan. Detecting interaction networks in the human microbiome with conditional granger causality. *PLoS computational biology*, 15(5):e1007037, 2019.

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

E. R. Mardis. Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

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

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*, 13(1):1–7, 2012.

F. A. Matsen, R. B. Kodner, and E. V. Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11:538, Oct. 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-538.

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

D. V. Meier, A. J. Greve, A. Chennu, M. R. van Erk, T. Muthukrishnan, R. M. M. Abed, D. Woebken, and D. de Beer. Limitation of Microbial Processes at Saturation-Level Salinities in a Microbial Mat Covering a Coastal Salt Flat. *Applied and Environmental Microbiology*, 87(17):e00698–21, Aug. 2021. doi: 10.1128/AEM.00698-21. URL https://journals.asm.org/doi/full/10.1128/AEM.00698-21. Publisher: American Society for Microbiology.

P. Mell, T. Grance, et al. The nist definition of cloud computing. 2011.

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*, 2014, 2014.

C. v. Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1): 258–261, 2003.

A. Y. Merkel, N. V. Pimenov, I. I. Rusanov, A. I. Slobodkin, G. B. Slobodkina, I. Y. Tarnovetckii, E. N. Frolov, A. V. Dubin, A. A. Perevalova, and E. A. Bonch-Osmolovskaya. Microbial diversity and autotrophic activity in Kamchatka hot springs. *Extremophiles*, 21(2): 307–317, Mar. 2017. ISSN 1433-4909. doi: 10.1007/s00792-016-0903-1. URL https://doi.org/10.1007/s00792-016-0903-1.

F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nature Methods*, pages 1–12, 2022.

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*, 37(5): 1530–1534, May 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa015. URL https://doi.org/10.1093/molbev/msaa015.

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*, 13(6):e0199609, 2018.

A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, et al. Mgnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1):D570–D578, 2020.

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*, pages 1–32, 2020.

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

S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, J. C. Sundaramurthi, J. Lee, M. Kandimalla, I.-M. A. Chen, N. C. Kyrpides, and T. Reddy. Genomes online database (gold) v. 8: overview and updates. *Nucleic Acids Research*, 49(D1):D723–D733, 2021.

J. H. Munson-McGee, E. K. Field, M. Bateson, C. Rooney, R. Stepanauskas, and M. J. Young. Nanoarchaeota, Their Sulfolobales Host, and Nanoarchaeota Virus Distribution across Yellowstone National Park Hot Springs. *Applied and Environmental Microbiology*, 81 (22):7860–7868, Nov. 2015. doi: 10.1128/AEM.01539-15. URL https://journals.asm.org/doi/full/10.1128/AEM.01539-15. Publisher: American Society for Microbiology.

G. Muyzer and A. J. M. Stams. The ecology and biotechnology of sulphate-reducing bacteria. *Nature Reviews Microbiology*, 6(6):441–454, June 2008. ISSN 1740-1534. doi: 10.1038/nrmicro1892. URL https://www.nature.com/articles/nrmicro1892. Number: 6 Publisher: Nature Publishing Group.

T. Nakamura, K. D. Yamada, K. Tomii, and K. Katoh. Parallelization of mafft for large-scale multiple sequence alignments. *Bioinformatics*, 34(14):2490–2492, 2018.

H. Narayanan and P. Srivastava. On the mixing time of coordinate hit-and-run, 2020.

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*, 2(1):1–10, 2019.

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*, 775:145818, 2021.

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

R. H. Nilsson, S. Anslan, M. Bahram, C. Wurzbacher, P. Baldrian, and L. Tedersoo. Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nature Reviews Microbiology*, 17(2):95–109, 2019a.

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*, 47(D1):D259–D264, 2019b.

NOAA. How much water is in the ocean?, 2021. URL https://oceanservice.noaa.gov/facts/oceanwater.html. [Online; accessed 07-January-2022].

M. S. Nobile, P. Cazzaniga, A. Tangherloni, and D. Besozzi. Graphics processing units in bioinformatics, computational biology and systems biology. *Briefings in bioinformatics*, 18(5):870–885, 2017.

D. Noble. *The music of life: biology beyond genes*. Oxford University Press, 2008.

E. Noor, S. Cherkaoui, and U. Sauer. Biological insights through omics data integration. *Current Opinion in Systems Biology*, 15:39–47, 2019.

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*, 98(20):11376–11381, 2001.

A. Noronha, J. Modamio, Y. Jarosz, E. Guerard, N. Sompairac, G. Preciat, A. D. Daníelsdóttir, M. Krecke, D. Merten, H. S. Haraldsdóttir, et al. The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic acids research*, 47(D1):D614–D624, 2019.

C. F. Norton, T. J. McGenity, and W. D. . Grant. Archaeal halophiles (halobacteria) from two British salt mines. *Microbiology*, 139(5):1077–1081, 1993. ISSN 1465-2080,. doi: 10.1099/00221287-139-5-1077. URL https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-139-5-1077. Publisher: Microbiology Society,.

S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, May 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.213959.116. URL https://genome.cshlp.org/content/27/5/824. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

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*, 7:1031, 2020.

J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. vegan: Community Ecology Package, 2020. URL https://CRAN.R-project.org/package=vegan.

B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a web browser. *BMC bioinformatics*, 12(1):1–10, 2011.

R. S. Oremland, P. R. Dowdle, S. Hoeft, J. O. Sharp, J. K. Schaefer, L. G. Miller, J. Switzer Blum, R. L. Smith, N. S. Bloom, and D. Wallschlaeger. Bacterial dissimilatory reduction of arsenate and sulfate in meromictic Mono Lake, California. *Geochimica et Cosmochimica Acta*, 64(18):3073–3084, Sept. 2000. ISSN 0016-7037. doi: 10.1016/S0016-7037(00)00422-1. URL https://www.sciencedirect.com/science/article/pii/S0016703700004221.

A. Oren. Thermodynamic limits to microbial life at high salt concentrations. *Environmental Microbiology*, 13(8):1908–1923, 2011. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2010.02365.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1462-2920.2010.02365.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1462-2920.2010.02365.x.

A. Oren. Cyanobacteria in hypersaline environments: biodiversity and physiological properties. *Biodiversity and Conservation*, 24(4):781–798, Apr. 2015. ISSN 1572-9710. doi: 10.1007/s10531-015-0882-z. URL https://doi.org/10.1007/s10531-015-0882-z.

J. D. Orth, I. Thiele, and B. Ø.. Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

R. Overbeek, R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, B. Parrello, M. Shukla, et al. The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic acids research*, 42(D1):D206–D214, 2014.

O. E. Owen, S. C. Kalhan, and R. W. Hanson. The key role of anaplerosis and cataplerosis for citric acid cycle function. *Journal of Biological Chemistry*, 277(34):30409–30412, 2002.

A. R. Pacheco, M. Moel, and D. Segrè. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature communications*, 10(1):1–12, 2019.

H. W. Paerl, J. L. Pinckney, and T. F. Steppe. Cyanobacterial–bacterial mat consortia: examining the functional unit of microbial survival and growth in extreme environments. *Environmental Microbiology*, 2(1):11–26, 2000. ISSN 1462-2920. doi: 10.1046/j.1462-2920.2000.00071.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1462-2920.2000.00071.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1462-2920.2000.00071.x.

E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390, 2013.

E. Pafilis, S. P. Frankild, J. Schnetzer, L. Fanini, S. Faulwetter, C. Pavloudi, K. Vasileiadou, P. Leary, J. Hammock, K. Schulz, et al. Environments and eol: identification of environment ontology terms in text and the annotation of the encyclopedia of life. *Bioinformatics*, 31(11):1872–1874, 2015.

E. Pafilis, P. L. Buttigieg, B. Ferrell, E. Pereira, J. Schnetzer, C. Arvanitidis, and L. J. Jensen. Extract: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. *Database*, 2016, 2016.

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*, page 100869, 2020.

M. J. Pallen, A. Telatin, and A. Oren. The next million names for archaea and bacteria. *Trends in Microbiology*, 29(4):289–298, 2021.

B. Ø.. Palsson. Metabolic systems biology. *FEBS letters*, 583(24):3900–3904, 2009.

B. Ø.. Palsson. *Systems biology*. Cambridge university press, 2015.

M. Papadaki, E. Kaitetzidou, C. C. Mylonas, and E. Sarropoulou. Non-coding rna expression patterns of two different teleost gonad maturation stages. *Marine Biotechnology*, 22(5):683–695, 2020.

D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, July 2015. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.186072.114. URL https://genome.cshlp.org/content/25/7/1043. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

D. H. Parks, M. Chuvochina, P.-A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz. A complete domain-to-species taxonomy for bacteria and archaea. *Nature biotechnology*, 38(9):1079–1086, 2020.

D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, and P. Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, Jan. 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab776. URL https://doi.org/10.1093/nar/gkab776.

E. Parmentier, D. Lanterbecq, and I. Eeckhaut. From commensalism to parasitism in carapidae (ophidiiformes): heterochronic modes of development? *PeerJ*, 4:e1786, 2016.

A. C. Parte, J. S. Carbasse, J. P. Meier-Kolthoff, L. C. Reimer, and M. Göker. List of prokaryotic names with standing in nomenclature (lpsn) moves to the dsmz. *International journal of systematic and evolutionary microbiology*, 70(11):5607, 2020.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, Apr. 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4197. URL https://www.nature.com/articles/nmeth.4197. Number: 4 Publisher: Nature Publishing Group.

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*, 1(1):1–13, 2018.

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*, 41:23–33, 2019.

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

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*, 5:e3687, 2017a.

C. Pavloudi, A. Oulas, K. Vasileiadou, G. Kotoulas, M. De Troch, M. W. Friedrich, and C. Arvanitidis. Diversity and abundance of sulfate-reducing microorganisms in a mediterranean lagoonal complex (amvrakikos gulf, ionian sea) derived from dsrb gene. *Aquatic Microbial Ecology*, 79(3):209–219, 2017b.

A. Y. Pei, W. E. Oberdorf, C. W. Nossa, A. Agarwal, P. Chokshi, E. A. Gerz, Z. Jin, P. Lee, L. Yang, M. Poles, et al. Diversity of 16s rrna genes within individual prokaryotic genomes. *Applied and environmental microbiology*, 76(12):3886–3897, 2010.

O. Perez-Garcia, G. Lear, and N. Singhal. Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Frontiers in microbiology*, 7: 673, 2016.

S. Pesant, F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, et al. Open science resources for the discovery and analysis of tara oceans data. *Scientific data*, 2(1):1–16, 2015.

B. K. Pierson and R. W. Castenholz. A phototrophic gliding filamentous bacterium of hot springs, Chloroflexus aurantiacus, gen. and sp. nov. *Archives of Microbiology*, 100(1): 5–24, Jan. 1974. ISSN 1432-072X. doi: 10.1007/BF00446302. URL https://doi.org/10.1007/BF00446302.

S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89, 2015.

M. Podar, K. S. Makarova, D. E. Graham, Y. I. Wolf, E. V. Koonin, and A.-L. Reysenbach. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biology Direct*, 8(1):9, Apr. 2013. ISSN 1745-6150. doi: 10.1186/1745-6150-8-9. URL https://doi.org/10.1186/1745-6150-8-9.

T. M. Porter. terrimporter/12SvertebrateClassifier: 12S Vertebrate Classifier v2.0.0-ref, Aug. 2021. URL https://doi.org/10.5281/zenodo.5157047.

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

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.

N. D. Price, J. Schellenberger, and B. O. Palsson. Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. *Biophysical journal*, 87(4):2172–2186, 2004.

C. M. Prieto-Barajas, E. Valencia-Cantero, and G. Santoyo. Microbial mat ecosystems: Structure types, functional diversity, and biotechnological application. *Electronic Journal of Biotechnology*, 31:48–56, Jan. 2018. ISSN 0717-3458. doi: 10.1016/j.ejbt.2017.11.001. URL https://www.sciencedirect.com/science/article/pii/S0717345817300738.

Q.-L. Qin, B.-B. Xie, X.-Y. Zhang, X.-L. Chen, B.-C. Zhou, J. Zhou, A. Oren, and Y.-Z. Zhang. A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology*, 196(12):2210–2215, June 2014. ISSN 1098-5530. doi: 10.1128/JB.01688-14.

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*, 41(D1):D590–D596, Jan. 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1219. URL https://doi.org/10.1093/nar/gks1219.

R. A. Quinn, J. A. Navas-Molina, E. R. Hyde, S. J. Song, Y. Vázquez-Baeza, G. Humphrey, J. Gaffney, J. J. Minich, A. V. Melnik, J. Herschend, J. DeReus, A. Durant, R. J. Dutton, M. Khosroheidari, C. Green, R. da Silva, P. C. Dorrestein, and R. Knight. From sample to multi-omics conclusions in under 48 hours. msystems 1: e00038-16. *Crossref, Medline*, 2016.

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)*, 17(3):228, 2017.

J. Raes and P. Bork. Molecular eco-systems biology: towards an understanding of community function. *Nature Reviews Microbiology*, 6(9):693–699, 2008.

S. W. Ragsdale and E. Pierce. Acetogenesis and the Wood-Ljungdahl Pathway of CO2 Fixation. *Biochimica et biophysica acta*, 1784(12):1873–1898, Dec. 2008. ISSN 0006-3002. doi: 10.1016/j.bbapap.2008.08.012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2646786/.

S. Ratnasingham and P. D. Hebert. Bold: The barcode of life data system (http://www.barcodinglife. org). *Molecular ecology notes*, 7(3):355–364, 2007.

J. L. Reed. Shrinking the metabolic solution space using experimental datasets. 2012.

J. Regalado, D. S. Lundberg, O. Deusch, S. Kersten, T. Karasov, K. Poersch, G. Shirsekar, and D. Weigel. Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves. *The ISME Journal*, 14(8):2116–2130, Aug. 2020. ISSN 1751-7370. doi: 10.1038/

s41396-020-0665-8. URL https://www.nature.com/articles/s41396-020-0665-8. Number: 8 Publisher: Nature Publishing Group.

L. C. Reimer, A. Vetcininova, J. S. Carbasse, C. Söhngen, D. Gleim, C. Ebeling, and J. Overmann. Bac dive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. *Nucleic acids research*, 47(D1):D631–D636, 2019.

K. Remoundou, P. Koundouri, A. Kontogianni, P. A. Nunes, and M. Skourtos. Valuation of natural marine ecosystems: an economic perspective. *environmental science & policy*, 12(7):1040–1051, 2009.

R. J. Roberts. Pubmed central: The genbank of the published literature, 2001.

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

S. Ronconi, R. Jonczyk, and U. Genschel. A novel isoform of pantothenate synthetase in the Archaea. *The FEBS Journal*, 275(11):2754–2764, 2008. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2008.06416.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-4658.2008.06416.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1742-4658.2008.06416.x.

L. Röttjers and K. Faust. From hairballs to hypotheses–biological insights from microbial networks. *FEMS microbiology reviews*, 42(6):761–780, 2018.

D. D. Roumpeka, R. J. Wallace, F. Escalettes, I. Fotheringham, and M. Watson. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in genetics*, 8:23, 2017.

V. Roy. Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, 7(1):387–412, 2020. doi: 10.1146/annurev-statistics-031219-041300.

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*, 17:e00547, 2019.

R. Ruvindy, R. A. White III, B. A. Neilan, and B. P. Burns. Unravelling core microbial metabolisms in the hypersaline microbial mats of Shark Bay using high-throughput metagenomics. *The ISME Journal*, 10(1):183–196, Jan. 2016. ISSN 1751-7370. doi: 10.1038/ismej.2015.87. URL https://www.nature.com/articles/ismej201587. Number: 1 Publisher: Nature Publishing Group.

P. A. Saa and L. K. Nielsen. ll-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks. *Bioinform.*, 32(15):2330–2337, 2016. doi: 10.1093/bioinformatics/btw132. URL https://doi.org/10.1093/bioinformatics/btw132.

A. Saghaï, A. Gutiérrez-Preciado, P. Deschamps, D. Moreira, P. Bertolino, M. Ragon, and P. López-García. Unveiling microbial interactions in stratified mat communities from a warm saline shallow pond. *Environmental microbiology*, 19(6):2405–2421, 2017.

A. Sahu, M.-A. Blätke, J. J. Szymański, and N. Töpfer. Advances in flux balance analysis by integrating machine learning and mechanism-based models. *Computational and Structural Biotechnology Journal*, 19:4626–4640, 2021.

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*, 32(1):305–332, 2001.

E. Sala and N. Knowlton. Global marine biodiversity trends. *Annu. Rev. Environ. Resour.*, 31:93–122, 2006.

J. Saldida, A. P. Muntoni, D. de Martino, G. Hubmann, B. Niebel, A. M. Schmidt, A. Braunstein, A. Milias-Argeits, and M. Heinemann. Unbiased metabolic flux inference through combined thermodynamic and 13c flux analysis. *bioRxiv*, 2020.

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

C. Sammut and G. I. Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

R. M. Samuel, R. Meyer, P. L. Buttigieg, N. Davies, N. W. Jeffery, C. Meyer, C. Pavloudi, K. Johnson Pitz, M. Sweetlove, S. Theroux, et al. Towards a global public repository of community protocols to encourage best practices in biomolecular ocean observing and research. *Frontiers in Marine Science*, page 1488, 2021.

A. Santos-Lopez, C. W. Marshall, M. R. Scribner, D. J. Snyder, and V. S. Cooper. Evolutionary pathways to antibiotic resistance are dependent upon environmental structure and bacterial lifestyle. *Elife*, 8:e47612, 2019.

G. Santoyo. Unveiling Taxonomic Diversity and Functional Composition Differences of Microbial Mat Communities Through Comparative Metagenomics. *Geomicrobiology Journal*, 38(7):639–648, July 2021. ISSN 0149-0451. doi: 10.1080/01490451.2021.1926600. URL https://doi.org/10.1080/01490451.2021.1926600. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01490451.2021.1926600.

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*, 6(12):gix108, 2017.

E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 49(D1):D10, 2021.

J. Schellenberger and B. Ø.. Palsson. Use of randomized sampling for analysis of metabolic networks. *Journal of biological chemistry*, 284(9):5457–5461, 2009.

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*, 36(7):1004–1013, 2020.

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*, 12(1):e0170037, 2017.

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*, 75(23): 7537–7541, 2009.

M. V. Schneider and R. C. Jimenez. Teaching the fundamentals of biological data integration using classroom games. *PLoS computational biology*, 8(12):e1002789, 2012.

C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020, 2020.

J. R. Schramski, A. I. Dell, J. M. Grady, R. M. Sibly, and J. H. Brown. Metabolic theory predicts whole-ecosystem properties. *Proceedings of the National Academy of Sciences*, 112(8):2617–2622, 2015.

K. Schuchmann and V. Müller. Energetics and Application of Heterotrophy in Acetogenic Bacteria. *Applied and Environmental Microbiology*, 82(14):4056–4069, June 2016. ISSN 0099-2240. doi: 10.1128/AEM.00882-16. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4959221/.

W. T. Scott, E. J. Smid, D. E. Block, and R. A. Notebaart. Metabolic flux sampling predicts strain-dependent differences related to aroma production among commercial wine yeasts. *Microbial cell factories*, 20(1):1–15, 2021.

T. Seemann. Barrnap: BAsic Rapid Ribosomal RNA Predictor, 2014a. URL https://github.com/tseemann/barrnap.

T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14): 2068–2069, July 2014b. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu153. URL https://doi.org/10.1093/bioinformatics/btu153.

N. Segata. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems*, 3(2): e00190–17, 2018. doi: 10.1128/mSystems.00190-17. URL https://journals.asm.org/doi/10.1128/mSystems.00190-17. Publisher: American Society for Microbiology.

H. Shaaban, D. A. Westfall, R. Mohammad, D. Danko, D. Bezdan, E. Afshinnekoo, N. Segata, and C. E. Mason. The microbe directory: An annotated, searchable inventory of microbes' characteristics. *Gates open research*, 2, 2018.

T. J. Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 2014. ISSN 1664-462X. URL https://www.frontiersin.org/article/10.3389/fpls.2014.00209.

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

S. S. Shishvan, A. Vigliotti, and V. S. Deshpande. The homeostatic ensemble for cells. *Biomechanics and Modeling in Mechanobiology*, 17(6):1631–1662, 2018.

I.-I. Shiung, M.-J. Chang, Y.-T. Chang, S.-L. Yeh, S.-J. Chang, C. Ying, and W.-L. Chao. Photosynthetic purple sulfur bacterium Marichromatium purpuratum RuA2 induces changes in water quality parameters, the occurrence of sulfonamide resistance gene and microbial community structure of marine aquaculture. *Aquaculture*, 493:68–78, Aug. 2018. ISSN 0044-8486. doi: 10.1016/j.aquaculture.2018.04.055. URL https://www.sciencedirect.com/science/article/pii/S0044848617307883.

W.-S. Shu and L.-N. Huang. Microbial diversity in extreme environments. *Nature Reviews Microbiology*, pages 1–17, 2021.

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*, 58(4):445–451, 2009.

H. Y. Simon, K. J. Siddle, D. J. Park, and P. C. Sabeti. Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794, 2019.

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

L. Sinclair, U. Z. Ijaz, L. J. Jensen, M. J. Coolen, C. Gubry-Rangin, A. Chroňáková, A. Oulas, C. Pavloudi, J. Schnetzer, A. Weimann, et al. Seqenv: linking sequences to environments through text mining. *PeerJ*, 4:e2690, 2016.

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*, 3:92, 2016.

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

J.-i. Sohn and J.-W. Nam. The present and future of de novo whole-genome assembly. *Briefings in bioinformatics*, 19(1):23–40, 2018.

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*, 105(36):13486–13491, 2008.

W.-Z. Song and T. Thomas. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics*, 33(12):1873–1875, 2017.

S. Spring, H. Lünsdorf, B. M. Fuchs, and B. J. Tindall. The Photosynthetic Apparatus and Its Regulation in the Aerobic Gammaproteobacterium Congregibacter litoralis gen. nov., sp. nov. *PLOS ONE*, 4(3):e4866, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0004866. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004866. Publisher: Public Library of Science.

L. J. Stal. Cyanobacterial Mats and Stromatolites. In B. A. Whitton, editor, *Ecology of Cyanobacteria II: Their Diversity in Space and Time*, pages 65–125. Springer Netherlands, Dordrecht, 2012. ISBN 978-94-007-3855-3. doi: 10.1007/978-94-007-3855-3_4. URL https://doi.org/10.1007/978-94-007-3855-3_4.

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*, 7(1):1–11, 2017.

A. D. Steen, A. Crits-Christoph, P. Carini, K. M. DeAngelis, N. Fierer, K. G. Lloyd, and J. Cameron Thrash. High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal*, 13(12):3126–3130, 2019.

J. L. Steenwyk, T. J. B. Iii, Y. Li, X.-X. Shen, and A. Rokas. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biology*, 18 (12):e3001007, Dec. 2020. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001007. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001007. Publisher: Public Library of Science.

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

A. Steinberger. asteinberger9/seq_scripts, 2020. URL https://doi.org/10.5281/zenodo.4270481.

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

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

R. Stevens, C. A. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Briefings in bioinformatics*, 1(4):398–414, 2000.

W. Stoeckenius, R. H. Lozier, and R. A. Bogomolni. Bacteriorhodopsin and the purple membrane of halobacteria. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics*, 505(3):215–278, Mar. 1979. ISSN 0304-4173. doi: 10.1016/0304-4173(79)90006-5. URL https://www.sciencedirect.com/science/article/pii/0304417379900065.

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

P. Sudhir and S. Murthy. Effects of salt stress on basic processes of photosynthesis. *Photosynthetica*, 42(4):481–486, Dec. 2004. ISSN 1573-9058. doi: 10.1007/S11099-005-0001-6. URL https://doi.org/10.1007/S11099-005-0001-6.

L. Suter, A. M. Polanowski, L. J. Clarke, J. A. Kitchener, and B. E. Deagle. Capturing open ocean biodiversity: comparing environmental dna metabarcoding to the continuous plankton recorder. *Molecular ecology*, 30(13):3140–3157, 2021.

N. Swainston, K. Smallbone, H. Hefzi, P. D. Dobson, J. Brewer, M. Hanscho, D. C. Zielinski, K. S. Ang, N. J. Gardiner, J. M. Gutierrez, S. Kyriakopoulos, M. Lakshmanan, S. Li, J. K. Liu, V. S. Martínez, C. A. Orellana, L.-E. Quek, A. Thomas, J. Zanghellini, N. Borth, D.-Y. Lee, L. K. Nielsen, D. B. Kell, N. E. Lewis, and P. Mendes. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12(7):109, June 2016. ISSN 1573-3890. doi: 10.1007/s11306-016-1051-4. URL https://doi.org/10.1007/s11306-016-1051-4.

D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

K. B. Sørensen, D. E. Canfield, and A. Oren. Salinity Responses of Benthic Microbial Communities in a Solar Saltern (Eilat, Israel). *Applied and Environmental Microbiology*, 70(3):1608–1616, Mar. 2004. doi: 10.1128/AEM.70.3.1608-1616.2004. URL https://journals.asm.org/doi/10.1128/AEM.70.3.1608-1616.2004. Publisher: American Society for Microbiology.

P. Taberlet, A. Bonin, L. Zinger, and E. Coissac. Analysis of bulk samples. In *Environmental DNA*, pages 140–143. Oxford University Press.

P. Taberlet, E. Coissac, M. Hajibabaei, and L. H. Rieseberg. Environmental DNA. *Molecular Ecology*, 21(8):1789–1793, 2012a. ISSN 1365-294X. doi: 10.1111/j.1365-294X.2012.05542.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2012.05542.x.

P. Taberlet, E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev. Towards next-generation biodiversity assessment using dna metabarcoding. *Molecular ecology*, 21(8): 2045–2050, 2012b.

K.-H. Tang, K. Barry, O. Chertkov, E. Dalin, C. S. Han, L. J. Hauser, B. M. Honchak, L. E. Karbach, M. L. Land, A. Lapidus, F. W. Larimer, N. Mikhailova, S. Pitluck, B. K. Pierson, and R. E. Blankenship. Complete genome sequence of the filamentous anoxygenic phototrophic bacterium Chloroflexus aurantiacus. *BMC Genomics*, 12(1):334, June 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-334. URL https://doi.org/10.1186/1471-2164-12-334.

Y. Tang, T. Dai, Z. Su, K. Hasegawa, J. Tian, L. Chen, and D. Wen. A tripartite microbial-environment network indicates how crucial microbes influence the microbial community ecology. *Microbial ecology*, 79(2):342–356, 2020.

S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. Differential expression in rna-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

I. Tavassoly, J. Goldfarb, and R. Iyengar. Systems biology primer: the basic methods and approaches. *Essays in biochemistry*, 62(4):487–500, 2018.

L. Tedersoo, M. Albertsen, S. Anslan, and B. Callahan. Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied and Environmental Microbiology*, 87(17):e00626–21, 2021.

M. Tessler, J. S. Neumann, E. Afshinnekoo, M. Pineda, R. Hersch, L. F. M. Velho, B. T. Segovia, F. A. Lansac-Toha, M. Lemke, R. DeSalle, C. E. Mason, and M. R. Brugler. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, 7(1):6589, July 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-06665-3. URL https://www.nature.com/articles/s41598-017-06665-3. Number: 1 Publisher: Nature Publishing Group.

F. Thanati, E. Karatzas, F. Baltoumas, D. J. Stravopodis, A. G. Eliopoulos, and G. Pavlopoulos. Flame: a web tool for functional and literature enrichment analysis of multiple gene lists. *bioRxiv*, 2021.

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

P. F. Thomsen and E. Willerslev. Environmental dna–an emerging tool in conservation for monitoring past and present biodiversity. *Biological conservation*, 183:4–18, 2015.

J. H. Tidwell and G. L. Allan. Fish as food: aquaculture's contribution. *EMBO reports*, 2 (11):958–963, 2001.

B. J. Tindall. Note: Proposals to update and make changes to the Bacteriological Code. *International Journal of Systematic and Evolutionary Microbiology*, 49(3):1309–1312, 1999. ISSN 1466-5034,. doi: 10.1099/00207713-49-3-1309. URL https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-49-3-1309. Publisher: Microbiology Society,.

L. Tong. Structure and function of biotin-dependent carboxylases. *Cellular and Molecular Life Sciences*, 70(5):863–891, Mar. 2013. ISSN 1420-9071. doi: 10.1007/s00018-012-1096-0. URL https://doi.org/10.1007/s00018-012-1096-0.

T. Tonon and D. Eveillard. Marine systems biology. *Frontiers in genetics*, 6:181, 2015.

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

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*, 12(7):263, 2020.

G. V. Uritskiy, J. DiRuggiero, and J. Taylor. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, 6(1):158, Sept. 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0541-1. URL https://doi.org/10.1186/s40168-018-0541-1.

M. Uschold, M. King, S. Moralee, and Y. Zorgios. The enterprise ontology. *The knowledge engineering review*, 13(1):31–89, 1998.

L. M. van der Loos and R. Nijland. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, 30(13):3270–3288, 2021. ISSN 1365-294X. doi: 10.1111/mec.15592. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15592. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15592.

H. van Gemerden. Microbial mats: A joint venture. *Marine Geology*, 113(1):3–25, July 1993. ISSN 0025-3227. doi: 10.1016/0025-3227(93)90146-M. URL https://www.sciencedirect.com/science/article/pii/002532279390146M.

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*, 13(4):e0194599, 2018.

P. Vangay, J. Burgin, A. Johnston, K. L. Beck, D. C. Berrios, K. Blumberg, S. Canon, P. Chain, J.-M. Chandonia, D. Christianson, et al. Microbiome metadata standards: Report of the national microbiome data collaborative's workshop and follow-on activities. *Msystems*, 6(1):e01194–20, 2021.

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*, (4), 2016.

O. S. Venturelli, A. V. Carr, G. Fisher, R. H. Hsu, R. Lau, B. P. Bowen, S. Hromada, T. Northen, and A. P. Arkin. Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular systems biology*, 14(6):e8157, 2018.

B. Veuger, A. Pitcher, S. Schouten, J. S. Sinninghe Damsté, and J. J. Middelburg. Nitrification and growth of autotrophic nitrifying bacteria and Thaumarchaeota in the coastal North Sea. *Biogeosciences*, 10(3):1775–1785, Mar. 2013. ISSN 1726-4170. doi: 10.5194/bg-10-1775-2013. URL https://bg.copernicus.org/articles/10/1775/2013/. Publisher: Copernicus GmbH.

L. Villanueva, A. Navarrete, J. Urmeneta, D. C. White, and R. Guerrero. Analysis of diurnal and vertical microbial diversity of a hypersaline microbial mat. *Archives of Microbiology*,

188(2):137–146, Aug. 2007. ISSN 1432-072X. doi: 10.1007/s00203-007-0229-6. URL https://doi.org/10.1007/s00203-007-0229-6.

P. T. Visscher and J. F. Stolz. Microbial mats as bioreactors: populations, processes, and products. In N. Noffke, editor, *Geobiology: Objectives, Concepts, Perspectives*, pages 87–100. Elsevier, Amsterdam, Jan. 2005. ISBN 978-0-444-52019-7. doi: 10.1016/B978-0-444-52019-7.50009-7. URL https://www.sciencedirect.com/science/article/pii/B9780444520197500097.

P. T. Visscher, K. L. Gallagher, A. Bouton, M. E. Farias, D. Kurth, M. Sancho-Tomás, P. Philippot, A. Somogyi, K. Medjoubi, E. Vennin, R. Bourillot, M. R. Walter, B. P. Burns, M. Contreras, and C. Dupraz. Modern arsenotrophic microbial mats provide an analogue for life in the anoxic Archean. *Communications Earth & Environment*, 1(1):1–10, Sept. 2020. ISSN 2662-4435. doi: 10.1038/s43247-020-00025-2. URL https://www.nature.com/articles/s43247-020-00025-2. Number: 1 Publisher: Nature Publishing Group.

C. Von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. String: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research*, 33(suppl_1):D433–D437, 2005.

P. D. Vouzis and N. V. Sahinidis. Gpu-blast: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, 27(2):182–188, 2011.

R. L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P. L. Buttigieg, N. Davies, D. Endresen, et al. Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. *PloS one*, 9(3):e89606, 2014.

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*, 73(16):5261–5267, 2007.

T. Ward, J. Larson, J. Meulemans, B. Hillmann, J. Lynch, D. Sidiropoulos, J. R. Spear, G. Caporaso, R. Blekhman, R. Knight, R. Fink, and D. Knights. Bugbase predicts organism-level microbiome phenotypes. *bioRxiv*, 2017. doi: 10.1101/133462. URL https://www.biorxiv.org/content/early/2017/05/02/133462.

R. Warwick and P. Somerfield. All animals are equal, but some animals are more equal than others. *Journal of Experimental Marine Biology and Ecology*, 366(1-2):184–186, 2008.

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*, 678:499–524, 2019.

D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS computational biology*, 14(2):e1005962, 2018.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Wikipedia contributors. Supercomputing in europe — Wikipedia, the free encyclopedia, 2021. URL https://en.wikipedia.org/w/index.php?title=Supercomputing_in_Europe&oldid=1009652575. [Online; accessed 7-January-2022].

A. Wilke, J. Bischof, T. Harrison, T. Brettin, M. D'Souza, W. Gerlach, H. Matthews, T. Paczian, J. Wilkening, E. M. Glass, et al. A restful api for accessing microbial community data for mg-rast. *PLoS computational biology*, 11(1):e1004008, 2015.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

H. L. Wong, A. Ahmed-Cox, and B. P. Burns. Molecular Ecology of Hypersaline Microbial Mats: Current Insights and New Directions. *Microorganisms*, 4(1):6, Mar. 2016. ISSN 2076-2607. doi: 10.3390/microorganisms4010006. URL https://www.mdpi.com/2076-2607/4/1/6. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

H. L. Wong, F. I. MacLeod, R. A. White, P. T. Visscher, and B. P. Burns. Microbial dark matter filling the niche in hypersaline microbial mats. *Microbiome*, 8(1):135, Sept. 2020. ISSN 2049-2618. doi: 10.1186/s40168-020-00910-0. URL https://doi.org/10.1186/s40168-020-00910-0.

D. E. Wood, J. Lu, and B. Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, Nov. 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1891-0. URL https://doi.org/10.1186/s13059-019-1891-0.

E. M. Wood-Charlson, D. Auberry, H. Blanco, M. I. Borkum, Y. E. Corilo, K. W. Davenport, S. Deshpande, R. Devarakonda, M. Drake, W. D. Duncan, et al. The national microbiome data collaborative: enabling microbiome science. *Nature Reviews Microbiology*, 18(6): 313–314, 2020.

Y.-W. Wu, B. A. Simmons, and S. W. Singer. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, Feb. 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv638. URL https://doi.org/10.1093/bioinformatics/btv638.

L. Wurch, R. J. Giannone, B. S. Belisle, C. Swift, S. Utturkar, R. L. Hettich, A.-L. Reysenbach, and M. Podar. Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nature Communications*, 7(1):12115, July 2016. ISSN 2041-1723. doi: 10.1038/ncomms12115. URL

https://www.nature.com/articles/ncomms12115. Number: 1 Publisher: Nature Publishing Group.

Z. Xu, D. Malmer, M. G. Langille, S. F. Way, and R. Knight. Which is more important for classifying microbial communities: who's there or what they can do? *The ISME journal*, 8(12):2357–2359, 2014.

C.-X. Xue, H. Lin, X.-Y. Zhu, J. Liu, Y. Zhang, G. Rowley, J. D. Todd, M. Li, and X.-H. Zhang. DiTing: A Pipeline to Infer and Compare Biogeochemical Pathways From Metagenomic and Metatranscriptomic Data. *Frontiers in Microbiology*, 12, 2021. ISSN 1664-302X. URL https://www.frontiersin.org/article/10.3389/fmicb.2021.698286.

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*, 56(1):73–81, 2013.

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*, 46:379–389, 2014.

C. Yang, D. Chowdhury, Z. Zhang, W. K. Cheung, A. Lu, Z. Bian, and L. Zhang. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, 19:6301–6314, 2021.

C. S. Yentsch and D. W. Menzel. A method for the determination of phytoplankton chlorophyll and phaeophytin by fluorescence. *Deep Sea Research and Oceanographic Abstracts*, 10(3):221–231, July 1963. ISSN 0011-7471. doi: 10.1016/0011-7471(63)90358-9. URL https://www.sciencedirect.com/science/article/pii/0011747163903589.

P. Yilmaz, J. A. Gilbert, R. Knight, L. Amaral-Zettler, I. Karsch-Mizrachi, G. Cochrane, Y. Nakamura, S.-A. Sansone, F. O. Gloeckner, and D. Field. The genomic standards consortium: bringing standards to life for microbial ecology. *The ISME journal*, 5(10): 1565–1567, 2011a.

P. Yilmaz, R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I. Karsch-Mizrachi, A. Johnston, G. Cochrane, et al. Minimum information about a marker gene sequence (mimarks) and minimum information about any (x) sequence (mixs) specifications. *Nature biotechnology*, 29(5):415–420, 2011b.

Y. Yue, H. Huang, Z. Qi, H.-M. Dou, X.-Y. Liu, T.-F. Han, Y. Chen, X.-J. Song, Y.-H. Zhang, and J. Tu. Evaluating metagenomics tools for genome binning with real metagenomic datasets and cami datasets. *BMC bioinformatics*, 21(1):1–15, 2020.

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*, 9(3): giaa022, 2020.

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*, 5:e69657, 2021a.

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. 2021b. URL https://doi.org/10.5281/zenodo.4665308.

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. 0s and 1s in marine molecular research: a regional HPC perspective. *GigaScience*, 10(8):giab053, Aug. 2021c. ISSN 2047-217X. doi: 10.1093/gigascience/giab053. URL https://doi.org/10.1093/gigascience/giab053.

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. Technical report, Zenodo, Dec. 2021d. URL https://zenodo.org/record/4665308.

H. Zafeiropoulos, S. Paragkamian, S. Ninidakis, G. A. Pavlopoulos, L. J. Jensen, and E. Pafilis. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. *Microorganisms*, 10(2):293, Feb. 2022. ISSN 2076-2607. doi: 10.3390/microorganisms10020293. URL https://www.mdpi.com/2076-2607/10/2/293. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

O. Zemb, C. S. Achard, J. Hamelin, M.-L. De Almeida, B. Gabinaud, L. Cauquil, L. M. Verschuren, and J.-J. Godon. Absolute quantitation of microbes using 16s rrna gene metabarcoding: A rapid normalization of relative abundances by quantitative pcr targeting a 16s rrna gene spike-in standard. *Microbiologyopen*, 9(3):e977, 2020.

M. L. Zeng and J. Qin. *Metadata, Second Edition*. Chicago : Neal-Schuman, 2016.

Y. Zeng and M. Koblizek. Phototrophic Gemmatimonadetes: A New "Purple" Branch on the Bacterial Tree of Life. In P. C. Hallenbeck, editor, *Modern Topics in the Phototrophic Prokaryotes: Environmental and Applied Aspects*, pages 163–192. Springer International Publishing, Cham, 2017. ISBN 978-3-319-46261-5. doi: 10.1007/978-3-319-46261-5_5. URL https://doi.org/10.1007/978-3-319-46261-5_5.

R. Zheng, R. Cai, C. Wang, R. Liu, and C. Sun. Characterization of the First Cultured Representative of "Candidatus Thermofonsia" Clade 2 within Chloroflexi Reveals Its Phototrophic Lifestyle. *mBio*, 13(2):e00287–22, Apr. 2022. doi: 10.1128/mbio.00287-22. URL https://journals.asm.org/doi/full/10.1128/mbio.00287-22. Publisher: American Society for Microbiology.

K. Zhuang, M. Izallalen, P. Mouser, H. Richter, C. Risso, R. Mahadevan, and D. R. Lovley. Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *The ISME journal*, 5(2):305–316, 2011.

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

J. Zoppi, J.-F. Guillaume, M. Neunlist, and S. Chaffron. Mibiomics: an interactive web application for multi-omics data exploration and integration. *BMC bioinformatics*, 22 (1):1–14, 2021.

Short CV

Education

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

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

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

Research projects - working Experience

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

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

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

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

Publications

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

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

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

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

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

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

Under review

- **Deciphering the functional potential of a hypersaline swamp microbial mat community** Pavloudi, C., **Zafeiropoulos, H.**
  Under reviw in *FEMS Microbiology Ecology*

[2] ZH and PS contibuted equally
[3] ZH and GA contributed equally

- **Automating the curation process of historical literature on marine biodiversity using text mining: the DECO workflow** Paragkamian, S., Sarafidou, G., Mavraki, D., .., **Zafeiropoulos H.**, Arvanitidis, C., Pafilis, E., Gerovasileiou, V.
  Under review in *Frontiers in Marine Science*
- **Metabolic models of human gut microbiota: what did we learn and what are the next steps** Garza, D.R., Gonze, D., **Zafeiropoulos, H.**, Liu, B., Faust, K.
  Under review in *Cell Systems*

In preparation

- Metagenome assembled genomes of novel prokaryotic taxa from a hypersaline marsh microbial mat
- dingo: a Python library for metabolic networks analysis

Awards

- **European Molecular Biology Organization Short-Term Fellowship** (2022)
  Project title: Exploiting data integration, text-mining and computational geometry to enhance microbial interactions inference from co-occurrence networks
  https://hariszaf.github.io/microbetag/
- **Mikrobiokosmos travel grant in memorium of Prof. Kostas Drainas** (2021)
- **Google Summer of Code** (2021)
  Project title: From DNA sequences to metabolic interactions: building a pipeline to extract key metabolic processes
  Report, GSoC archive
- **Federation of European Microbiological Societies Meeting Attendance Grant** (2020)
  for joining the *Metagenomics, Metatranscript- omics and multi 'omics for microbial community studies* Physalia course
- **Short Term Scientific Mission (STSM) - DNAqua-net COST action** (2019)
  Project title: A comparison of bioinformatic pipelines and sampling techniques to enable benchmarking of DNA metabarcoding
  Report
- **Best Poster Award @ Hellenic Bioinformatics conference** (2018)
  for *PEMA: a Pipeline for Environmental DNA Metabarcoding Analysis*

Selected presentations

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

- **1st DNAQUA International Conference** (2021)
  PEMA v2: addressing metabarcoding bioinformatics analysis challenges, oral talk - video
- **Federation of European Microbiological Societies - FEMS2020** (2020)
  “Mining literature and -omics (meta)data to associate microorganisms, biological processes and environment types” - video poster
- **PyData Global PyData2020**
  “Geometric and statistical methods in systems biology: the case of metabolic networks”, oral talk - video
- **8th International Barcode of Life Conference** - 2019
  "P.E.M.A.: a pipeline for environmental DNA metabarcoding analysis" (flashtalk)

Participation in proposal writing

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

Contact

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