TÉCNICO LISBOA

Tackling the diversity of the marine microbial rare biosphere: methodological challenges and ecological insights

Francisco Daniel de Oliveira Pascoal

Thesis to obtain the Master of Science Degree in

Microbiology

Supervisors: Prof. Rodrigo da Silva Costa and
Prof. Catarina Maria Pinto Mora de Magalhães

Examination Committee

Chairperson: Prof. Isabel Maria De Sá Correia Leite de Almeida
Supervisor: Prof. Rodrigo da Silva Costa
Members of the Committee: Dra. Tina Keller Costa

November 2019

"*Who can explain why one species ranges widely and is very numerous, and why another allied species has a narrow range and is rare? Yet these relations are of the highest importance, for they determine the present welfare, and, as I believe, the future success and modification of every inhabitant of this world.*"

Darwin C., *On the Origin of Species*, 1859

Preface

The work presented in this thesis was performed at the Biological Sciences Research Group, Department of Bioengineering, Instituto Superior Técnico (Lisbon, Portugal) and at the Interdisciplinary Center of Marine and Environmental Research, University of Porto (Porto, Portugal), during the period of September 2018 to October 2019, under the supervision of Prof. Rodrigo da Silva Costa and Prof. Catarina Maria Pinto Mora de Magalhães. The work was financially supported by the Portuguese Science and Technology Foundation (FCT) through a grant to Francisco Pascoal on behalf of NITROLIMIT project (PTDC/CTS-AMB/30997/2017). This research was also partially supported by the "Programa Operacional Regional de Lisboa" (Project N. 007317) and the Strategic Funding UID/Multi/04423/2013 and UID/BIO/04565/2013 through national funds provided by FCT and European Regional Development Fund (ERDF), in the framework of the "PT2020" program.

This work contributed for a scientific poster presentation at Dalhousie University, for the Marine Microbes mini symposium, entitled "Thinking marine microbial rarity". This work also resulted in the submission of a scientific article for publication, named "The link between the ecology of the prokaryotic rare biosphere and its biotechnological potential"

Declaration

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

This thesis project was possible due to the financing from FCT, through the project PTDC/CTS-AMB/30997/2017 and by the strategic funding UID/Multi/04423/2013 and UID/BIO/04565/2013. The work was also supported by “Programa Operacional Regional de Lisboa” (Project N. 007317). The space and scientific environment provided by both the Instituto Superior Técnico and the Interdisciplinary Center for Marine and Environmental Research, are also acknowledge.

The scientific knowledge produced in this work was possible thanks to the expertise and experience from the senior researchers that supervised this Master of Science thesis, namely, by the Prof. Rodrigo da Silva Costa, from the Biological Sciences Research Group, Department of Bioengineering, Instituto Superior Técnico, University of Lisbon and by the Prof. Catarina Maria Pinto Mora de Magalhães, from the EcoBiotech Research group, at the Interdisciplinary Center of Marine and Environmental Research, University of Porto.

Some reflections in this work were only possible with the exchange of previously unpublished data and opinions with the team from the Microbial Ecology cluster, Genomics Research in Ecology and Evolution in Nature, of the Groningen Institute for Evolutionary Life Sciences, University of Groningen. Specifically, through personal communications with Dra. Xiu Jia, Dr. Francisco Dini-Andreote and Dra. Joana Falcão-Sales. The opinions of Dr. António Sousa and Dra. Maria Paola were also important, both from the EcoBiotech Research group, at the Interdisciplinary Center of Marine and Environmental Research, University of Porto. Prof. Luís Torgo is acknowledge for support in R programming, from the Faculty of Computer Science, Dalhousie University.

The personal acknowledgments go to both Prof. Catarina Maria Pinto Mora de Magalhães and Prof. Rodrigo da Silva Costa, for setting me in the microbial rare biosphere research field, allowing me to work in a friendly, comfortable and scientific environment, always available to help. A final personal acknowledgment to Prof. António Manuel Nunes dos Santos, retired member of the Department of Social Sciences, from the NOVA School of Science and Technology, for setting me in the path to science and for teaching me how science history and philosophy are necessary to understand scientific knowledge.

Abstract

The microbial rare biosphere represents the bulk of microbial diversity in virtually all environments. This is historically recognized in general biology and confirmed in microbiology due to the emergence of high throughput sequencing of the small subunit of the ribosome gene. The number of studies on the microbial rare biosphere has been growing every year, allowing for the recognition that, despite their low abundance, they contribute to ecosystem functioning and are important to understanding microbial community assembly. Currently there is no coherent, unifying definition of microbial rarity, as the concepts in use vary greatly and commonly lack biological meaning. To approach this hurdle from a statistical standpoint, the Multivariate Cutoff Level Analysis (MultiCoLA) algorithm was recently proposed for determining abundance thresholds from where microbial rarity could be delineated across distinct microbiomes with their own unique community structures. This algorithm was tested in the present study, where it generated coherent results across independent marine datasets, but it is not able to give support to a non-subjective definition of microbial rarity. Nevertheless, using the rare prokaryotic communities identified with the later method, it was possible to explore how different metagenomic strategies and seawater sampling methodologies affect the structure of the so-defined marine microbial rare biosphere. Ecological insights from the Arctic Ocean data and from the *Spongia officinalis* (marine sponge) microbiome data corroborate existing knowledge on the marine prokaryotic rare community assembly processes. Furthermore, this study integrates both stochastic and deterministic mechanisms in the process of marine prokaryotic rare biosphere assembly, with water masses and host-associated relationships playing key roles. Finally, this work provides methodological guidelines for optimal sampling of the seawater rare biosphere.

Keywords: community assembly; microbial dark matter; microbial ecology; rarity definition; seawater sampling; sponge-associated microbiome.

Resumo

A biosfera rara microbiana representa a maior parte da diversidade microbiana em praticamente todos os ambientes. Isso é historicamente reconhecido na biologia geral e confirmado na microbiologia devido ao surgimento da sequenciação em massa do gene da subunidade pequena do ribossoma. O número de estudos sobre a biosfera rara microbiana vem crescendo a cada ano, permitindo o reconhecimento de que, apesar de sua baixa abundância, contribui para o funcionamento do ecossistema e é importante para entender a estrutura das comunidades microbianas. Até hoje não existe uma definição coerente de raridade microbiana e as definições utilizadas na literatura são variadas e sem significado biológico. Para resolver isso, o algoritmo de Análise Multivariada de Limites (MultiCoLA) foi proposto recentemente. Este trabalho testou esse algoritmo, constatando que fornece resultados coerentes quando comparando dados de diferentes amostragens do ambiente marinho, mas não fornece uma definição não subjetiva da biosfera rara microbiana. Ainda assim, usando as comunidades procarióticas raras descritas com esse método, foi possível explorar como diferentes estratégias de metagenómica e metodologias de amostragem de água do mar influenciam a descrição da biosfera rara procariota marinha. Informações ecológicas dos dados do oceano Ártico e dos dados do microbioma de *Spongia officinalis* (esponja marinha), permitiram corroborar o conhecimento existente sobre os processos de formação de comunidades procarióticas raras marinhas. Além disso, este trabalho sugere como integrar mecanismos estocásticos e determinísticos na estruturação das comunidades procarióticas raras no ambiente marinho, com massas de água e simbiose desempenhando papéis-chave. Finalmente, este trabalho fornece sugestões metodológicas para a amostragem da biosfera rara procariótica na água do mar.

Palavra-chave: Montagem de comunidades; matéria negra microbiana; ecologia microbiana; definição de raridade; amostragem de água do mar; microbioma associado a esponjas marinhas

List of Contents

List of Figures

.

List of Tables

List of Abbreviations

AW. Atlantic Water

bp. Base pair

cDNA. complementary DNA

CPR. Candidate Phyla Radiation

CRT. Conditionally Rare Taxa

DCA. Detrended Correspondence Analysis

DOC. Dissolved Organic Carbon

EMOSE. EuroMarine Open Science Exploration

HGT. Horizontal Gene Transfer

HTS. High Throughput Sequencing

KtW. Killing the Winner hypothesis

MAG. Metagenome Assembled Genome

MAW. Modified Atlantic Water

mcrA. Methyl coenzyme-M reductase

MetaB 18S. Amplicon sequencing for the V9 region of 18S rRNA gene

MetaB 16S large. Amplicon sequencing for the V4-V5 region of 16S rRNA gene, with library sizing for sequences with more than 600bp

MetaB 16S nS. Amplicon sequencing for the V4-V5 region of 16S rRNA gene, without library sizing

MetaB 16S small. Amplicon sequencing for the V4-V5 region of 16S rRNA gene, with library sizing for sequences with less than 600bp

MetaG 16S. Total-Community DNA shotgun sequencing data for 16S rRNA gene sequences

MultiCoLA. Multivariate Cutoffs Level Analysis

n. Number of individuals

NA. Not applicable

NICE. Norwegian Young Sea Ice Expedition

OTU. Operational Taxonomical Unit

PAH. Polycyclic Aromatic Hydrocarbons

PCA. Principal Components Analysis

PCR. Polymerase Chain Reaction

PSW. Polar Surface Water

PRT. Permanently Rare Taxa

RAC. Rank Abundance Curve

rDNA. Ribosomal DNA sequence

rRNA. Ribosomal RNA

SAD. Species Abundance Distribution

Si. Number of Species with i'th individuals

SSU. Small Subunit

St. Total number of species

t. Rarity threshold

TC-DNA. Total Community DNA

V4-V5. SSU rRNA sequence from the hypervariable region 4 to 5

V6. SSU rRNA sequence from the hypervariable region 6

V9. SSU rRNA sequence from the hypervariable region 9

wPSW. Warm Polar Surface Water

1. Introduction

1.1 Overview

It is now known that the Species Abundance Distribution (SAD) of natural microbial communities follows a universal pattern consisting of a few high abundance taxa and a very large number of low abundance taxa (1–5). The low abundance microbial taxa are known as the microbial rare biosphere (6). It follows that the microbial rare biosphere represents a small number of cells and, counterintuitively, also represents the bulk of microbial diversity in a given environment or sample (3). Since the pioneering study by Sogin et al. (6), the biological processes explaining microbial rarity have been subjected to intense research. It is now accepted that the microbial rare biosphere is heterogeneous from the perspective of (i) spatial and temporal dynamics (7,8), (ii) activity (9–11), (iii) ecological role (12) and (iv) rarity mechanisms (3,8,9,13). From the first perspective (i), different environmental conditions are expected to produce biological abundance variations across time and spatial scales and other parameters such as habitat type and heterogeneity. The possible variations in biological abundances divide the concept of rare biosphere in different types of rarity (8). The currently most accepted division is (8,13): Permanently rare taxa (PRT); Permanently rare taxa, with variation; Conditionally rare taxa (CRT) and Transiently rare taxa. From the activity point of view (ii), it has been hypothesized that rare microbes can be dying cells, dead cells (3) or inactive, but viable, dormant cells (10). Notwithstanding, it was found that a significant component of the rare biosphere is not only active, but can also be more active than their abundant neighbors (11,14–16). From the ecological role view (iii), despite their low numbers, their high diversity represents a ‘genomic reservoir’ or ‘pool of diversity’ (3,4,8,12,17–20). This diversity can work as a ‘seed bank’ (21), assuming the existence of CRT that can respond to specific environmental alterations and/or stressors. The latter mechanism works by clonal amplification of a previously rare, but viable, cell (7). Another suggested process assumes the existence of low abundance cells with disproportionately high activity for a specific set of metabolic functions, with consequences for the overall community (9,22,23). Finally, it was recently proposed that rare, but active, bacteria can also use Horizontal Gene Transfer (HGT) to transfer useful genes to other bacteria, in response to specific stressors (24). This hypothesis is mainly supported by the finding of mobile plasmids with specific functional genes in rare biosphere members, with infrequently used genes in the sampled environment (24). From the perspective of rarity mechanisms (iv), a rare microbe can be unfit to grow in a specific environment, or can have optimal growth conditions, but be outcompeted or predated (1,3). From the Killing the Winner (KtW) theory perspective (25), low abundance can work as a defensive strategy to protect bacteria against bacteriophage attack, because the probability of a virus to find a rare host is lower than for an abundant host. Despite previous contradicting evidence emerging from host-associated communities where dense microbiomes are found to possess lower viral abundances, due to suppressed lysis, favoring temperate dynamics (bacteriophages that switch between dormant and productive phases) (26). The rare microbes can also have intrinsic metabolic limitations preventing them from growing abundant, independently of the conditions (1,3). Recently, the molecular mechanisms behind high activity, in optimal conditions, with slow to null growth, are starting to be addressed (9).

The microbial rarity patterns are found in all domains of life, with more studies focusing on prokaryotes (6,14,15,17,18,20,23,27–41) than eukaryotes (11,16,42–46). They are also found within functional groups (47). Some rare taxa are phylogenetically close to the abundant members, whereas others are phylogenetically distant (40). The microbial rare biosphere includes a considerable amount of genetic novelty (4,38,40) that has been related with unknown and unclassified taxa (8). To date, the study of the microbial rare biosphere has been mostly dependent of High Throughput Sequencing (HTS) strategies, mostly based on 16S and/or 18S amplicon sequencing, e.g. Sogin et al. (6). Those methods are necessary to have a general view on the entire rare community within a given sample, even though other methods, such as Total-Community DNA (TC-DNA) shotgun sequencing, should be used for a deeper understanding of the metabolism and full extent of the rare biosphere (9,12,48). Understanding the impact of different methodologies on metagenomic results and the resulting bias in relative abundance estimations (49) will be a key factor when studying the microbial rare biosphere.

This work addresses rarity from earlier studies in general biology and how different discoveries allowed for the finding of a virtually ubiquitous microbial rare biosphere. It covers the current knowledge of microbial rarity in ecology and then focuses in the marine environment, exploring missing gaps in current methodologies, both on conceptual and technical perspectives. Finally, this work addresses questions regarding the ecology and assembly patterns of the marine prokaryotic rare biosphere using host-associated and planktonic microbial communities as model systems, presenting insights into host-driven selection and spatiotemporal dynamics of low abundance communities in marine settings, and into the stochastic and deterministic forces underlying their assembly mechanisms.

1.2 Historical perspective of biological rarity

Rarity is a general concept that has been previously studied in general ecology, but only recently it was successfully applied to the microbial world (6). Naturalists, as early as in the XIX century, observed that different species have different abundance distributions, with most species being rare. In fact, Darwin stated that “rarity is the attribute of a vast number of species of all classes, in all countries” (50). The prevalence of rarity was merely based on observations and not on statistical treatment of data.

In the early XX century, some studies focused on the distribution of the frequency of abundances of different species, resulting in several SAD and similar arithmetic curves for different datasets. For all types of animals and plants tested, the SAD curves always had the same hollow-like shape (51). Figure 1 illustrates the typical shape of a SAD curve. In the context of current microbiology, the equivalent to ‘species’ in Figure 1 would be Operational Taxonomical Units (OTUs), defined in current metagenomics as a cluster of sequences with 97% (or more) similarity, with the number of reads assigned to each OTU used to calculate abundances, as proxy to individuals in macroecological communities.

For instance, Corbet et al. (52) described, in 1942, a dataset of butterflies across different geographic points. By plotting the SAD, they found that, for the rare species, the relationship between the number of species (Si) with i’th number of individuals (n), can be described according to equation 1:

$$Si = \frac{c}{n^m}$$

Equation 1

This relationship originated from the logarithm of the line equation, with C as a constant dependent on the dataset and m as a constant from the slope of the line. Because there are more rare species than abundant species (illustrated in Figure 1), it follows that the slope is negative, resulting in equation 1.

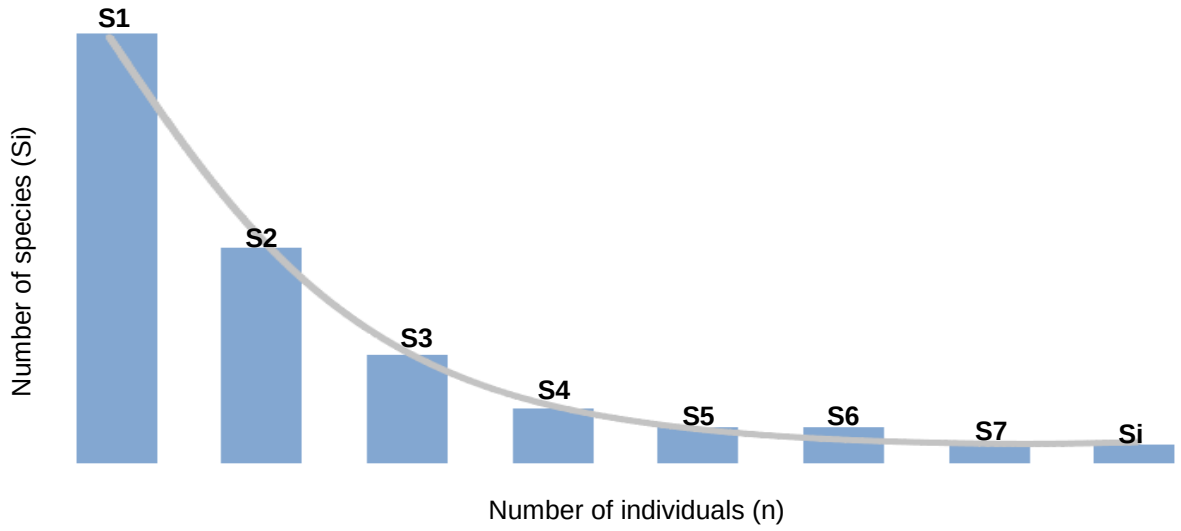

Figure 1. Hypothetical species abundance distribution, illustrating the typical hollow shape curve. Blue histograms for the number of species with i't individuals and grey line representing the typical hollow shape curve.

From this dataset of butterflies, Fisher et al. (53) tried to establish a relationship between the number of individuals and the number of species within the sampling universe (equivalent to all samples). Where the total number of species in the universe of samples (Total Species, St), would be given by equation 2, resembling the harmonic series:

$$St = C \sum_{n=1}^{n=+\infty} \frac{1}{n}$$

Equation 2

Naturally, it is impossible, in biological systems, to have infinite species or an infinite number of individuals, thus, equation 2 has to be constrained (53). Preston (54) proved that equation 2 did not fit all data, because it would result in an even distribution of species, suggesting the Poisson distribution instead of the harmonic series. These first studies were followed by many others, with the main objective of designing a statistical model to explain the shape of the arithmetic curve of the SAD plot for the entire range of abundances observed in any community. Briefly, the first models were based on the lognormal distribution (54), where rare species have high proportions; the logseries distribution (53), with a lower proportion of rare species; the geometrical model, by Motomura, cited in (51), where communities are uneven in terms of abundance of different species; and the broken stick model (55), with even

abundances across communities. From the 1970's, new models were produced to explain and predict SADs, in different contexts and at different taxonomical levels. Overall, different datasets fitted for different models and there was more effort on making new models, than on disproving and adapting previous models. For a comprehensive review, see McGill et al. (51).Nevertheless, those first studies were the first to reveal the widespread existence of highly diverse communities of rare species using an statistical analysis framework (52,53). They allowed the establishment of definitions that are still used today, such as the usage of distinguishing 'abundance' terminologies (54). Abundance can be 'global abundance', if it refers to the total number of individuals in a given universe (set of samples) or 'local abundance', if it refers to the number of individuals found in a specific sample. Thus, 'global abundance' is equivalent to the whole dataset, whereas the 'local abundance' is equivalent to specific samples. Finally, abundance is often approached in terms of 'relative abundance', meaning the proportion of individuals of a given species relative to the total number of individuals in a community (54). From now on, unless stated otherwise, abundance is referred to as 'local abundance', that is, the abundance that technically derives from a sample, which can be expressed as absolute values, or proportions ('relative abundance') of species (or OTUs, in the context of this work) in a given community.

After the first evidences in support of the highly diverse rare biosphere, early works on rarity focused solely on plants and animals (56). Rabinowitz et al. (57), described rarity from the perspective of geography, habitat and local abundance. By integrating the data this authors defined seven types of rarity, using results acquired from plant communities (57). Those findings are important today, as they form the conceptual basis on which the definition of rarity among microorganisms can be built (13). Furthermore, it was found that as body size was inversely correlated with relative abundance for most animals, the capacity of asexual reproduction and maintenance of small body sizes could be an advantage for rarity, in animals (58). By extrapolation it could be inferred that rarity would be advantageous for microbial species (59). As the SAD curve can be applied to any group of living organisms, including microorganisms, it could be predicted that the SAD curve would display the same behavior when microbial communities could be effectively sampled at appropriate scales. New methods based on the massive sequencing of the Small Subunit (SSU) of the rRNA gene (6,60), addressed thoroughly in the following sections, enabled appropriate description of natural microbial communities with unprecedented accuracy, opening new avenues to the study of rarity among microorganisms.

1.3 Microbial rare biosphere

1.3.1 First views on the role of the microbial rare biosphere in the ecosystem

With the discovery of the microbial rare biosphere, several questions raised related to its geographical distribution, ecosystem-level functions and processes explaining low abundance (6). Sogin et al. (6) hypothesized the microbial rare biosphere to be a source of genomic innovation, possibly associated with community resilience (the ability to recover after a perturbation).

The realization of the existence of high genetic diversity among low abundance populations in heterogeneous environments highlighted the relevance of better understanding the microbial rare biosphere in terms of structure and function (61). To study the diversity and phylogenetic novelty of the

prokaryotic rare biosphere, Elshahed et al. (40) used near full length sequencing of 16S rRNA gene clones from soil samples, and quantified the phylogenetic distance between the rare and abundant OTUs identified. Both rare taxa phylogenetically close (named non-unique rare biosphere), and distant (named unique rare biosphere) to abundant taxa were found, suggesting much genetic novelty associated with the microbial rare biosphere (40). This genetic novelty concept would later be integrated into the dark biosphere concept (8).

The 'seed bank' hypothesis, proposed for the rare biosphere, provides a framework for the existence of a large number of different low abundance microorganisms in natural biomes, with a possible role in ecosystem functioning (21). Pedrós-Alió (21) also proposed that the microbial rare biosphere should be used to rethink the dictum of "Everything is everywhere, but the environment selects" by Baas Becking (see reference (62) for details). If this dictum holds, the microbial rare biosphere is dispersed across all environments: in environments with unfavorable conditions rare taxa remain rare, but viable, becoming abundant with changing conditions (21). For prokaryotes, growth by clonal amplification would work as an advantage to promote rarity, as singletons would not be dependent on the existence of mates in the same local environment (59), and with low abundances it would also be easier to escape predators (6,21). Using freshwater samples, Szabó et al. (39) tested the seed bank hypothesis by progressively removing the low abundance taxa. For that purpose, they diluted the original samples different times and for each dilution the rare prokaryotic communities were lowered. This is because with each dilution the probability of rare prokaryotes to remain is lower than that of abundant prokaryotes. Those dilutions worked as initial community inoculum for growth media with phenol or humic substances (mixture of bioavailable, but recalcitrant substances). By comparing the composition and functioning of the new prokaryotic communities, it was found that resistance to perturbations decreased with rare prokaryotes loss (39). In addition, it was also found that the most abundant members did not grow with the addition of phenol, explaining why the complete communities (rare and abundant) kept their functions after perturbation. Thus, the low abundance microbial species could confer functional redundancy to the microbial community, meaning that rare species could have overlapping functions with other abundant species, at the same time and space, as opposed to functional complementarity, where each species carries out a specific set of non-overlapping functions (39). Another study using freshwater samples, found that some prokaryotes with relative abundances close to 0.3% contributed to approximately 40% of ammonium uptake, suggesting that some rare prokaryotes can contribute disproportionately to ecosystem functioning (63). Despite the usual consensus around the seed bank theory, Galand et al. (41), in a time series study of the marine rare prokaryotic community of the Arctic ocean, provided evidence against the hypothesis. Their point of view was that the seed bank hypothesis required both a cosmopolitan distribution of the rare communities and the existence of CRT. Instead, they found that rare communities were associated with specific biogeographical patterns and not cosmopolitan distributions. Furthermore, they also did not find evidence for CRT in their samples. They differentiated the rare and abundant communities as influenced by distinct selective pressures, with the marine microbial rare biosphere being subject to water masses. These findings were further supported by another study in the Arctic ocean (27) and in Chinese lake water samples (32).

From the seed bank theory it can also be expected that a significant number of dormant taxa make part of the rare biosphere (10). Dormant taxa are expected to have low abundance and low activity. To understand if the OTUs identified in a sample are active or inactive, one possibility is to compare the ratio of rDNA to rRNA sequences (cDNA). This method assumes that more metabolically active cells will have more ribosomes (10). In this context, Jones and Lennon (10) predicted the existence of both active and inactive rare taxa, based on dormancy models from plants adapted to microbiology, and confirmed the finding with rRNA:rDNA data from two lake samples. Later, Campbell et al. (14) used time series of coastal water samples, determining activity and abundance through time, finding that both activity and abundance oscillate with time, thus contradicting the findings by Galand et al. and Kirchman et al. (27,64). These differences might be due to different time series size (65) or can be intrinsic to the different environments tested. It was also found that some rare taxa can be more active than abundant taxa (10,14), and that some taxa decrease activity with increasing abundance, possibly because they enter in a stationary phase (14) and/or are predated more easily, from the KtW perspective (25). Despite some limitations in this methodology (66,67), the acceptance that rare microbes can be active and change activity trough different conditions, supports the existence of the seed bank theory. Also, the existence of active, but permanently rare taxa was confirmed as well by the identification of *Desulfosporosinus* spp., a rare taxon (0.006% relative abundance) found to have the most influence on sulfate reduction in the microbial communities of peatland soil samples, in a study coupling 16S rRNA amplicon gene sequencing and isotopic labeling of Sulphur (22).

These early findings suggested that the microbial rare biosphere could have different ways of impacting the local ecosystem, inducing new studies on the ecological role of the microbial rare biosphere (4). Sjöstedt et al. (68), tested how the same initial community would respond to different Dissolved Organic Carbon (DOC) concentrations and salinity. Results revealed significant differences in the overall community structure after exposure to different DOC concentrations, meaning that different OTUs were abundant and rare before and after the disturbances. Importantly, despite the differences in community composition, the overall functions of the whole community remained the same, with similar growth yields. A later study, using seawater samples from the Mediterranean Sea, tested the response of microbial communities to phenanthrene, a Polycyclic Aromatic Hydrocarbon (PAH), where rare taxa showed to work as keystone species, supported by *in situ* activity assays, and the microbial community composition of rare taxa worked well with the seed bank theory, with most PAH tolerant bacteria belonging to the prokaryotic rare biosphere (31).

Different ecological roles are associated with different ecological strategies and this aspect is key to understand how the microbial rare biosphere behaves with changing conditions. One of the first studies that considered the existence of different patterns of abundance and activity in the same rare community was by Hugoni et al. (15), distinguishing three types of rarity: ‘local seed bank’, ‘non local seed bank’ and ‘active, but always rare taxa’. This division includes active and inactive taxa, with inactive taxa being considered as members of the seed bank. Within that seed bank, some taxa are specific (and well adapted) to the local environment, whereas others randomly appear in the same local environment. The division in different types of rarity allows to explain why the microbial rare biosphere

can have both biogeographical and cosmopolitan patterns. The apparent contradiction can be understood in the light of the interaction between stochastic and deterministic mechanisms in microbial community assembly, recently reviewed by Zhou and Ning (5). Briefly, stochastic patterns follow the neutral theory, from general ecology, to explain SAD's (69), whereas deterministic patterns follow the niche partitioning theory (70), the latter being the most studied. Gobet et al. (28), using the microbiome of coastal sands across different time points, found that bacterial community turnover over time was highly influenced by a permanently rare biosphere, as a consequence of deterministic patterns, i.e. as a response to specific environmental variables. Caporaso et al. (71), while comparing a time series of 6 years of seawater samples, added that the deterministic factors are responsible for a core rare biosphere, which includes PRT and CRT, as they found a common group of rare microbes specifically adapted to their local environment. A different study used high sampling frequency (384 seawater samples), to understand what taxa were frequently found (present in most samples) or infrequently found (present in a few samples) (29). While no spatiotemporal patterns associated with infrequent rare taxa were observed, such patterns were found for frequent rare taxa, providing evidence for the coexistence of both stochastic and deterministic forces in shaping the structure of low abundance microbial communities. The results from Vergin et al. (29) were in agreement with the findings of Caporaso et al. (71), since the 'core microbiome' in the latter study (71) is the equivalent to the 'frequently rare taxa' of the former (29). Congruently, Ai et al. (72) established a model simulation, where community dynamics of rare taxa were better understood when combining deterministic and stochastic processes. The transition from a descriptive to a modeling approach requires the disentanglement of stochastic from deterministic processes in the rare biosphere (5,13,29).

In contrast with the ever-growing literature on rare prokaryotes, there are few studies on the microbial eukaryotic rare biosphere. These were firstly addressed in the hypothesis article by Caron and Countway (1) and verified experimentally in later studies (44,45,73). In marine environments, there was no evidence for stochastic processes, as the distribution of different rare taxa was significantly correlated with biogeography (16). The patterns of rare and abundant protists are similar, despite functional differences, indicating functional complementarity (16). A later study, on the protist rare biosphere in freshwater environments, found a permanently rare, but active, protist community (11). Also finding that rare microbial eukaryotes can be more active than the abundant one, as was found for rare prokaryotes (14). Weise et al. (42), focusing on rare ciliates, a subset of the microbial eukaryotic rare biosphere, suggested that it is important to differentiate between effective and non-effective dispersal. Furthermore, they suggested that high dispersal rates would explain the insurgence of 'accidentally rare taxa' (or transiently rare taxa), whereas other OTUs could be temporarily rare and later follow the abundant species patterns (42).

Despite the different approaches to explain rarity patterns within the microbial rare biosphere (15,71), it was important to confirm the existence of CRT to support the seed bank hypothesis. The work by Shade et al. (7) compared microbial communities of 9 different ecosystems (air, ocean water, lake water, stream, human skin, human tongue, adult human gut, infant human gut, wastewater) across a long-time series. They described CRT in all ecosystems studied and found that those CRT contributed

significantly to overall community structure. Across a wide range of different ecosystems, they found that the CRT communities corresponded from 1% to 28% of the rare taxa and could account for up to 98% of community variation (7). The reason why some studies indicate that the rare biosphere is permanently rare, is because they might not have enough sequencing power of the used marker gene and/or a sufficiently long-time series, as it has been shown that the identification of CRT is highly dependent on the temporal scale (7,65). The relevance of understanding CRT from the perspectives of stochastic and deterministic processes has been stressed in many studies (29,65,71,72). In fact, CRT can be associated with deterministic processes because one selective pressure (condition) induces growth or death. However, the existence of one CRT in a specific site can also result from random events of dispersal (stochastic factor). Hence, it is important to distinguish between resident taxa that are persistently found in that site from transient taxa that randomly appear on that same site (65). From this perspective, the persistent rare biosphere would be relative to the core microbial community, meaning the taxa specifically adapted to that site (7,65). In a study of soil CRT (33), a correlation between CRT and pulses of ecosystem activity was found, and CRT were considered not to be constricted to the seed bank, as they can also be active while rare (33). A further study by Baltar et al., (34) tested the microbial community response to nutrient enrichment and lowered pH through different seasons. It was found that the rare taxa were not always associated with the response to environmental perturbation. They disagree with previous studies possibly due to the definition of rarity used, where 0.1% to 1% relative abundance was considered common. While most studies use rarity thresholds around this range, by converting Baltar et al., (34) definition to a more common threshold, they would have close to 50% of CRT. With the establishment of the existence of CRT in a range of 1% to 28% relative abundance (7), and other findings regarding the spatiotemporal dynamics of the rare biosphere abundance (15,28,29,71), Lynch and Neufeld (8) established a conceptual framework to divide the different types of rarity: ‘permanently rare taxa’, ‘permanently rare taxa, with variation’, ‘conditionally rare taxa’ and ‘transiently rare taxa’ (8), later adapted to fit community assembly theory (13). The relationship between the different types of rarity and the community assembly model is explored in section 1.4.1 of this thesis.

1.3.2 Current view on the ecological role of the microbial rare biosphere

Current studies are focusing on the ecological role of the rare biosphere at different levels, such as response to perturbations (2,20,24,74), host-symbiont interaction (43,75–79) and biogeochemical functions (36,80). In general, they support previous hypotheses. Relevant findings include the study by Kalenitchenko et al. (23), where singletons found on HTS studies were found to also have ecological roles. There is also the perspective of functional groups, as approached in a study by Yang et al. (47), which showed that methanogenic taxa could be permanently rare. Interestingly, the genes for methane reduction found were present in plasmids known to be mobile, thus leading to the hypothesis that permanent rare taxa can confer functional redundancy to the ecosystem through HGT. Wang et al. (24) also found evidence for this hypothesis. Thus, the microbial rare biosphere is potentially linked with mechanisms of conditional clonal growth, disproportionately high activity and/or transfer of useful genes, stored in its genomic pool. Some examples of how these mechanisms contribute to the ecosystem are

provided in this section, where the ecological role of the microbial rare biosphere is divided according with the following topics: biogeochemical cycles, community assembly and host associated interactions.

1.3.2.1 Microbial rare biosphere and biogeochemical cycles

From the biogeochemical point of view, it has been shown that rare taxa can contribute to specific pathways, despite their low abundance, due to disproportionally high activity and/or functional redundancy (1,9,22,23,36,63,80–82), supported by the finding that rare taxa can be more active than abundant taxa (9,10,14,15,83,84).

For the sulfur cycle, it was found that the genus *Desulfosporosinus* significantly contributed to sulfate reduction, despite having a relative abundance of 0.006% (22). Another study by Hausman et al. (80) tested the influence of sulfate enrichment on sulfate reduction reactions, and found that the rare genus *Desulfosporosinus* remained growing slowly, despite increasing activity as measured by rRNA:rDNA ratios, confirmed by qPCR. These findings show that activity is not necessarily related with growth and that rare taxa might have strategies to maintain high activity of some metabolic processes without being abundant (80). Those mechanisms were recently studied at the molecular level (9). The hypothesis that low abundant microorganisms can influence the sulfur cycle at the ecosystem level was further supported by Kalenitchenko et al. (23), where it was demonstrated that singletons in HTS datasets, with relative abundances equivalent to 0.00000002% and thus termed ‘ultrarare’, were active and responded to environmental shifts, contributing to the sulfur biogeochemical cycle (23). Another study regarding the sulfur cycle (85), using amplicon sequencing of (bi) sulfite reductase genes, found that a significant component of sulfate reducers belonged to unclassified taxa, thus representing genetic novelty. Although this study did not focus on the prokaryotic rare biosphere, it found that most of the genetic novelty was associated with the core microbiome that represented taxa with relative abundances of around 1%. However, most studies approaching the rare biosphere use rarity thresholds lower than 1% to diagnose low abundance populations.

Regarding the methane cycle, rare aerobic methane oxidizing bacteria were found to be responsible for the whole methane consumption in the microbial community analyzed (81). That study combined methane isotopic labeling and metagenomics of riparian flood samples (81). A further study based on archaeal methane production, using methyl coenzyme-M reductase (mcrA) as gene marker, found that some methanogens were CRT (47). Also, some methanogens were found to be permanently rare, and since the mcrA genes of rare methanogens were associated with mobile genetic elements, the hypothesis has been raised that permanently rare methanogens contribute to the methane production through HGT processes (47). These results represent a new perspective on how the genetic pool of rare species can contribute to the overall community, in accordance with Wang et al. (24). Some evidence on the role of rare prokaryotes in the nitrogen cycling was also found in studies not focusing on the microbial rare biosphere. Griffiths et al. (86) found a relation with diversity loss and nitrification, in a dilution to extinction experiment. In addition, it was demonstrated that low abundance prokaryotes contributed to up to 40% ammonium consumption, contributing to nitrogen uptake (63). The loss of denitrification processes, where nitrate is reduced to dinitrogen, have been also associated with diversity loss, possibly related with rare prokaryotes (87). Finally, using both metagenomics and

metatranscriptomics, the view that the rare biosphere contributes to functional redundancy and that nitrogen fixation can be performed by rare taxa was also supported (82). In spite of all the above-mentioned activities, it has been suggested that the carbon cycle would be hardly influenced by the rare biosphere, as it is mostly related with biomass gain and loss, and the rare biosphere, by definition, will not have a significant contribution to biomass (3).

1.3.2.2 Microbial rare biosphere and community assembly

The role of the rare biosphere can also be considered from the point of view of community assembly, usually reflecting the resistance and resilience of microbial communities when facing perturbations. A perturbation can be defined as a disturbance strong enough to get a response from the microbial community (29). In this case, resistance is the strength of the community against the perturbation and resilience is the ability to recover after the perturbation. The work by Fernandez-Gonzalez et al. (36) tested the influence of different energy inputs on prokaryotic communities. That study used methane as an input of energy and measured methane oxidation. They incubated the same original inoculum with two different strategies, one with constant methane input (control group) and another set with cyclical methane input, for a long-time period (36). The structure of the microbial community changed, but, despite the differences in energy inputs, the methane oxidation rate and growth rates become similar with time. Those results show, on one side, that the prokaryotic rare biosphere can actively respond to different methane contexts, but a more general view shows that prokaryotic communities remain stable with different energy inputs over time, because of the functional redundancy provided by the prokaryotic seed bank (36).

Other studies have shown the ability of rare prokaryotes to degrade pollutants (20,31,74), also reflecting the resistance provided by the low abundance taxa. One of these studies (31) tested the effect of phenanthrene (one type of PAH) high concentrations on a microbial community, finding that rare taxa are able to degrade phenanthrene and grow to become abundant, fitting well with the seed bank theory (31). Another study tested the prokaryotic response to an oil spill in soil, showing that some hydrocarbon degraders are CRT (20). Rare eukaryotes were also shown to be relevant for the recovery of microbial communities after oil spills in soil (46). In soil microcosms it was found that alkane degraders, corresponding to less than 0.1% relative abundance, were main contributors for the metabolism of long chain alkanes (88). This is further supported by Wang et al. (24), where they tested the response of a microbial community, isolated from a lake, to a group of organic compounds, such as 2,4-dichlorophenoxyacetic acid, 4-nitrofenil and caffeine. Again, those compounds were degraded by CRT and the genes associated with 2,4-dichlorophenoxyacetic acid degradation were present in plasmids, known to be mobile, thus indicating HGT as a strategy for ecosystem resistance and resilience (24). This result is particularly important in the context of perturbation response, because 2,4-dichlorphenoxiacetic acid was not found in the lake, meaning that the rare biosphere can have (apparently) unnecessary functions for long periods of time and, if needed, can transfer by HGT those functions (24). In agreement, it was also reported that, in denitrifying sludge communities, the complete degradation of cholesterol, through the 2,3-seco pathway, can be carried out by the microbial rare

biosphere (74). Other studies on microbial responses to pollutant degradation indirectly suggest, as well, that the rare biosphere has a role in pollutant degradation (89–91).

1.3.2.3 Microbial rare biosphere and host-associated interactions

The rare biosphere can work as a source for horizontal transfer of symbionts to their hosts (acquisition of microbial symbionts from the surrounding environment), as reported for rhizosphere microbiomes where most symbionts were recruited from the rare soil biosphere (77). This pattern was also found for the amphibian skin (92) and marine sponge microbiomes (75,93,94). Additionally, the rare biosphere can contribute to the overall symbiotic community and host health. Some studies addressing plant-microbiome interactions reported on a relationship between the rare biosphere and plant health (76,77,95). In works testing plants with or without rare microbes in the soil, the plants presented less defenses, despite producing more nutrient content, when low abundance microorganisms were removed, possibly meaning that the rare biosphere helps stimulating plant immune response (76,95). The prokaryotic rare biosphere has also been described as a component of the coral microbiome (96) and, recently, rare dinoflagellates were associated with host-symbiont resilience in corals (43). Some studies indicate a role of the rare biosphere in the human microbiota, for example, Horz et al. (97) suggested that archaea in the human microbiota are overlooked because of their low abundance, but with an increasing recognition that archaea might be more important than previously thought. A study of the lung microbiota of Cystic Fibrosis patients showed that many pathogens associated with Cystic Fibrosis infections were present in low abundances (98). Regarding the mouth microbiota, it was found that low abundance species in biofilms contribute to the inflammatory process of periodontitis (99).

1.4 Defining the microbial rare biosphere

The definition of microbial rarity is subjective, and it depends on the methodology used. The most general definition is that the microbial rare biosphere corresponds to the OTUs that are included in the ‘long tail’ of the Rank Abundance Curve (RAC, Figure 2) (3). The problem is on the exact threshold at which a specific OTU is considered rare, or ‘where is the beginning of the RAC long tail?’ (Figure 2, dashed circle). Most studies use a random threshold of relative abundance per sample, usually 0.1% (9,11,22,24,29,34,35,68,77,80,83,84,100–102) or 0.01% (7,37,38,41,103–107), meaning that OTUs with relative abundances inferior to those arbitrary thresholds are considered rare, in a given sample. Alternatively, the same arbitrary threshold can be expressed in absolute values, e.g. (5). If a threshold of 10 reads per sample is considered, it means that OTUs with less than 10 reads are rare, in a given sample. These thresholds are frequently used, but they can be considered ambiguous and artificial. Recently, it has been proposed another method to categorize the rare biosphere in a more meaningful way (13) based on the use of Multivariate Cutoff Level Analysis (MultiCoLA), developed by Gobet et al. (108), thus “exploring the effect of rarity on community structure” (13). Briefly, this approach compares the communities produced by different thresholds, using absolute abundance thresholds (number of reads, per sample) (13,108). The algorithm produces a new Table for each threshold (truncated Table), then each truncated Table is compared with the original community, using correlation values from 0 to 1. For that purpose, there is the non-parametric Spearman rho correlation coefficient (109) and the

Procrustes correlation method (110). From the perspective of Jia et al. (13), the rarity threshold is decided based on the abrupt change of correlation, expected to reflect the distinction between two different groups: the abundant and the rare. Figure 3 illustrates the MultiCoLA expected results, from the perspective of the RAC curve from Figure 2.

Figure 2. Hypothetical rank abundance curve, as expected by the species abundance distribution. All different taxa are ordered from the most abundant to the least abundant in the Rank axis, with a plot of their relative abundance. The dashed circle with the question mark illustrates the ambiguity regarding the beginning of the long tail of the curve. Different blue shades are used to illustrate how different taxa in the same curve are sampled according with different methods (less-resolving TC-DNA fingerprinting and cloning-and-sequencing methods compared with high throughput sequencing technologies). The waves in the end illustrate the unknown diversity, elusive to current methods.

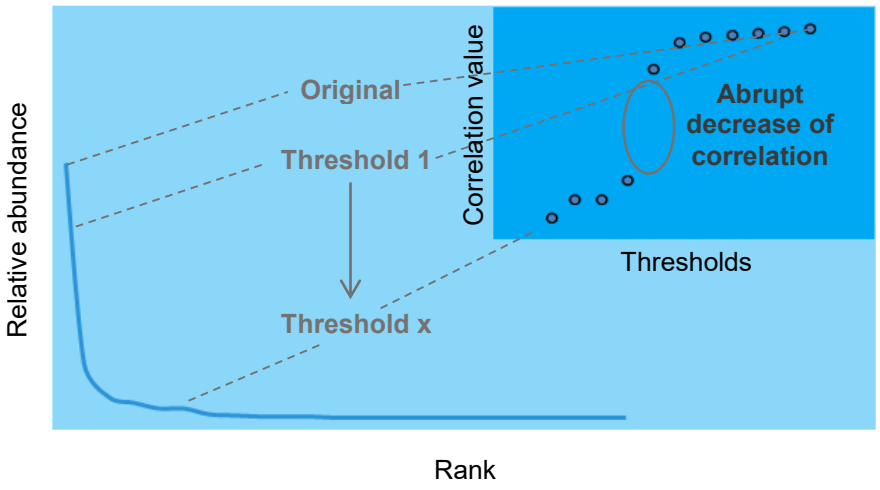

Figure 3. Schematic representation of the MultiCoLA algorithm. Different abundance thresholds are applied resulting in different communities, as illustrated by the rank abundance curve in blue and by the dashed lines in gray. For each new community, a correlation value is plotted that is a measure of the resemblance between the original and the new community, and the correlation values are expected to decrease abruptly when the correct rarity threshold is selected.

The MultiCoLA method is intended to define rarity in a non-arbitrary, context-dependent perspective, despite that, as illustrated in Figure 3, it requires the selection of a threshold based on a set of correlation values. If the resulting data is not as objective as the data expected in Figure 3, where the decrease in correlation is very evident, then the choice of the threshold can be considered subjective. MultiCoLA can be the best option if it succeeds to give an absolute value threshold of reads per sample that is coherent with the magnitude of different, independent datasets. Furthermore, it has the potential to differentiate the abundant and rare communities as separate groups, because the threshold is decided based on correlation values. To our knowledge, the MultiCoLA approach, when used to define the microbial rare biosphere, has not yet been challenged as proof-of-concept in a contextual and interpretation-based manner. This method has been used (and constructed) previously for determining the effect of rare OTUs removal on the total community (28,108), not to define rarity. Because most studies apply thresholds of 0.1% to 0.01% relative abundance per sample, there is the possibility of considering those values as the consensus in the literature, but they, so far, do not rely on any biological justification. Thus, it is relevant to test new approaches.

1.4.1 Defining different types of microbial rarity

Previous studies (addressed in the 1.3.1 section of this work) associated biogeographic patterns of the microbial rare biosphere with deterministic mechanisms and cosmopolitan patterns with stochastic mechanisms, realizing that different mechanisms contribute to the abundance variation of the microbial rare biosphere (29,71,72). Cumulative knowledge of the abundance dynamics of the microbial rare biosphere allowed for the distinction of different types of rarity, based on the review by Lynch and Neufeld (8). The different types of rarity were connected with specific deterministic and stochastic mechanisms, integrating the types of rarity into the community assembly model (13). Deterministic processes work through selective pressures: if the selective pressure corresponds to a set of constant conditions across time/space[1], it is ‘homogenizing selection’; if the selective pressure is due to a varying condition across time/space, it is ‘variable selection’. Conceptually, a rare cell under homogenizing selection does not change abundance across time/space because it is under the same conditions, thus belonging to the permanently rare biosphere. Alternatively, if the same cell is under variable selection, it can grow abundant by changing conditions and vice-versa, thus belonging to the CRT (13). Stochastic processes are based on random events and are harder to predict as they are influenced by drift, random death and growth of cells, selection and dispersal (5). Dispersal can produce different outcomes depending on its limitation, if dispersal is not limited, it dilutes the cells randomly in a process of ‘homogenizing dispersal’. A rare OTU under homogenizing dispersal will remain rare through time/space, but it will receive and loose members randomly, so it will produce permanent rarity, with variation (13). If dispersal is limited and different cells cannot randomly disperse, there will be significant differences across different time/space points. In the latter case, if a cell randomly appears in a given environment, it can be the only one, disappearing without new ones coming from another site, that is associated with transiently rare OTUs. This framework, proposed by Jia et al. (13), is consistent with the community assembly theory from general ecology (111). Therefore, by using community assembly theory to predict the major

[1] For simplicity purposes, only time and space are considered, but the model can be applied to any type of variables.

processes influencing a specific community, the different types of rarity within that community can also be identified.

Another approach is to identify the types of rarity from the perspective of each OTU across samples, e.g. (7), it is less consistent with the community assembly theory, but it is mathematically simpler to determine. For the identification of CRT, for example, the coefficient of bimodality has been used to analyze if each OTU is CRT or not (7). Some difficulties might arise, as it is not always evident the distinction between stochastic and deterministic mechanisms. Consider that the observer point of view represents the known variables. The observer might identify that a group of OTUs are transiently rare across different sampling sites and explain them as a result of stochastic mechanisms (13). Those mechanisms are attributed to dispersal limitation, diversification and/or drift, but the distinction is not linear (5). For instance, if diversification (associated with speciation), is produced due to a set of selective pressures, then it is a deterministic mechanism, but if those selective pressures are unknown (and unpredictable), then the process is identified by the observer as stochastic. Not because it is itself stochastic, but because the deterministic cause of the effect is unknown to the observer (5). The best example of this apparent contradiction is the drift process, defined as the random death and birth of cells - it is random to the observer, because the observer does not know the variables responsible for each birth and death in a given environmental sample. Drift and diversification are difficult to correctly analyse in metagenomic assessment of natural microbial communities (5). This distinction between stochasticity and determinism is easier to solve using the mathematical model of community assembly (13) than trying to identify the types of rarity from the perspective of individual OTUs across samples, as it is biased by the observer's point of view.

1.5 Methodological developments that allowed the study of the microbial rare biosphere

Microbial diversity can be addressed using two major methodological branches: culture-dependent and culture-independent methods. While the former approach may lead to the domestication of rare phylotypes in the laboratory (see Hardoim et al. (79), for an example from sponge symbiotic communities). The latter approach allows the study of the microbial rare biosphere in a comprehensive fashion, enabling the determination and comparison of the taxonomic composition and diversity of both abundant and rare fractions within the community (6).

Culture-dependent methods possess well-known caveats in the assessment of microbial diversity in any given environment, leading Staley and Konopka (112) to coin the term 'great plate count anomaly' to refer to the fact that only a minor fraction of the total microbial communities inhabiting natural ecosystems are cultivable (112). The 'great plate count anomaly' is based on the difference, by several orders of magnitude, of cell counts found in culturing methods (estimated via counting of colony forming units) versus the cell counts found with alternative, cultivation-independent methods such as direct observation with fluorescence (112). For a recent study on the 'great plate count anomaly' and the study of uncultured microorganisms see (79,113). The existence of not (yet) cultivable microorganisms required the assessment of microbial diversity with culture-independent methods. Those methods have

their roots in early works by Woese and Fox (114), where 16S and 18S rRNA gene sequences were used to reorganize the main divisions of life (at the time, prokaryotes and eukaryotes) to three domains of life: Bacteria, Archaea and Eukarya (114). To do so, a phylogenetic signal comparable across all living organisms was necessary. At the time, the ribosome was the only unit known to be universal and to evolve slowly, thus working as a phylogenetic signal (114). At the same time, methods to sequence DNA were being developed, such as chain termination DNA sequencing (115). Early approaches to understand prokaryotic natural communities (116,117) used both the knowledge of 16S rRNA gene sequences as a phylogenetic signal and the ability to sequence DNA. Giovannoni et al. (116), in a study of bacterioplankton in the Sargasso sea, showed that the phylogenetic analysis of 16S rRNA gene sequences, assessed by clone libraries, allows the study of microbial diversity in a culture independent way (116). This way, SSU rRNA DNA sequences became a tool to study natural microbial communities. At the same time, Ward et al. (117) also developed culture-independent methods to access microbial diversity with 16S rRNA gene sequences, where they showed its usefulness to study previously uncultured taxa. The general approach of cloning libraries allowed for a better estimation of the microbial biodiversity than the ones possible with traditional fingerprinting methods or culture-dependent methods (118,119). Indeed, one clone library study with high resolution described a SAD curve that resembled the ones previously discussed, suggesting the existence of a large group of species with very low abundance, from 9 coastal water samples, from different geographical points (120). Some fingerprinting methods such as Desaturating Gradient Gel Electrophoresis of rRNA genes (121) or automated Ribosomal Intergenic Spacer Analysis (122) were useful in describing natural microbial communities (123). Although these methods were powerful in enabling comparative analyses of multiple microbial communities simultaneously, and thus robust statistical assessments, they were limited in the number of different taxa identified, constraining the extent of the corresponding SAD curves (Figure 2), and consequently failing to have a comprehensive view on the microbial rare biosphere (3).

With the availability of HTS, it became possible to sequence up to millions of sequences at the same time, with a read length of 200 base pairs (bp) (124). Soon after, Kysela et al. (60) developed a method for the assessment of phylogenetic diversity based on HTS of the hypervariable region 6 (V6) of the SSU of rDNA sequences. This method was also called tag sequencing, because it used smaller sequences of specific small regions of the SSU rRNA, considered enough to assign taxonomy (60). In this case, the analytical pipelines used did not deliver lists of prokaryotic species, but of OTUs, that technically represent clusters of rRNA gene sequences sharing a given percent homology (usually, the OTU definition threshold is 97% 16S rRNA gene similarity), which are then taxonomically classified and used as proxies for species, in spite of continuous debate around the use of this approach as a proxy to estimate ‘true’ microbial species richness in the environment (125). With tag sequencing of the V6 region of 16S rRNA sequences amplified from DNA extracted from deep ocean samples, Sogin et al. (6) found that most prokaryotic OTUs are rare, highlighting the long tail of the RAC, thus proving the existence (and coining the term) of ‘rare biosphere’ in microbiology. This was possible because of the higher sequencing power of the then emerging pyrosequencing technology, resulting in a more complete view of the true extent of microbial biodiversity in seawater samples (6). Figure 2 illustrates how different

methodologies influence the view of the RAC for the same community, revealing why HTS based methods were necessary to characterize the microbial rare biosphere.

1.6 Marine microbial rare biosphere

The marine microbial rare biosphere follows the same ecological patterns addressed in sections 1.3 and 1.4.1. In fact, most of these studies focus on the analysis of seawater samples (6,7,14–16,27–29,34,35,41,43,71,75,83,84,100,101,104,126–128), with less studies on other environments, such as soil (37,40,77,80,95,103).

In the ocean, the barriers to dispersal are not obvious, so it is tempting to think that microbes, due to their small size, are universally dispersed and simply change their relative abundance in the face of different conditions, as could be supported by the existence of the microbial rare biosphere (21). Rare marine microbes may display biogeographic distributions if they are from the core local microbiome or may be transiently rare displaying a cosmopolitan distribution (16,41). The component of the microbial rare biosphere that has a specific biogeography, in the ocean, has been associated with deterministic selective pressures, such as water masses (16,27,41). Anderson et al. (35) suggested that the abundant microbes are more cosmopolitan in the ocean than the rare ones, as it is easier for abundant taxa to be dispersed. The work also suggested that rare bacteria are more restricted geographically than archaea, as indicated by the cosmopolitan distribution of abundant archaea such as Marine Group I and II (35). The existence of conflicting views on the effect of selection in marine environments might be due to the ecological gradients tested (29,71). For example, Galand et al. (41) only found permanent rarity in the Arctic ocean, whereas Kirchman et al. (27), working on seawater samples from the same ocean, despite also finding that permanent rarity was prevalent, found a small percentage of coexisting CRT. Notwithstanding, permanent and conditional rarity are both linked to deterministic selective pressures (13). Recently, Troussellier et al. (78) proposed the ‘sustaining the rare hypothesis’, arguing that the distribution of rare microbes in the ocean is also connected to interactions with macro organisms, such as animals. Mobile animals, like fish, have a gut microbiota that is constantly being dispersed to the water as they travel long distances (78). On the other side, sessile animals, like marine sponges, work as water filters, accumulating large amounts of microbes and thus working as a source of microorganisms that are rare in the surrounding waters. For instance, Webster et al. (100) found rare taxa in seawater microbial communities that are present in the sponge microbiome as sponge-specific symbionts. This suggests that there are species that need the host associations to live, but can persist in the seawater as viable free-living cells, at low abundances, to be later filtered by other sponges (100). The presence of rare symbionts in seawater was also reported for long geographical distances, maybe explaining why different sponge species can have similar specific symbionts independently of geographic distributions (94).

1.6.1 Marine microbial rare biosphere assessment, methodological aspects

After seawater sampling for the retrieval of cells, the next steps can be ‘culture dependent’ or ‘culture independent’. It is worth noting that marine ecosystems include benthic systems, thus, sediment samples are also important, but this section will focus on the seawater component. For culture-

independent methods, seawater sampling is based on the filtration of water, with the objective of retrieving cells from the total community present in a specific sampled volume, for posterior DNA extraction. Filters are usually of two types, membrane filters, e.g. in (14,16,29,41) or Sterivex filters, e.g. in (15,18,27,41,83,104), with pore sizes of approximately 0.22 μm. Sometimes, pre-filtration steps are used, e.g. in (11,27,101,129), to remove suspended particles and lower eukaryotic contamination, when assessing for marine rare prokaryotic diversity. Also, it is possible to use size fractioned filtration, where different pore sizes are sequentially used, for example, a pore of 20 μm, a pore of 3 μm and a pore of 0.22 μm. In this case, there are three samples, the first expected to have cell sizes superior to 20 μm, the second expected to have cells ranging between 20 μm and 3 μm, and the last expected to have cells ranging from 3 μm to 0.22 μm. After seawater sampling, TC-DNA is extracted from the retrieved cells. Next, the community may be characterized by amplicon sequencing of target genes (e.g. 16S and 18S rRNA gene to generate taxonomic profiles for prokaryotes and eukaryotes, respectively) or by TC-DNA shotgun sequencing (taxonomic and functional profiling of the total community). Intriguingly, a recent study found that the marine planktonic rare biosphere is more sensitive to different DNA extraction protocols than the abundant biosphere (130).

For the past ten years or so, HTS has been mostly performed using 454 Pyrosequencing (124) (nowadays phased out) and Illumina (131) technologies. In 454 Pyrosequencing, genomic DNA is fragmented, fragments attach to beads in an emulsion of water and oil for emulsion Polymerase Chain Reaction (PCR) with a detection chip. DNA polymerase releases pyrophosphate at each nucleotide addition and a chip detects each pyrophosphate released. The first reaction uses one nucleotide, if the nucleotide is added to the sequence, then a signal is detected, after the reaction the remaining nucleotides are washed and another reaction, with another nucleotide, begins. Reactions continue until the final sequence is obtained (124,132). Whereas in Illumina, despite being also based on a polymerization reaction, DNA fragments are attached into a slide. Clusters are made with PCR amplification, for each round, reversibly fluorescently labeled nucleotides are added and the nucleotides not used are washed away. With the end of the round, a laser beam detects which nucleotides are added, then a new round can begin and this process is repeated for the entire sequence (132,133). The resulting sequences are pair ended, meaning they are sequenced from both directions. Most microbial rare biosphere studies published to this date used 454 pyrosequencing (6,16,29,39,86,101). This technology was initially preferred for microbial taxonomy assessments because it provided longer sequences than Illumina, but 454 pyrosequencing has a higher error rate and is known to have more PCR bias (134). The error rate is relevant for microbial rare biosphere studies, since it is important to know if one low abundance read is real or the result of sequencing errors (23,135). Notwithstanding, results from both technologies are valid to study the rare biosphere. The unprocessed sequences or raw reads have to be quality-filtered to guarantee that all analyzed data possess biological meaning, for example, using the MGnify platform (136) or other bioinformatic processing pipelines. For data generated with the Illumina chemistry, paired end reads are merged and tested for quality. Amplicon datasets are used for taxonomic information, shotgun sequencing datasets also identify genes and their potential functions. For taxonomy, SSU rRNA gene sequences are identified and taxonomically classified, according with some database, e.g. the Silva database (137), resulting in the assignment of

OTUs. For functional assessment, functional genes are predicted using a database, e.g. the InterPro database(138). The final output is a list of annotated and unannotated putative coding DNA sequences.

Amplicon sequencing-based methods are vulnerable to PCR bias and to the occurrence of artificial sequences, leading to the discussion on whether the microbial rare biosphere was an artifact. To test that, Neufeld et al. (61) compared commonly-used short sequences of the 16S rRNA gene with near full length sequencing of the 16S rRNA gene, finding perfect matches between the rare short sequences and the full length ones. Hamp et al. (139), showed that primer design has more influence on low abundance prokaryotes, but they also found that differences across primers were not significant. However, the strategy of amplicon sequencing and the strategy of TC-DNA shotgun sequencing, without amplification, results in significant differences (139) as expected by the fact that there are much less rRNA genes in a metagenome dataset than in an amplicon dataset. The PCR step could also lead to overestimations of the rare biosphere, but Huse et al. (135) showed that the rare biosphere identified in such context is real and not a result of artificial overestimations. Also related with PCR, Gonzalez et al. (140) showed that the prokaryotic rare biosphere can be underrepresented in comparison with the abundant biosphere, as it is more difficult to amplify low abundance sequences. On the other hand, they suggest that PCR has no bias in taxonomic identification, and that factors such as DNA quality after extraction are also important (140). It is generally recognized that the sequencing power is important for the study of the microbial rare biosphere (6,83), with conflicting views on the need to improve sequencing power or not (23,83). A recent study developed a model to estimate the bias associated with different methodological steps in metagenomics. From the model it is predicted that the relative abundance of the identified taxa can change dramatically (49). Therefore, the understanding of how methodological options influence the view of relative abundance of the microbial communities is important to have a correct view of the microbial rare biosphere.

Culture-dependent methods are also useful as they are necessary to explore the metabolism of microorganisms and are a prerequisite to classify new species. Studies testing the assessment of the microbial rare biosphere by culturing methods and HTS based methods, proved that some microorganisms randomly retrieved with culturing methods were not found with HTS, meaning that culturing methods can successfully identify undetected rare taxa (48,141–144). It was also confirmed that some traditionally cultured microorganisms belong to the microbial rare biosphere (48,83), as culture media can represent unique conditions, allowing the growth of very rare taxa. ‘Culturomics’ has also shown to be useful to complement metagenomic studies e.g. (145,146).

To answer ecological questions regarding the microbial rare biosphere, it might be necessary to go beyond the diversity assessment and functional prediction encoded in genes, and integrate those methodologies with other ones (12). Such methods include synthetic communities (147), for example, to test the effect of order of arrival in community dynamics. Manipulation of microbial communities, for example, using mesocosm experiments to test the effect of varying conditions in an original community, testing the effect of rare microbes loss (39), or the microbial rare biosphere behavior through environmental changes (2). One interesting approach to understand what is the role of specific rare taxa on biogeochemical cycles, or in other functions, is the use of stable isotope probing, for example to test

which microorganisms contribute the most to sulfur cycling (22). Another promising approach is the use of Metagenome Assembled Genomes (MAG), to have an integrated view of the metabolism of rare and not yet cultivable taxa (148). An example of the integration of multiple methods, including MAGs, is the recent study by Hausman et al. (9), where the molecular mechanisms of activity at low abundance are studied in depth, for *Candidatus* Desulfosporosinus infrequens, despite being uncultivable so far. The mentioned taxon, is responsible for dissimilatory sulfate reduction in peatland soils (80)

1.7 Objectives overview

This work aims at both the methodological challenges associated with the correct assessment of the marine microbial rare biosphere, as well as the study of the ecology of low abundance microbial communities. For the methodological challenges, this work explores the definition of rarity, using the MultiCoLA approach proposed by Jia et al. (13), across different datasets and how different seawater sampling methodological steps influence the view of the prokaryotic rare biosphere. From the ecological perspective, this work explores the different types of rarity in the environment in relation with community assembly theory, in the context of the Arctic Ocean, and in relation with host-symbiont interaction.

2. Methodology

2.1 Datasets description

2.1.1 EuroMarine Open Science Exploration 2017 dataset

The main objective of the EuroMarine Open Science Exploration (EMOSE) 2017 was to study the different methodologies available for metagenomics-based studies of marine microbial communities, using the methods employed by large-scale marine microbial diversity surveys such as Tara Oceans (149), Malaspina (150) and the Ocean Sampling Day (151) by applying different filtration methods, filtered seawater volumes and library preparation strategies. Sequences generated with the EMOSE initiative were processed using MGnify (access number MGYS00001935) and are available at ENA (project PRJEB87662). Metadata information is available at PANGEA (152). Sampling was performed by the EMOSE 2017 team.

The samples were collected during 3 days, but this work only uses the samples collected on the first day to avoid small environmental variations. The sampling point was at latitude 42.486º and longitude 42.492º, the air temperature was 17.5ºC and the water temperature was 15.5ºC, salinity was 38 psu and the depth of sampling was at 3 meters. There were six different groups of volumes: 1L, 2.5L, 10L, 100L, 496L and 1000L. Due to methodological constraints during seawater sampling, within the 1000L group the effectively filtered seawater volume included two samples of 716L and one sample of 776L. Two filtration techniques were used: Whole water filtration (> 0.22 µm) by Sterivex filter units (ref: SVGPB1010) or membrane filter unit (ref: GPWP14250) and Size fractionated filtration (>20 µm, 3-20 µm and 0.22-30 µm) just for the membrane filters. For the group of 496L, there were no replicates for the small and medium fraction and for the 1000L group there were two replicates for the small and medium size fractions, but three replicates for the large size fraction. Different approaches were used to retrieve taxonomical information: TC-DNA shotgun sequencing and SSU rRNA gene amplicon

sequencing, using the hypervariable region 9 (V9) of the 18S rRNA gene, with the 1391F/EukB set of primers, for eukaryotic diversity. For prokaryotic diversity, the hypervariable region 4 to 5 (V4-V5) of the 16S rRNA gene was amplified using the primers 515F-Y/926R (153,154), modified from (155,156), to avoid underestimation of the SAR11 clade and overestimation of Gammaproteobacteria. This set of primers can also amplify 18S rRNA gene sequences at homologous regions, but with amplicons approximately 180 bp longer (153). After PCR, prokaryotic amplicons are expected to have approximately 450 bp, whereas eukaryotic amplicons are expected to have around 600 bp. After library preparation, with addition of overhangs for the sequencing machine, the library size is expected to vary from 450 to 850 bp, with prokaryotic sequences presumably in the range of 450 to 650 bp and eukaryotic sequences on the range of 650 to 850 bp. Three different sequencing strategies were used for the 16S rRNA gene amplicon sequencing, considering the given set of primers: (i) sequencing of the entire library (no sizing, MetaB16S nS) and (ii) sequencing of the two different range values separately (MetaB 16S sizing), for the 450 to 650 bp (MetaB 16S small) and for the 650 to 850 bp (MetaB 16S large) library size range (Table 1). The sequence platform for the TC-DNA shotgun sequencing samples was Illumina HiSeq 4000 and for the amplicon sequencing (for both sets of primers) was Illumina HiSeq 2500. The metagenomic strategies are summarized in Table 1.

Table 1. Overview of the different metagenomic strategies used in the EMOSE 2017 dataset. MetaG is 'shotgun sequencing of TC-DNA', MetaB 18S is 'amplicon sequencing for 18S V9 region of rRNA gene', MetaB 16S nS is 'amplicon sequencing of 16S V4-V5 region of rRNA gene, without sizing', MetaB 16S small is 'amplicon sequencing of 16S V4-V5 region of rRNA gene, with sizing for 400bp' and MetaB 16S large is 'amplicon sequencing of 16S V4-V5 region of rRNA gene, with sizing for 600bp'. Sequencing platform, read length, primers and library size are indicated (if applied).

Metagenomic strategy	Primers (if applied)	Library size (if applied)	Read length
MetaG 16S	NA	NA	150 bp
MetaB 18S	1391F/EukB (157) V9 region of 18S rRNA gene	290 - 300 bp	150 bp
MetaB 16S nS	515F-Y/926R (153) V4-V5 region of 16S rRNA gene (also targets 18S rRNA gene)	450 – 850 bp	250 bp
MetaB 16S small		450 – 650 bp	250 bp
MetaB 16S large		650 – 850 bp	250 bp

2.1.2 *Spongia officinalis* 2014 dataset

The *Spongia officinalis* 2014 dataset resulted from TC-DNA shotgun sequencing from *Spongia officinalis* associated samples (158). Samples of sponge tissue (10g, 4 samples), surrounding seawater (1m away, 2L, 3 samples) and surrounding sediment (1m away, 50g, 3 samples). Sampling was performed during May in 2014, at the coast of Pedra da Greta, Algarve. The water conditions were 18ºC, 8.13 pH and 36.4‰ salinity. Samples of *S. officinalis*, seawater and sediments were collected through SCUBA diving.

The samples were collected at 20m depth following the sampling methodology described by Hardoim et al. (159). Nitrocellulose membranes (0.22 uM) were used for seawater filtration. DNA extraction was performed using UltraClean Soil DNA isolation kit for all samples. For the sponge specimens, DNA was extracted from the inner sponge body according with the methodology from Hardoim et al. (79). Sequencing was performed on an Illumina Hiseq 2500 apparatus. Sequences were processed using the MGnify platform (accession number MGYS00000563) and sequences are available at ENA (Project number PRJEB11585).

2.1.3 Norwegian Young Sea Ice expedition 2015 dataset

The Norwegian young sea Ice Expedition (NICE), performed between February to June 2015, had as an objective to monitor the effect of thinning of Arctic sea ice during winter to spring transition, collecting data for different research fields (160). The research vessel was fixed to the ice and drifted along with the ice in the region north of Svalbard, between the Nansen Basin, the Yermak Plateau and a Transitional region. These regions are surrounded by Yermak and Svalbard branches, representing an inflow of warmer and saltier Atlantic waters during the winter spring transition (161). Nine samples were collected during the transition from winter to spring, in March, April and June, ranging from a darker period to a lighter period and at different depths, from surface (5m) to subsurface (25m and 50m) and mesopelagic (250m) depths (162). Different samples also represent different water masses, namely: Polar Surface Water (PSW), warm Polar Surface Water (PSWw), Atlantic Water (AW) and Modified Atlantic Water (MAW). Seawater sampling was performed using whole water filtration, with Sterivex filter units (0.22 µm). Filtered volumes ranged from 3L to 11L. DNA extraction was performed with PowerWater DNA isolation kit protocol. The same samples were used for TC-DNA shotgun sequencing and SSU rRNA amplicon sequencing. Sampling information is summarized in Table 2.

Table *2.* Summary of the sampling conditions from the NICE 2015 dataset. Samples are numbered from 1 to 9 by order of the date of sampling and depth. The water masses listed are: Polar Surface Water (PSW), warm Polar Surface Water (PSWw), Modified Atlantic Water (MAW), Atlantic Water (AW). The sampled volume for each sample is also listed.

Sample	Date of sampling	Depth (m)	Ocean region	Water mass	Volume (L)
NICE_1	09/03/2015	5	Nansen Basin	PSW	5.7
NICE_2	09/03/2015	50	Nansen Basin	PSW	3.7
NICE_3	09/03/2015	250	Nansen Basin	MAW	4.5
NICE_4	27/04/2015	5	Transitional Region	PSW	11.0
NICE_5	27/04/2015	50	Transitional Region	PSW	11.0
NICE_6	27/04/2015	250	Transitional Region	MAW	9.2
NICE_7	16/06/2015	5	Yermak Plateau	PSW	3.0
NICE_8	16/06/2015	20	Yermak Plateau	PSWw	3.3

NICE_9	16/06/2015	250	Yermak Plateau	AW	4.0

Amplicon sequencing of the 16S rRNA gene was performed with the primer set 545F-Y/926R, previously described in (153,154), for the hypervariable regions V4-V5. For 18S rRNA gene sequences, the set of primers TAReuk454FWI/TAReukREV3 modified for the hypervariable region V4 were used, as described in Stoeck et al. (44). Amplicon sequences were processed on MGnify (accession number MGYS00001922) and are available on the ENA database (project number PRJEB21950). TC-DNA sequences were processed on MGnify (accession number MGYS00001869) and are available on the ENA database (project number PRJEB15043).

2.2 Bioinformatic processing of raw reads, by the MGnify platform

The raw reads from all the datasets (EMOSE 2017, *Spongia officinalis* 2014 and NICE 2015) were submitted to the MGnify platform, to have a standardized processing of raw reads across different datasets, as described in Mitchel (136). Briefly, the steps performed by the MGnify institute were: merging of raw reads with SeqPrep (136), then checked for quality with Trimmomatic (163). Reads from TC-DNA shotgun sequencing were divided in rRNA and non rRNA encoding sequences, using Infernal (164) with the Rfam database (165). For the SSU rRNA gene amplicon sequencing it is not necessary to distinguish from rRNA and non rRNA sequences. The SSU rRNA gene sequence was used for the cluster of OTUs, at 97% identity cutoff, using the program MapSeq (166) with the Silva database (137). All the steps described in the bioinformatic processing of raw reads were performed by the MGnify team.

2.3 Downstream analysis

In this work, downstream analysis is referred to as the handling of data resulting from the bioinformatic processing of raw reads, with the objective of extracting biological meaning from the datasets. In this work, when analyzing TC-DNA shotgun sequencing taxonomical information, unless stated otherwise, the focus was on the prokaryotes alone, excluding eukaryotes. Unless stated otherwise, the downstream analysis was performed in the R statistical environment (167).

2.3.1 Multivariate Cutoff Level Analysis, adapted to define microbial rarity

This work adapted the scripts developed for the MultiCoLA algorithm, available from Gobet et al. (108), based on the conceptual framework by Jia et al. (13), resulting in the scripts available in Annex 1. MultiCoLA can make the analysis ‘sample by sample’ (type = SAM), or for ‘all samples’ (type = ADS) at the same time. In ‘sample by sample’, the different thresholds are applied for each sample individually, whereas for the ‘all samples’ approach, the thresholds are applied at the same time to the all samples, as if they were combined into one. To define a threshold of rarity, ‘sample by sample’ is the best option. With this sample-based analysis, for each dataset it is necessary to calculate the maximum value of reads in all samples, and then select the sample with the lowest maximum. Furthermore, the original script focused on the abundant component (typem = abundant), but in this study it focus on the rare component (typem = rare). The correlation between dissimilarity matrices was measured using the non-parametric Spearman’s coefficient and the Procrustes coefficient. The final output is the number of reads resulting after each threshold is applied and the comparison of different correlation coefficients resulting

after each of those thresholds. In all analyses performed with MultiCoLA, the entire list of OTUs was used ('whole OTUs'), thus including unclassified (taxonomical NAs) OTUs. One important difference between the script used in this work and the original (108) is that the second part, regarding the addition of environmental variables, was not used, as it is not necessary for the definition of rarity.

2.3.2 Alpha diversity

The R package phyloseq (168) was used to calculate alpha diversity, with custom R commands (167). The metrics chosen include total number of OTUs per sample, total number of reads per sample and Shannon index (169), taking therefore the dominance of each OTU into consideration in the estimation of diversity evenness. The same metrics were applied to the total, rare and abundant communities, for each dataset. To estimate p-values for significant changes, One-Way ANOVA was used with R general commands, the script is available in Annex 2.

2.3.3 Multivariate Ordination and Beta diversity

For beta diversity, ordination analysis was used with custom R commands, vegan package (170). To decide whether to utilize linear or unimodal methods, it is necessary to calculate the gradient length, using Detrended Correspondence Analysis (DCA) (171). For linear methods and unconstrained analysis, Principal Components Analysis (PCA) (172) was used. For the *Spongia Officinalis* 2014 dataset, beta diversity was also explored from the perspective of shared and specific OTUs (total, abundant and rare) across the different samples, using Venn diagrams, with VennDiagram R package (173).

2.4 Defining different types of rarity

For the NICE 2015 dataset, this work developed an R function (types.r, in Annex III) to distinguish the different types of rarity across sampling depth and sampling month. The function needs two inputs: a previously defined rarity threshold (t) and a complete OTU Table. This function is adapted for the NICE dataset and not generalized for any dataset of the same type. It compares three points for each OTU, in this context, compares three samples one after the other. If the sum of OTUs=0 in all samples, is labeled as absent. If OTUx>t in all samples, is labeled as abundant. From this point on, if it was not labeled as abundant, it is rare. If the rare OTUx=0 in at least one sample, it is labeled transiently rare. If the rare OTUx>t in at least one sample, it is labeled CRT. If the rare OTUx<t in all samples, it is labeled as permanently rare. If the variation of the permanently rare OTUx is high, it is labeled permanently rare OTU, with variation. The last step was later ignored and merged in the permanently rare label, as there is no correct value of variation to distinguish between the categories. The final output is a list of all OTUs and their respective label (absent, abundant, transiently rare, PRT, permanently rare with variation (later omitted) or CRT).

2.5 Data visualization with Circos

Circos is a software for circular visualization of big data in an esthetic and compact way (174). This work used Circos to have a qualitative view of diversity, using custom Perl commands. For the *Spongia*

Officinalis 2014 dataset, it includes the taxonomy of all OTUs, their respective abundance in all samples and links to visualize the rare OTUs shared across samples. For the NICE 2015 dataset, one Circos was produced for the TC-DNA shotgun sequencing taxonomic profile and another Circos, with the same script, for the 16S rRNA gene amplicon sequencing. The NICE Circos illustrates not only the abundance and full list of OTUs, but also the different types of rarity as defined by the types.r function.

3. Results

3.1 Defining microbial rarity

3.1.1 Testing MultiCoLA on the EMOSE 2017 dataset

The EMOSE 2017 dataset was divided into five metagenomic strategies (Table 1). After applying MultiCoLA, each subset reflected a different threshold of rarity. Notwithstanding, by turning the absolute value of reads used to define rarity into their relative abundance equivalent, the mean value was 0.161%, ranging from 0.047% to 0.514% (Table 3).

Table 3. Prokaryotic and eukaryotic rarity thresholds obtained by MultiCoLA, for the EMOSE 2017 dataset. Rarity thresholds are in absolute abundance (number of reads per sample) and the average of the equivalent relative abundance across samples.

Sub dataset	Absolute abundance threshold	Relative abundance threshold (average per sample)	Number of reads clustered into OTUs (average per sample)	Number of samples
MetaG 16S	6	0.097%	60 014	50
MetaB 18S	197	0.047%	1 492 134	47
MetaB 16S nS	154	0.055%	1 920 794	68
MetaB 16S small	972	0.094%	1 367 111	53
MetaB 16S large	7899	0.514%	1 810 973	53

Higher sequencing powers, in terms of the number of reads obtained from the marker genes assessed, were associated with amplicon sequencing strategies. The resulting rarity thresholds for those approaches were two orders of magnitude superior than that of the TC-DNA shotgun sequencing strategy data, when considering the prokaryotic taxa identified (Table 3). The threshold selected for MetaG 16S was 6 reads per sample, selected from one of the first thresholds below 0.9 correlation value in both Procrustes and non-parametric Spearman's coefficient analyses (Figure 4A). In each sample, that absolute value threshold was on average less than 0.097%. For MetaB 18S, the threshold decided, with similar reasoning, was 197 reads per sample, corresponding to the lowest correlation value in the non-parametric Spearman's correlation coefficient. With the Procrustes correlation, all values were close to 0.9 (Figure 4B). This absolute threshold corresponded to a relative abundance of 0.047% on average. For MetaB 16S nS, the threshold selected was 154 reads per sample, corresponding to 0.85 correlation value in the non-parametric Spearman's correlation coefficient and 0.77 with the Procrustes correlation coefficient (Figure 4C). For MetaB 16S small, the threshold selected was 972 reads per sample, corresponding to 0.9 and 0.69 correlation values for the non-parametric Spearman and Procrustes correlation coefficients, respectively (Figure 4E). For this threshold, the average relative abundance per sample was 0.094%, corresponding to an average sequencing power of 1 367 111 reads (Table 3) per sample when both prokaryotic and eukaryotic OTUs retrieved with this specific set of primers (153) were

taken into account. By separating the OTUs identified as prokaryotes from those identified as eukaryotes, the prokaryotes corresponded to an average of 1 005 209 reads per sample, with an average relative abundance threshold per sample of 0.13%. For the OTUs identified as eukaryotes, the average number of reads was 361 902, and the average relative abundance threshold would be 1.9%. Applying the same reasoning for the large library sizing (MetaB 16S large), the total number of reads classified into OTUs per sample was, on average, 1 810 973, corresponding to an average relative abundance per sample of 0.514%. Here, the average number of reads and relative abundance per sample was 454 897 and 7.19% for prokaryotes, respectively, while for eukaryotes these values equaled 1 356 076 reads and 4.9%, respectively. Thus, by applying the sizing strategy for the smaller amplicon size (MetaB 16S small) mostly prokaryotic reads were obtained, while sizing for the large size (MetaB 16S large) resulted in the retrieval of mostly eukaryotic sequences. When no sizing was used, resulting in the analysis of both prokaryotic and eukaryotic reads at the same time, a much different rarity threshold was observed in comparison with small and large sizing (Table 3).

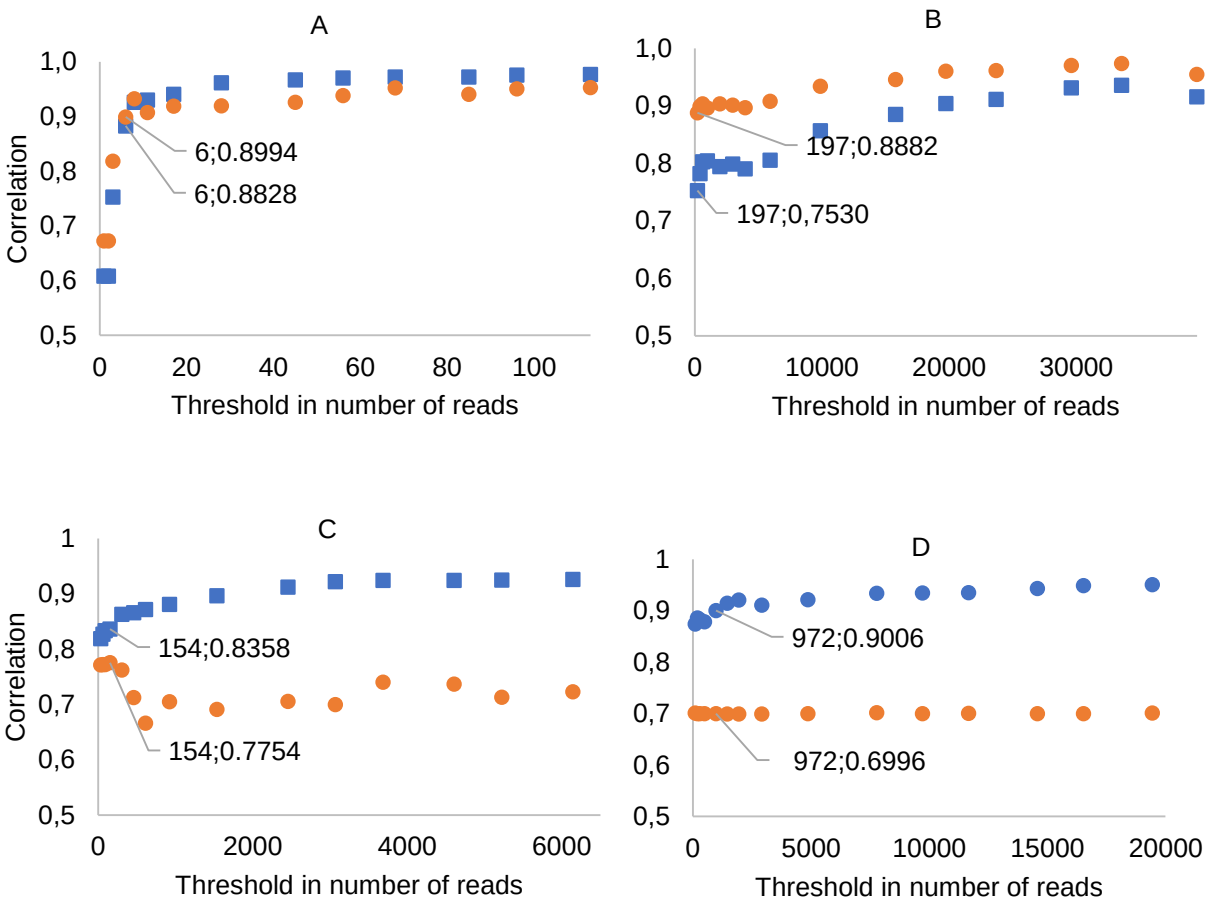

Figure 4 is continued next page.

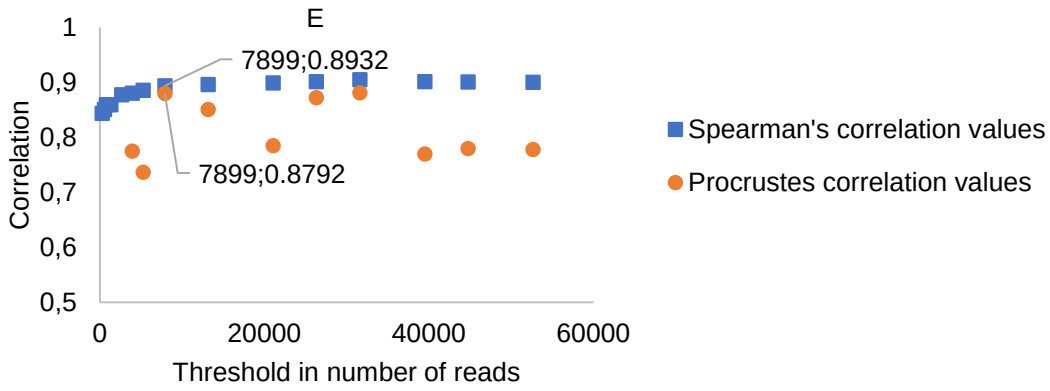

Figure 4. MultiCoLA results for the EMOSE 2017 dataset. Correlation values between the truncated community and the original community for each threshold tested. Thresholds are presented in number of reads per sample. Correlations are given by the non-parametric Spearman's correlation coefficient (blue squares) and Procrustes correlation coefficient (orange circles). 4A – Results for MetaG 16S, for prokaryotic data; 4B – Results for MetaB 18S; 4C – Results for MetaB 16S nS; 4D – Results for MetaB 16S small; 4E – Results for MetaB 16S large.

Furthermore, the count of rare OTUs identified in MetaB 16S large was superior than that obtained for MetaB 16S small, due to the difference in thresholds obtained (Figure 5). When the OTUs identified as prokaryotes and eukaryotes were separated within the MetaB 16S large and small datasets, the count of rare prokaryotic OTUs was superior for MetaB 16S small. Thus, library sizing improves rare prokaryotic diversity assessment for this set of degenerate primers.

The pattern identified in the correlation values of the MultiCoLA results (Figure 4) was not the same as expected (Figure 3), because the correlation values decrease in a gradual, instead of drastic, way suggesting that there is no objective method to decide the exact threshold without a subjective choice.

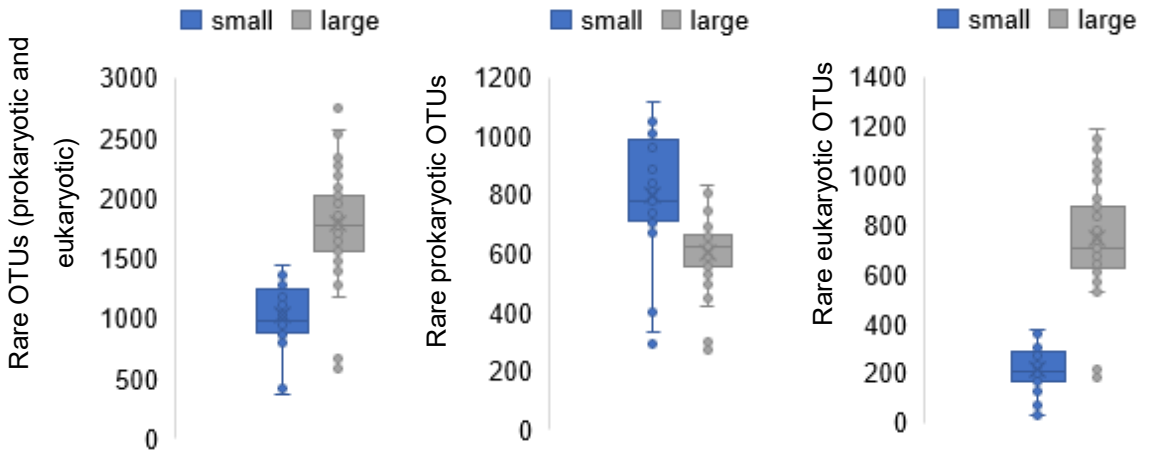

Figure 5. Comparison of rare OTUs count for MetaB 16S small and large sizing, for prokaryotes and eukaryotes combined and separated. Box plots with mean value, quartiles and outliers for the number of rare OTUs counted (rare prokaryotic and eukaryotic OTUs, rare prokaryotic OTUs and rare eukaryotic OTUs).

3.1.2 Testing MultiCoLA on the *Spongia officinalis* 2014 dataset

For the *Spongia Officinalis* 2014 dataset, following the same reasoning as previously, the rarity threshold decided was 13 reads per sample (Figure 6), representing an average of 0.44% relative abundance per sample. This threshold was similar to the threshold obtained with similar methodologies in the previous dataset (MetaG 16S from EMOSE 2017), with 3224 reads per sample, on average. This dataset compares three different types of samples (sediment, seawater and sponge tissue) with characteristic communities (158). MultiCoLA is expected to have different behaviors according with the number of samples used and is expected to give different results according with the number of reads (13,108). Thus, to use all samples or to use different groups of samples (i.e. sediment, seawater and sponge tissue) separately can induce different results. To test that, the algorithm was also applied separately for each group of samples (Figure 7). The sponge tissue samples resulted in a threshold of 36 reads per sample, representing a relative abundance threshold ranging from 1.15% to 1.6%. For the sediment and seawater samples, the thresholds were 7 and 5 reads per sample, respectively and ranging from 0.26% to 0.36% relative abundance (in sediment) and 0.1% (in seawater). Figure 7 shows that values are more consistent when using all samples simultaneously rather than separately for each individual microhabitat.

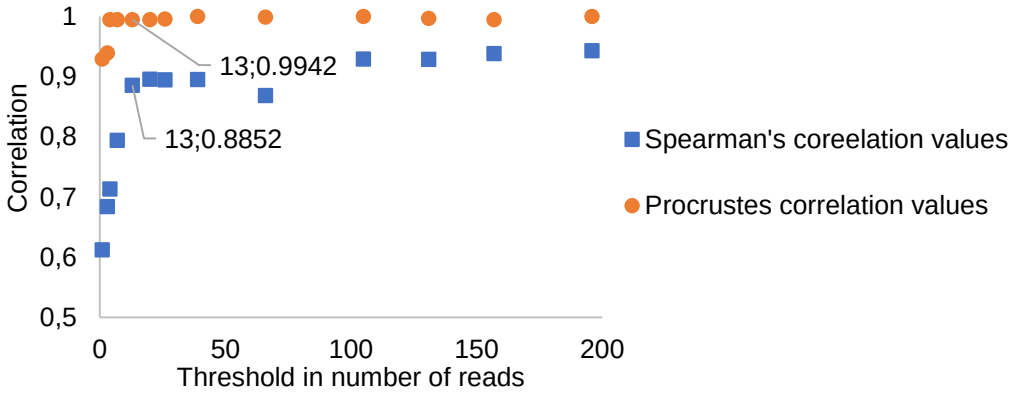

Figure 6. MultiCoLA results for the *Spongia officinalis* 2014 dataset. Correlation values between the truncated community and the original community for each threshold tested. Thresholds are presented in number of reads per sample. Correlations are given by the non-parametric Spearman's correlation coefficient (blue squares) and Procrustes correlation coefficient (orange circles).

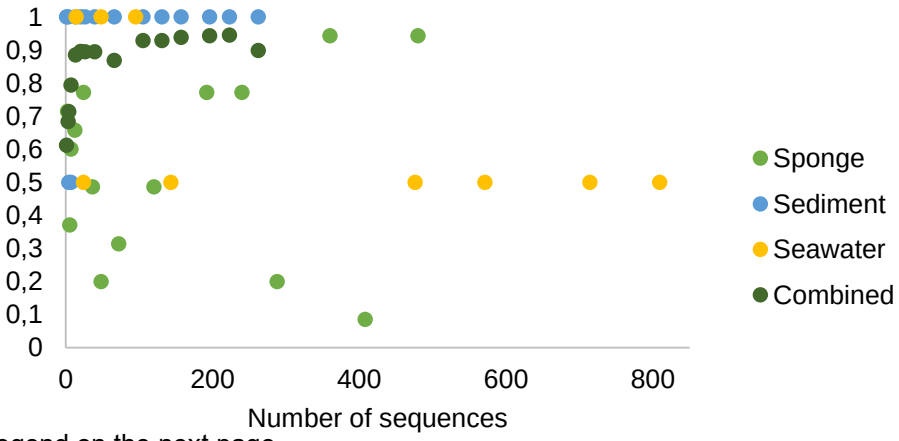

Figure 7 legend on the next page.

Figure 7. Comparing MultiCoLA across different types of samples in the *Spongia officinalis* 2014 dataset. Non-parametric Spearman's correlation values for each group of samples (sediment in blue, seawater in yellow or sponge tissue in green) and for all samples combined, in dark green.

3.1.3 Testing MultiCoLA on the NICE 2015 dataset

For the TC-DNA shotgun sequencing data of the NICE 2015 dataset, MultiCoLA resulted in a threshold of 42 reads per sample (Figure 8A), representing a mean relative abundance of 1.12%, ranging from 0.86% to 1.6%, per sample. For the 16S amplicon sequencing data, the absolute value threshold was 1054 reads per sample (Figure 8B), representing a mean relative abundance of 0.6% ranging from 0.24% to 0.95% per sample. The thresholds obtained in absolute values were higher than the values obtained in the EMOSE and *Spongia officinalis* datasets, when comparing similar metagenomic strategies. Also, the values for TC-DNA shotgun sequencing and 16S rRNA gene amplicon sequencing differed due to different number of reads delivered by each sequencing strategy. In the TC-DNA shotgun sequencing data, there were on average 3488 16S rRNA gene reads per sample, while in the 16S rRNA gene amplicon sequencing data, there were on average 212 365 reads per sample.

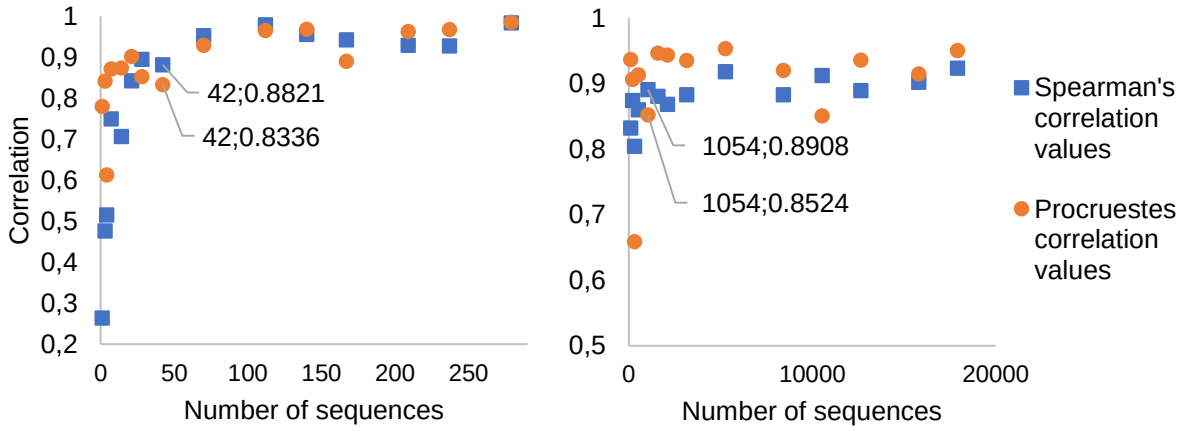

Figure 8. MultiCoLA results for the NICE 2015 dataset. Correlation values between the truncated community and the original community for each threshold tested. Thresholds are presented in number of reads per sample. Correlations are given by the non-parametric Spearman's correlation coefficient (blue squares) and Procrustes correlation coefficient (orange circles).

3.2 Methodological assessment of the marine prokaryotic rare biosphere

3.2.1 Seawater sampling effect on the prokaryotic rare biosphere, on the EMOSE 2017 dataset

The MetaB 16S nS data from EMOSE 2017 was selected to study how the prokaryotic rare diversity is affected by the type of filter unit (Sterivex vs membrane), the filtering methodology (whole water vs size fractioned filtration), the filtered volume and size fractioning (sometimes equivalent to pre filtration), of seawater samples. The subset without library sizing is the equivalent approach to other studies in the literature. The effect of different sampling methodologies in recovering the prokaryotic rare biosphere

diversity were first evaluated by comparing alpha diversity metrics, according with the grouping of variables presented in Table 4. Alpha diversity metrics (number of OTUs, number of reads and Shannon diversity index) were applied to the rare, abundant and total community. The number of rare OTUs represents between 75% and 95% of all OTUs present, while the relative abundance per sample of those rare OTUs ranges from 0.57% to 3.3%.

The diversity, as measured by the Shannon index, was always higher in the prokaryotic rare biosphere than in the total community (for example, Figure 9). There were no significant differences in rare prokaryotic richness and diversity between 2.5L and 10L samples for the Sterivex filter unit (Table 5, group i). For 10L volumes, when comparing Sterivex and membrane filtering units, the values were similar as well, except for the number of OTUs, which were higher with the membrane filtering method, though not significantly (Table 5, group ii). For 10L, across the small (0.22-3µm) and medium fractions (3-20µm) there were significant differences (Table 5, group iii), with the small fraction displaying more reads for the total and abundant prokaryotes, but less reads for the rare ones. For 100L volumes, comparing for the small, medium and large fractions (3µm to 20µm), the differences in number of reads were not significant for the small and medium fractions, for the total and abundant communities, but there was a significant decrease for the large fraction (Table 5, group iv and figure 9). For the rare prokaryotic community, the number of reads increased with size fraction (Figure 9). For the number of OTUs and Shannon index there was a general increase in the prokaryotic rare biosphere, by increasing the size fraction (Figure 9). For 496L and superior volumes the patterns are like 100L (Table 5, groups v and vi). It is noteworthy that the 496L samples did not include replicas for the small and medium fractions, due to sampling constraints of very large volumes. Also, the values above 496L include samples with 716L and 760L, instead of the desired 1000L, due to sampling constraints. For the small fraction, across all volumes, except for the high variance in the number of reads, there were no evident differences in diversity values (Table 5, group vii, small fraction). For the medium fraction, across all volumes, 10L had less reads than the remaining, but beyond that, there was no increase in reads after 100L (Table 5, group vii, medium fraction). For the OTUs number and Shannon index, there were no differences overall. For the large fraction, the only different pattern is the increase in OTUs number and Shannon index for 1000L (Table 5, group vii, large fraction). Overall, significant changes are associated with size fractionation (Table 5). For the alpha metrics, the samples with 100L are shown as illustrative of the most important patterns found (Figure 9).

Table 4. Summary of the variables studied across samples in the EMOSE 2017 dataset. The type of filter, filtering methodology and filtered volume are listed for each group, with the variable analyzed.

Group number	Type of filter	Filtering method	Filtered volume (L)	Variable analysed
i	Sterivex	Whole water	2.5 – 10	Volume
ii	Sterivex and membrane	Whole water	10	Type of filter
iii	Membrane	Size fraction	10	Size fraction
iv	Membrane	Size fraction	100	Size fraction
v	Membrane	Size fraction	496L	Size fraction
vi	Membrane	Size fraction	716, 760 and 1000	Size fraction

vii	Membrane	Small size fraction	10 – 1000	Volume
	Membrane	Medium size fraction	10 – 1000	Volume
	Membrane	Large size fraction	10 – 1000	Volume

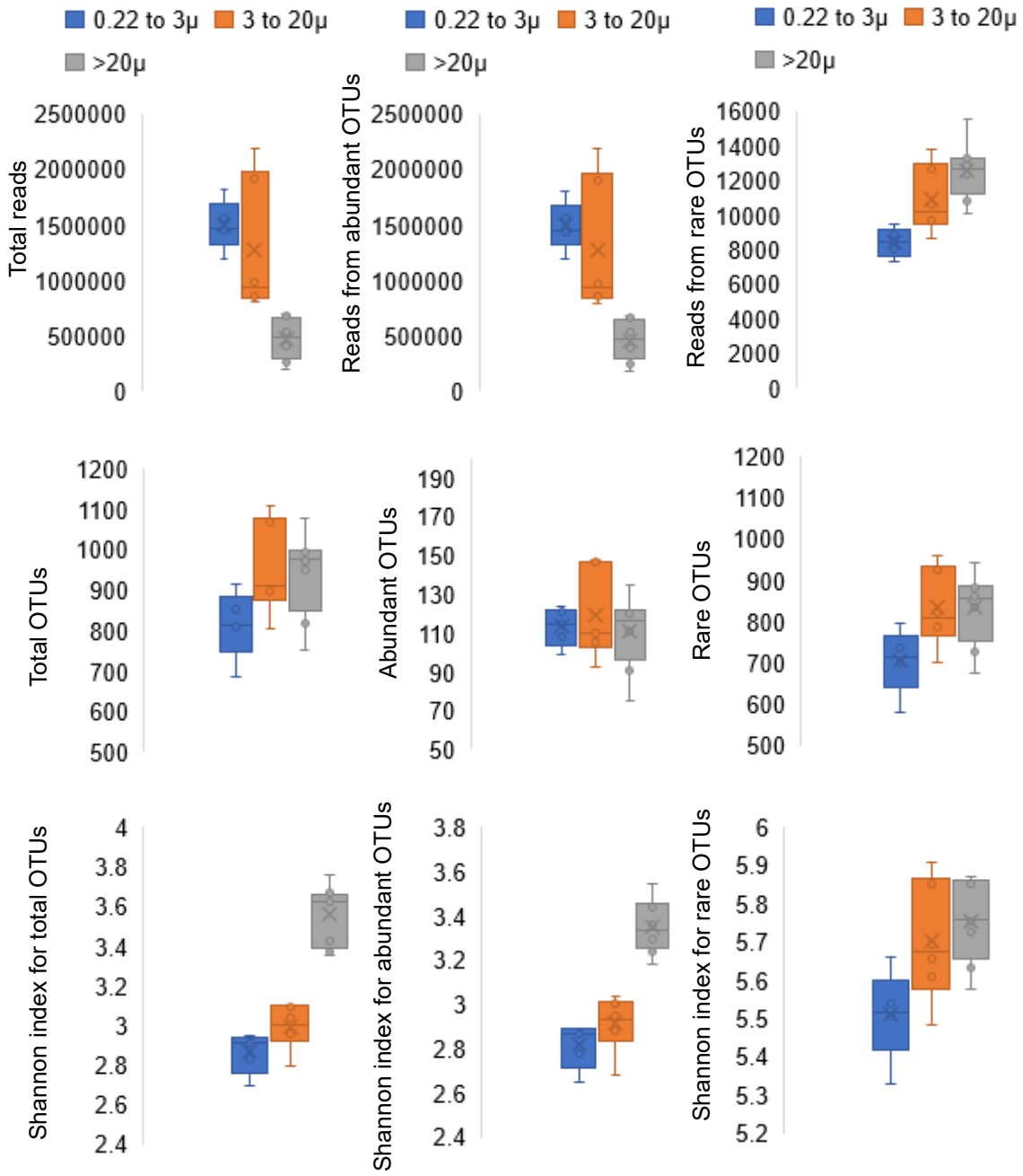

Figure 9. Alpha diversity plots for 100L, comparing for small (0.22 to 3 µm), medium (3 to 20 µm) and large size fractioning (more than 20 µm). Alpha metrics applied are the number of reads, the number of OTUs and the Shannon index. All metrics were applied separately to the total community, for the abundant community and for the rare community.

Table 5. Significance values for alpha diversity differences across the compared variables in Table 4, from the EMOSE 2017 dataset, regarding rare prokaryotic OTUs. One-way ANOVA test for total, abundant and rare communities, for alpha metric values. Rare and abundant communities were divided using a threshold of 154 reads per sample, from section 3.1.1. P-values<0.05 are in **bold**.

Variable compared (group number)	Alpha metric	Total OTUs	Abundant OTUs	Rare OTUs
Volume (i)	Nº of OTUs	0.922	0.983	0.92
	Shannon index	0.471	0.483	0.632
	Nº of reads	1	0.551	0.85
Type of filter (ii)	Nº of OTUs	0.331	0.698	0.571
	Shannon index	0.201	0.263	0.095
	Nº of reads	0.272	0.271	0.572
Size fraction (iii)	Nº of OTUs	**0.0151**	0.083	**0.0139**
	Shannon index	**0.0000109**	**0.0000105**	**0.00131**
	Nº of reads	**0.0014**	**0.00144**	**0.00077**
Size fraction (iv)	Nº of OTUs	**0.0132**	0.807	**0.0267**
	Shannon index	**2.52E-10**	**5.36E-09**	**0.00066**
	Nº of reads	**0.0000448**	**0.0000409**	**0.0313**
Size fraction (v)	Nº of OTUs	0.118	0.528	**0.00019**
	Shannon index	**4.42E-13**	**3.57E-08**	**0.0056**
	Nº of reads	**0.000549**	**0.0005**	**0.00055**
Size fraction (vi)	Nº of OTUs	**0.0159**	0.101	**0.00092**
	Shannon index	**0.0045**	**0.0316**	**0.0005**
	Nº of reads	0.175	0.169	**1.52E-07**
Volume - small fraction (vii)	Nº of OTUs	0.0762	0.129	0.06
	Shannon index	0.731	0.651	0.388
	Nº of reads	**0.0349**	**0.0349**	**0.024**
Volume - medium fraction (vii)	Nº of OTUs	0.801	0.15	0.933
	Shannon index	**7.08E-13**	**1.1E-10**	0.548
	Nº of reads	**0.0005**	**0.0048**	**0.0073**
Volume - large fraction (vii)	Nº of OTUs	0.172	**0.086**	**0.0362**
	Shannon index	**0.00599**	**0.000976**	**0.0033**
	Nº of reads	0.148	0.146	0.0847

The previous analysis was not informative from the point of view of community composition. To define the variables responsible for changes in community composition, ordination analysis was used. The gradient length, as measured by DCA, is of 1.25, meaning linear methods should be used. With an unconstrained analysis, by PCA (Figure 10), it is possible to identify different patterns. The Sterivex filter unit samples are grouped in a small area completely covered by the membrane filter unit samples area (Figure 10, compare squares with black and purple border). Thus, membrane filter samples have a broader composition (in rare prokaryotic OTUs) than the Sterivex ones, but both filters have different range of volumes. Within the same range of volumes, both types of filter are in the same area. For 10L, 100L, 496L and 1000L samples cover a broad area, whereas the 2.5L samples are restrained within a

smaller area (Figure 10, compares yellow squares with remaining squares). For the filtering methodology, three separate groups clearly represent each size fraction, with the whole water filtering within the area of the small fraction size (Figure 10, compares black, gray and pink lines). Furthermore, the areas covered are much narrower and specific, indicating that each size fraction represents a different rare community.

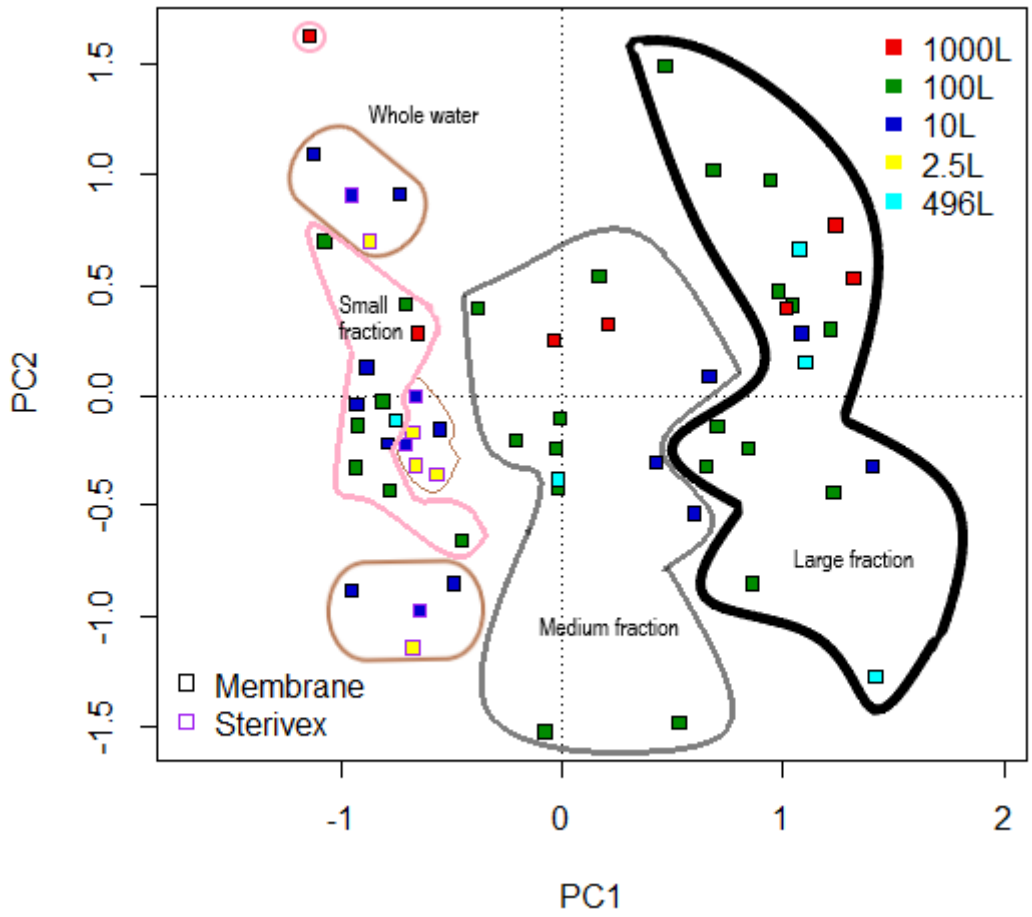

Figure 10. PCA of MetaB 16S nS data from the EMOSE 2017 dataset. Samples are illustrated according with volume (squares colored in red for 1000L, green for 100L, blue for 10L, yellow for 2.5L and cyan for 496L), filter type (square border in black for membrane filter units and purple for Sterivex filter units). Different areas are highlighted according with the filtration method (black line for the large fraction, gray line for the medium fraction, pink line for the small fraction and brown line for the whole water filtration).

3.3 Sponge-associated prokaryotic rare biosphere (*Spongia officinalis* 2014 dataset)

Within the *Spongia officinalis* 2014 dataset, the prokaryotic rare biosphere was more diverse than the abundant and total biosphere in all types of samples (sponge tissue, seawater and sediment), despite of the lower number of reads from rare OTUs (Table 6). For the abundant community, the main number of reads is on the sediment, followed by seawater and sponge tissue. For the number of OTUs and for

the Shannon index, the sediment was always superior to seawater and sponge tissue samples, for both the total, abundant and rare biosphere. The differences on diversity are significant across different types of samples (Table 7).

Table 6. Alpha diversity for the *Spongia officinalis* 2014 dataset samples. Values for alpha diversity are the Shannon index, number of OTUs and number of reads (of 16S rRNA gene sequences used for taxonomy). For each sample, communities are divided in total, abundant and rare, according with the OTUs abundance, using the rarity threshold of 13 reads per sample, from section 3.1.2.

Sample	Abundance	Shannon index	Number of OTUs	Number of reads
Sponge_1	Rare	3.96	71	142
	Abundant	2.42	19	3016
	Total	2.67	90	3158
Sponge_2	Rare	4.21	93	183
	Abundant	2.36	18	2662
	Total	2.71	111	2845
Sponge_3	Rare	3.80	66	177
	Abundant	2.37	15	2075
	Total	2.75	81	2252
Sponge_4	Rare	4.05	83	182
	Abundant	2.35	21	2948
	Total	2.67	104	3130
Sediment_1	Rare	5.06	220	511
	Abundant	3.02	32	2039
	Total	3.93	252	2550
Sediment_2	Rare	4.75	163	402
	Abundant	2.89	26	1553
	Total	3.78	189	1955
Sediment_3	Rare	5.15	240	577
	Abundant	3.15	36	2163
	Total	4.08	276	2740
Seawater_1	Rare	4.34	107	278
	Abundant	2.47	20	4335
	Total	2.81	127	4613
Seawater_2	Rare	4.40	111	269
	Abundant	2.47	22	3891
	Total	2.84	133	4160
Seawater_3	Rare	4.56	137	327
	Abundant	2.49	24	4511
	Total	2.88	161	4838

Table 7. Significance values for alpha diversity differences across sponge tissue, sediment and seawater samples, from the *Spongia officinalis* 2014 dataset. One-way ANOVA test for total, abundant and rare communities, for alpha metric values. Rare and abundant communities were divided using a threshold of 13 reads per sample, from section 3.1.2. P-values<0.05 are in **bold**.

Samples compared	Shannon index	Number of OTUs	Number of reads	
Sponge vs sediment	**0.00017**	**0.000291**	**0.000122**	Rare
Sponge vs seawater	**0.00718**	**0.00758**	**0.000381**	
Sediment vs seawater	**0.0151**	**0.023**	**0.0192**	
	Shannon index	Number of OTUs	Number of reads	
Sponge vs sediment	**2.21E-05**	**0.00152**	**0.0284**	Abundant
Sponge vs seawater	**0.000852**	**0.0846**	**0.000731**	
Sediment vs seawater	**0.00189**	**0.0405**	**0.000888**	
	Shannon index	Number of OTUs	Number of reads	
Sponge vs sediment	**1.39E-06**	**0.000324**	0.193	Total
Sponge vs seawater	**0.00186**	**0.00865**	**0.000522**	
Sediment vs seawater	**0.00026**	**0.0243**	**0.00236**	

Because the sponge tissue harbors a diverse microbiome highly influenced by the surrounding environment (79), it is important to know how many different OTUs are shared across different types of samples and how many are specific of each type of sample. Venn diagrams were used to illustrate those patterns (Figure 11). Different patterns emerge when dividing the community in total, rare and abundant, according with the threshold decided in section 3.1.2 (13 reads per sample). From a logical standpoint, the total community is equal to the rare community plus the abundant community and that applies to the Venn diagrams obtained. But it only applies to the entire Venn, meaning that it is not possible to add sponge specific rare OTUs and sponge specific abundant OTUs, to get the number of sponge specific OTUs from the total community. This is well illustrated with the example of CRT: consider OTU 77, assigned to the genus *Rubrobacter*. This OTU is rare in the sponge and abundant in seawater and sediment. Thus, it will be considered shared across all types of samples in the total community Venn diagram. But, when using the rare community, the abundant OTUs are removed, thus, it will be absent (not biologically absent, but absent as a rare OTU) in the seawater and sediment samples. From this logical constraint, because it is a CRT, it will be considered sponge-specific in the rare community Venn diagram and it will be considered shared across sediment and seawater in the abundant community. Despite that, when considering the total number of OTUs in the rare and abundant communities, the sum was equal to the sum in the total community. For those reasons, the Venn diagrams for analyzing patterns of shared and specific OTUs across different abundance categories should be accompanied with the types of rarity, to get a more reasonable interpretation (data not available). Seawater had more shared OTUs than specific OTUs for the rare and total biosphere, but not for the abundant biosphere, where most OTUs are specific. For the rare and total communities, within the seawater shared OTUs, most are shared with sediment or with sediment and sponge simultaneously, the shared OTUs between seawater and sediment are a minority. For the rare and total biosphere, sediment OTUs are mostly specific, and most of the shared OTUs are shared with seawater. Sponge tissue rare OTUs are mostly

shared with sediment. Total community OTUs, from the sponge tissue, are mostly shared with sediment and seawater simultaneously (Figure 11).

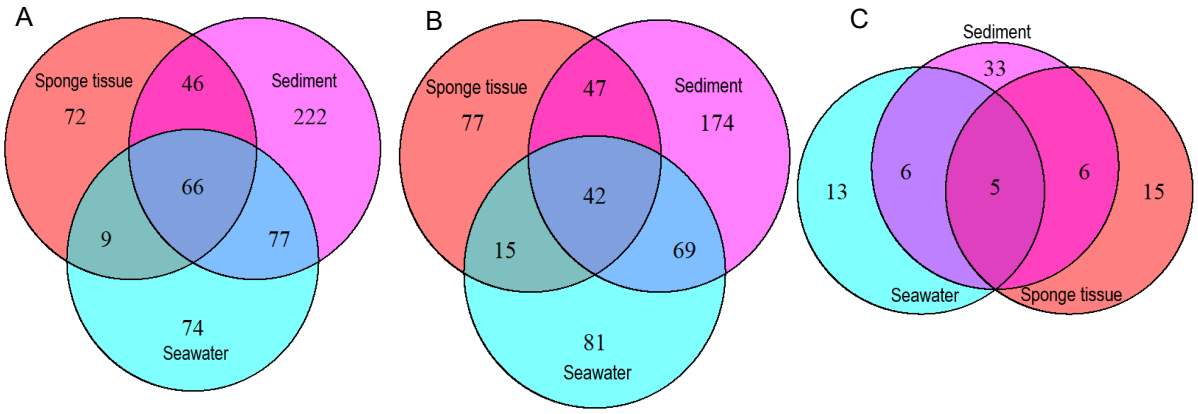

Figure 11. Venn diagrams for shared and specific prokaryotic OTUs across different samples, in the *Spongia officinalis* 2014 dataset. A – Shared and specific OTUs, from the total community, across sponge tissue, sediment and seawater samples; B – Shared and specific OTUs, from the rare community, with abundance <13 reads per sample, across sponge tissue, sediment and seawater samples; C – Shared and specific OTUs, from abundant community, with abundance ≥13 reads per sample, across sponge tissue, sediment and seawater samples. The rarity threshold was selected based on section 3.1.2.

Ordination analysis can be used to understand if the quantitative patterns found with the shared OTUs listing and alpha diversity comparisons also apply for community composition, for the prokaryotic rare biosphere. The gradient of variance, as measured by DCA, is 2.27, thus selecting linear methods. From the PCA analysis (Figure 12), seawater and sponge samples are clustered together, indicating similar rare prokaryotic OTUs composition. Whereas the sediment sample are not clustered in one specific area and are further apart from the sponge and seawater samples.

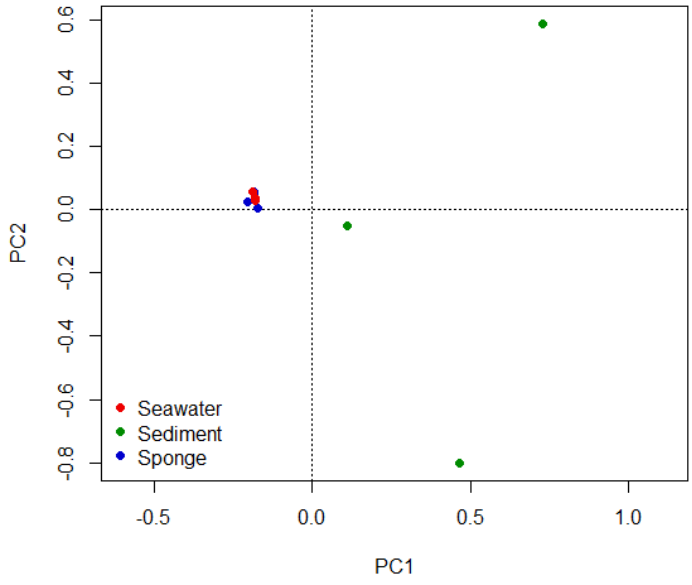

Figure 12. Legend description on next page.

Figure 12. PCA of TC-DNA shotgun sequencing from the *Spongia officinalis* 2014 dataset, for the prokaryotic rare biosphere across sponge tissue, sediment and seawater samples. Sponge tissue samples are colored in blue, sediment samples are colored in green and seawater samples are colored in red.

To view diversity from a qualitative point of view, a circular Figure was produced with Circos software (Figure 13). This Figure allows to visually understand and relate the relative abundance and taxonomic diversity of all OTUs, without removing essential information. There are, in total, 44 different prokaryotic phyla, not considering OTUs not assigned to any phylum, with 15 bacterial candidate phyla. From Figure 13, it is evident that the phyla with higher OTU richness are Proteobacteria, Bacteroidetes, Actinobacteria, Verrucomicrobia and Firmicutes. By further dividing each phylum in different classes, those major phyla are divided from 5 to 7 different classes. Interestingly, some of the less OTU-rich phyla were represented by many different classes, meaning they have few different OTUs in each specific class, for example, the phylum Chloroflexi contained OTUs classified in 7 different classes, with only 1 to 2 different OTUs per class. Most phyla were represented by few classes each containing many different OTUs while possessing many classes represented by few different OTUs. For example, the phylum Firmicutes has 17 OTUs belonging to the class Clostridia and 10 OTUs divided into five other classes.

In the heatmaps from Figure 13, abundance is proportional to the color intensity, turning visible that most OTUs, across all samples, are rare and only a few are abundant. This is in accordance with the previous analysis, but in Figure 13 no numerical threshold is used to differentiate between rare and abundant OTUs, rarity is rather inferred from the gradient of colors. An important component of the pool of abundant OTUs are those not classifiable at phylum or class levels (Figure 13). Those OTUs belong to the microbial “dark matter” and are probably sub divided in other phyla and classes, meaning that an important component of diversity remains elusive (175). It is also noteworthy that, as expected from the previous analysis, some rare OTUs are specific of each environment (sponge tissue, sediment or seawater) and other rare OTUs are shared and/or have variable abundance from one environment to the other. More importantly, some phyla clearly have low abundance OTUs and absent OTUs in sponge tissue (in comparison with sediment and seawater), namely the phyla: Bacteroidetes, all candidate phyla (except for Candidatus Poribacteria), Ignavibacteriae, Planctomycetes, Rhodothermaeota, Synergistetes, Tenericutes and Verrucomicrobia. Proteobacteria, despite showing many different OTUs in sponge tissue samples, had lower abundances when compared with sediment and seawater samples. Overall, sediment samples displayed more intense colors compared with seawater and sponge tissue samples, despite having mostly rare OTUs. The sediment samples also had the lowest number of white spots (absent OTUs), indicating that it is the most diverse group of samples. There are some exceptions, for example, the phylum Candidatus Poribacteria is clearly abundant in sponge tissues and rare in sediment and seawater samples. The phyla with more OTUs in sponge tissue, when compared to other samples, include Thaumarchaeota, Acidobacteria, Chloroflexi, Gemmatimonadetes and in some groups within the classes Alpha and Gammaproteobacteria, within the Proteobacteria phylum. All candidate phyla in the sponge tissue, except for Candidatus Poribacteria, are always rare, suggesting a

relationship between the rare biosphere and the microbial “dark matter”. The sediment samples, despite having mostly rare OTUs, have some phyla with some abundant OTUs, such as Bacteroidetes, Planctomycetes and Proteobacteria (with exceptions). Other phyla, despite rare, are more abundant in sediment than in seawater or sponge. For the seawater samples, the major exceptions to rarity are in the phyla Bacteroidetes and some classes in Proteobacteria, such as Alpha and Epsilonproteobacteria, with the remaining phyla being mostly rare. Overall, most of the prokaryotic diversity was found in sediments, followed by seawater and sponge tissue (see gray section of Figure 13 and Table 6).

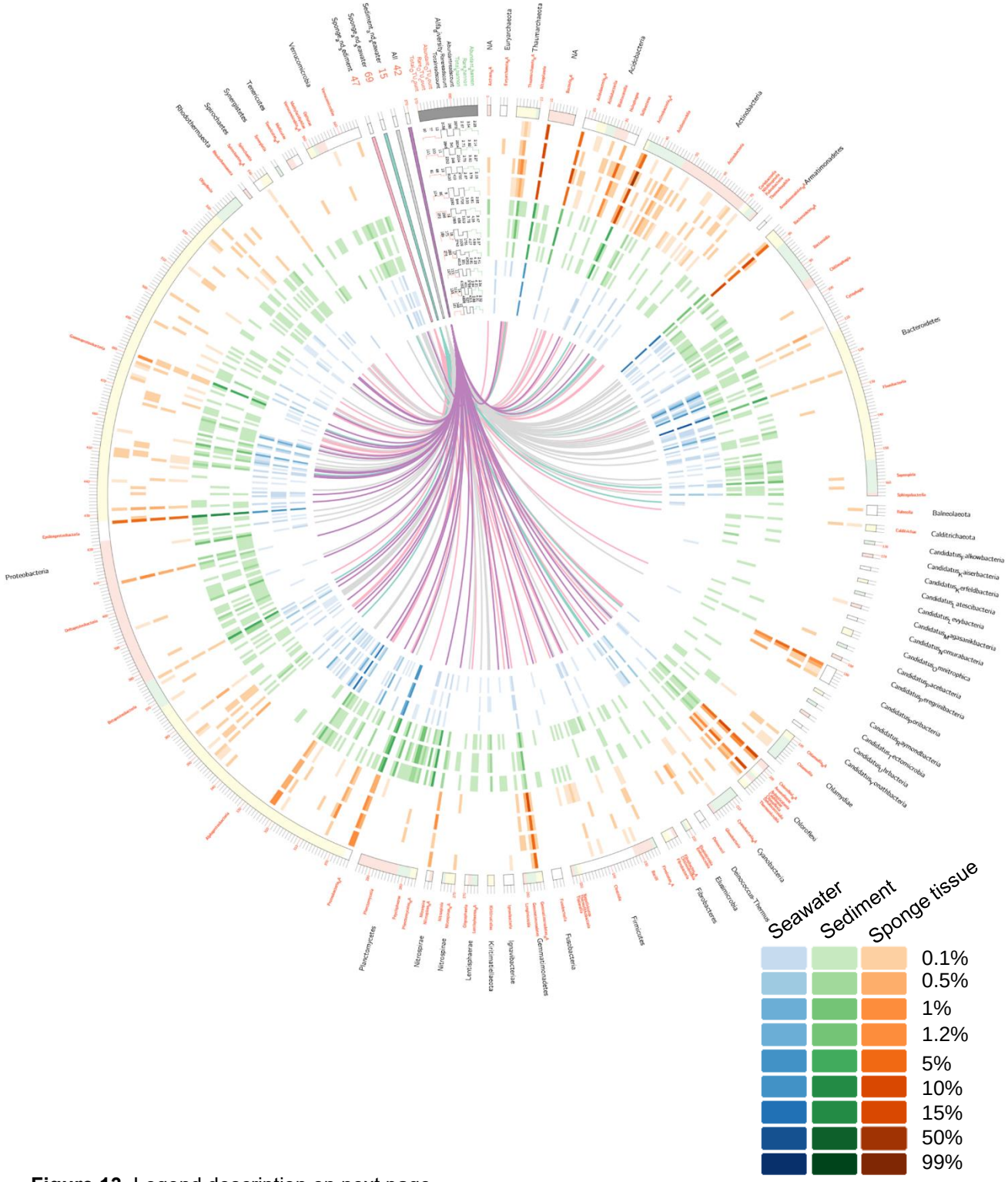

Figure 13. Legend description on next page.

Figure 13. Circular visualization of the *Spongia officinalis* 2014 dataset, for prokaryotes. Figure produced using Circos software. Read the Circos Figure clockwise, from outside to inside. All OTUs found at least in one sample are numbered from 1 to 566. OTUs are organized at phylum level by separated groups. Within each phylum section, colored bars are used for different classes and names are labeled in red. Heatmaps represent the abundance of each OTU in a given sample, with sponge tissue samples represented in orange, sediment samples represented in green and seawater samples represented in blue. White spaces in the heatmaps represent absence. The color gradient is highlighted at the bottom-left side of the Figure. Links highlight which OTUs are shared across: sponge tissue and sediment (pink), sponge tissue and seawater (light green), sediment and seawater (gray) and OTUs present in all types of samples (purple link). The gray slice, at the end of the Circos Figure, summarizes alpha diversity metrics, comparing the OTU count, reads count, Chao1, Shannon index and inverse of Simpsons for the total, rare and abundant OTUs. The gray slice area distinguishes rare from abundant OTUs using the threshold of 13 reads per sample, from section 3.1.2. This Figure can be better visualized using the virtual rather than the printed version one of this thesis.

3.4 Spatiotemporal and depth effects on the marine prokaryotic rare biosphere of the Arctic ocean, on the NICE 2015 dataset

The NICE 2015 dataset was used to study how the marine prokaryotic rare biosphere behaves through spatiotemporal variation (seasonal transition from winter to spring, corresponding to March, April and June samples, with spatial drift along the ice) and depth (transition from surface, to middle and bottom layer, corresponding to 5m, 25m or 50m and 250m samples). It was also used to study the influence of water masses on the prokaryotic communities. This dataset is divided in TC-DNA shotgun sequencing and 16S rRNA gene amplicon sequencing.

3.4.1 TC-DNA sequencing data from the NICE 2015 dataset

For TC-DNA shotgun sequencing data, there was a total of 31 400 16S rRNA gene amplicon reads, ranging from 2617 reads to 4851 reads per sample (Table 8). Most OTUs belong to the prokaryotic rare biosphere, but as they were represented by low numbers of sequences, abundant OTUs accounted for most of the reads obtained in the dataset, as usual. In total, 13 to 20 abundant prokaryotic OTUs versus 221 to 301 rare prokaryotic OTUs were found, when considering all samples (Table 8). Diversity was always higher within the rare component, confirmed by the Shannon index values (Table 8). When comparing samples from March, April and June for the number of reads, there were more total and abundant reads in June than in March and April (Table 8). For the prokaryotic rare biosphere, March was lower in diversity than April and June, but the latter two displayed similar values. For the number of OTUs, there was some increase during the March to June transition for the total community, but not significant (Table 9). For the abundant and rare biosphere, the number of OTUs across months displayed no significant differences (Table 9). When comparing samples from Surface, Middle and Bottom water layers, there were no significant differences across depth, for all abundances and considering all alpha diversity metrics (Table 9). When comparing different water masses, despite the different number of

samples compared, all water masses showed different alpha diversity values. The more diverse water masses were PSWw and AW.

Table 8. Alpha diversity for the NICE 2015 dataset samples, for prokaryotic data identified from TC-DNA shotgun sequencing. Values for alpha diversity are the Shannon index, number of OTUs and number of reads (of 16S rRNA gene sequences used for taxonomy). For each sample, communities are divided in total, abundant and rare, according with the OTUs abundance, using the rarity threshold of 42 reads per sample, from section 3.1.3.

Sample	Abundance	Shannon index	Number of OTUs	Number of reads
NICE_1	Rare	4.86	257	1111
	Abundant	2.15	13	1913
	Total	3.80	270	3024
NICE_2	Rare	4.89	264	1124
	Abundant	2.18	14	2130
	Total	3.76	278	3254
NICE_3	Rare	4.76	221	908
	Abundant	2.24	12	1994
	Total	3.65	233	2902
NICE_4	Rare	4.93	277	1228
	Abundant	2.54	18	2180
	Total	4.05	295	3408
NICE_5	Rare	4.95	269	1268
	Abundant	2.26	14	1779
	Total	4.06	283	3047
NICE_6	Rare	4.88	256	1081
	Abundant	2.32	13	1536
	Total	4.05	269	2617
NICE_7	Rare	4.92	262	1089
	Abundant	2.23	15	2993
	Total	3.53	277	4082
NICE_8	Rare	5.10	301	1359
	Abundant	2.50	20	3496
	Total	3.82	321	4855
NICE_9	Rare	4.92	278	1241
	Abundant	2.32	13	2970
	Total	3.69	291	4211

Table 9. Alpha diversity differences across samples compared in the NICE 2015 dataset, for prokaryotic data identified from 16S rRNA gene sequencing. One-way ANOVA tests were performed for total, abundant and rare communities to determine if alpha metric values are significantly different. Rare and abundant communities were divided using a threshold of 42 reads per sample, from section 3.1.3. P-values<0.05 are in **bold**.

Variables compared	Shannon index	Number of OTUs	Number of reads	Abundance
Date/site	0.131	0.17	0.223	Rare
Depth	0.263	0.373	0.304	
Water mass	**1.46E-08**	**2.63E-07**	**3.76E-07**	
	Shannon index	Number of OTUs	Number of reads	
Date/site	0.204	0.42	**0.0017**	Abundant
Depth	0.987	0.296	0.884	
Water mass	**0.0029**	**7.75E-07**	**4.44E-06**	
	Shannon index	Number of OTUs	Number of reads	
Date/site	**0.0061**	0.174	**0.0047**	Total
Depth	0.861	0.349	0.782	
Water mass	0.965	**3.42E-06**	**1.08E-07**	

To understand what variables determine the structure of the prokaryotic rare biosphere within the NICE 2015 dataset, ordination analysis was used. The DCA value was 1.78, thus supporting the selection of linear multivariate ordination models for the analysis of this dataset. From the PCA diagram (Figure 14), samples from March and April were clustered together while June samples displayed a different prokaryotic rare biosphere composition. Also, June samples were distant from each other. The differences within June samples were due to different water masses, that are present at different depths, because of seasonal variation (161).

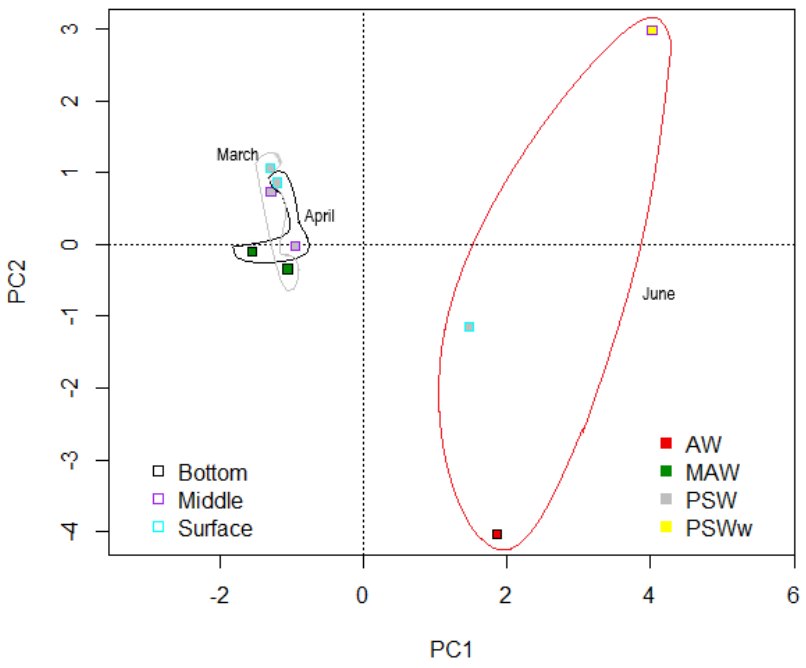

Figure 14. Legend description on next page.

Figure 14. PCA of TC-DNA shotgun sequencing data from the NICE 2015 dataset, for rare prokaryotes. Samples are illustrated according with water masses (squares colored in red for AW, green for MAW, gray for PSW and yellow for PSWw), sampling depth (square border in black for Bottom, purple for Middle and cyan for Surface) and different areas are highlighted according with the date of sampling (gray line for March, black line for April and red line for June).

Each OTU was labeled according with the type of rarity, if rare, using the function types.r, in R (Annex 3). The types of rarity were divided in spatiotemporal rarity and depth rarity, as the types of rarity were defined by comparing different variables. For spatiotemporal rarity, different depths were compared for the same date/site, whereas for rarity through depth, different dates/sites were compared for the same depth. Spatiotemporal and depth patterns were similar; therefore, in Figure 15 two illustrative examples are shown for (1) depth variation using March samples and (2) spatiotemporal variation using surface samples. All types of rarity were found, with transient rarity being the most prevalent, followed by permanent rarity. CRT were always the smallest group of rare OTUs, both from a spatiotemporal and across-depth perspective (Figure 15). Abundant OTUs represent the minority of OTUs, as previously described (larger number of reads corresponding to fewer OTUs). There were more permanently rare OTUs across depths than space/time. Considering that transient rarity is within the permanently rare biosphere, because those OTUs never grow abundant, results showed that rare OTUs remain rare in the Arctic ocean.

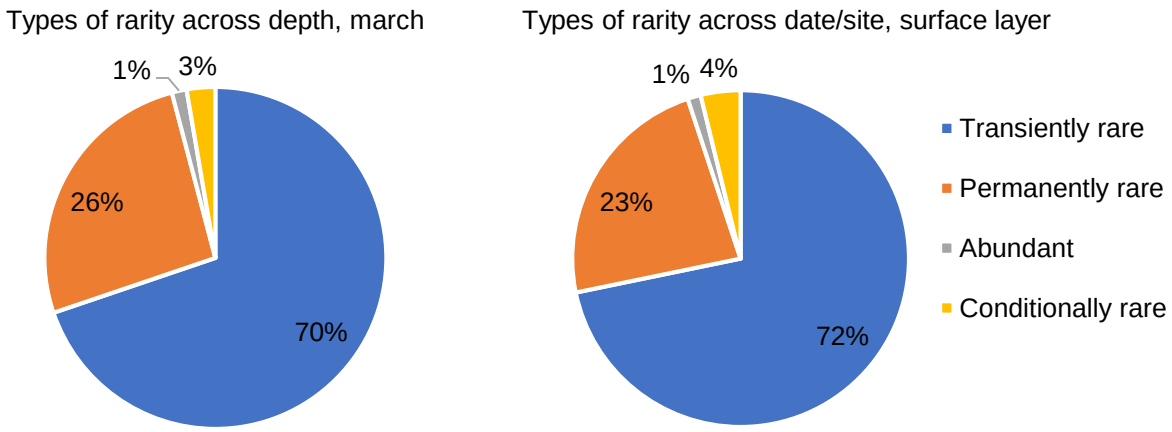

Figure 15. Percentage of prokaryotic rare OTUs (for each type of rarity) and abundant OTUs, identified in the TC-DNA shotgun sequencing data from the NICE 2015 dataset. The types of rarity are calculated according with the variables compared: Across depth, for march on the left; across date/site, for surface on the right. The algorithm types.r was used according with Annex 3.

With a circular visualization of the prokaryotic OTUs abundance across phyla and classes, it is possible to understand the different types of rarity from a qualitative perspective. In Figure 16, 76 different prokaryotic phyla are represented, not counting unidentified phyla (Bacterial and Archaeal NA's), with more than half being candidate phyla. From the inner heatmap, in blue, it is suggested that most of the candidate phyla are rare across depths and space/time. The inner blue heatmap, despite having low resolution, illustrates that most of the prokaryotic diversity is at low abundance, independently of the samples compared, as previously estimated in Table 8. Within more abundant phyla, e.g.

Bacteroidetes, Proteobacteria and Firmicutes, there were a few classes with high abundance, but most of the other classes were rare.

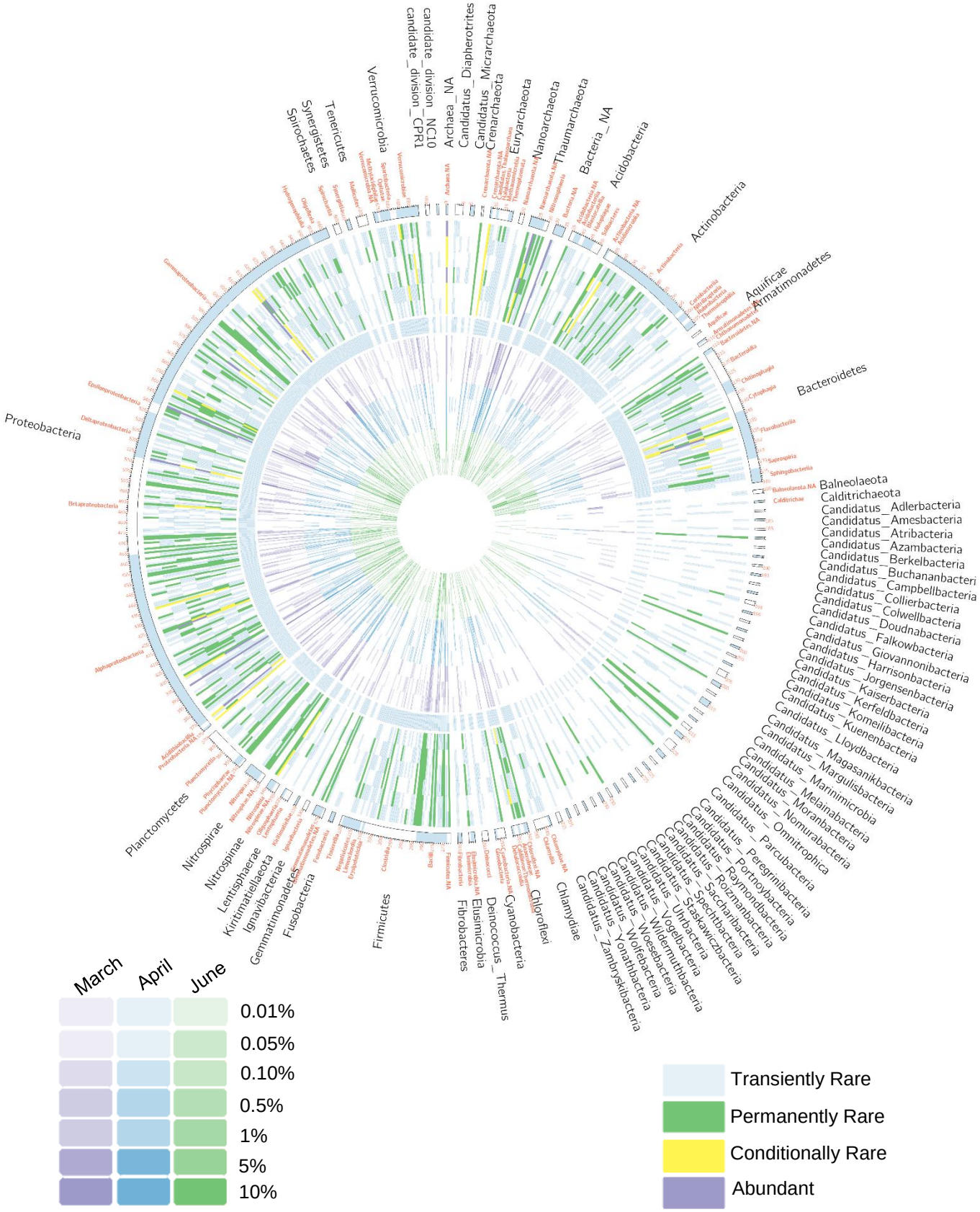

Figure 16. Legend description on next page.

Figure 16. Circular visualization of the TC-DNA shotgun sequencing data from the NICE 2015 dataset, for prokaryotes. Figure produced using Circos software. Read the Circos Figure clockwise, from outside to inside. All OTUs found at least in one sample are numbered from 1 to 697. OTUs are organized at phylum level by separated groups, with phylum labels in black. Within each phylum, different colored bars represent different classes, and class names are labeled in red. The outer heatmaps represent the types of rarity of each OTU, if rare, or abundant or absent, with the color code on the bottom right. White spaces in the heatmaps represent absence. The color gradient for the types of rarity is highlighted in the superior right side of the Figure. The inner heatmaps represent abundance, where white spaces represent absence. The color gradient for abundance is on the bottom left side of the gradient. This Figure can be better visualized using the virtual rather than the printed version one of this thesis.

Furthermore, from the inner heatmap it is immediately suggested that transient rarity is present, because there are many white spots (indicating absence) followed by light colored spots (indicating low abundance). Those patterns are confirmed by the outer heatmap, with a color code for each type of rarity, illustrating the calculations on Figure 15. Transient rarity is represented by light blue spots, present in all phyla. Candidate phyla were essentially rare, with some exceptions being permanently rare (green spots). A common pattern was that phyla with high numbers of transiently rare OTUs also had high numbers of permanently rare OTUs, phylogenetically close. For the conditionally rare biosphere, it was mostly represented by phyla such as Proteobacteria, Verrucomicrobia, Euryarchaeota, Actinobacteria, Bacteroidetes, Nitrospinae and Chloroflexi. For Verrucomicrobia and Bacteroidetes, most CRT belong to a single class. For example, the class Flavobacteria encompasses all CRT from Bacteroidetes phylum.

3.4.2 16S rRNA gene amplicon sequencing data from the NICE 2015 dataset

There was a total of 1 911 285 reads within the 16S rRNA gene amplicon sequencing from the NICE 2015 dataset, with 212 365 reads per sample on average. The same patterns were found as previously (TC-DNA shotgun sequencing, NICE 2015), with most of the OTUs being rare in all samples, but with the minority of abundant OTUs representing the bulk of the sequences (Table 10). When comparing spatiotemporal variation, from March, April and June, along drifting ice, the number of reads was generally higher in June samples for the total and abundant prokaryotes. For the rare prokaryotes, April and June samples displayed higher values than March. For the number of rare OTUs, there was a general non-significant increase from March to June (Table 11). Shannon diversity measures were similar across space and time, except for April, in the total biosphere, where there was an increase in diversity. When comparing different sampling depths, all values were similar, with no significant differences. When comparing water masses, alpha metrics were significantly different for each water mass, considering all abundance types, with PSWw and AW having usually higher values.

Table 10. Alpha diversity for the NICE 2015 dataset sample, for prokaryotic data identified from 16S rRNA gene amplicon sequencing. Values for alpha diversity are the Shannon index, number of OTUs and number of reads (of 16S rRNA gene sequences used for taxonomy). For each sample, communities are divided in total, abundant and rare, according with the OTUs abundance, using the rarity threshold of 1054 reads per sample, from section 3.1.3.

Sample	Abundance	Shannon index	Number of OTUs	Number of reads
NICE_1	Rare	3.844	182	15707
	Abundant	2.009	16	139105
	Total	2.523	198	154812
NICE_2	Rare	3.899	214	18969
	Abundant	2.276	22	166227
	Total	2.773	236	185196
NICE_3	Rare	4.065	204	14808
	Abundant	1.368	8	134741
	Total	1.958	212	149549
NICE_4	Rare	4.403	355	21733
	Abundant	2.329	18	88706
	Total	3.233	373	110439
NICE_5	Rare	4.302	283	25100
	Abundant	2.169	18	124744
	Total	2.978	301	149844
NICE_6	Rare	4.093	246	18879
	Abundant	2.586	27	154203
	Total	3.095	273	173082
NICE_7	Rare	4.204	289	24530
	Abundant	2.499	26	406976
	Total	2.814	315	431506
NICE_8	Rare	4.27	323	21461
	Abundant	2.716	33	368549
	Total	3.015	356	390010
NICE_9	Rare	3.623	193	7893
	Abundant	1.872	15	158954
	Total	2.145	208	166847

To understand which variables were determining community composition for the rare biosphere in the 16S rRNA gene amplicon sequencing data of the NICE 2015 dataset, ordination analysis was used. The gradient length, as measured by DCA, was 2.4, supporting selection of linear ordination methods. According with the PCA plot (Figure 17), there was no obvious pattern in favor of a specific variable. Samples from June were further away from March and April. Samples from April were very distant from each other, while PSW water masses were generally close to each other as well as bottom samples. In June, each depth corresponded to a different water mass, which might be responsible for the differences across June samples. By changing season, different depths corresponded to different water masses, thus the combination of all variables explains better the patterns found.

Table 11. Significance values for alpha diversity differences across samples compared in the NICE 2015 dataset, for prokaryotic data identified from 16S rRNA gene amplicon sequencing. One-way ANOVA test for total, abundant and rare communities, for alpha metric values. Rare and abundant communities were divided using a threshold of 1054 reads per sample, from section 3.1.3. P-values<0.05 are in **bold**.

Variables compared	Shannon index	Number of OTUs	Number of reads	Abundance
Date/site	0.29	0.144	0.509	Rare
Depth	0.509	0.44	0.142	
Water mass	**2.20E-06**	**0.000190**	**1.30E-07**	
	Shannon index	Number of OTUs	Number of reads	
Date/site	0.299	0.351	0.053	Abundant
Depth	0.444	0.518	0.759	
Water mass	**0.00340**	**0.000320**	**0.000120**	
	Shannon index	Number of OTUs	Number of reads	
Date/site	0.141	0.138	0.069	Total
Depth	0.321	0.422	0.718	
Water mass	**0.00320**	**8.40E-05**	**9.60E-05**	

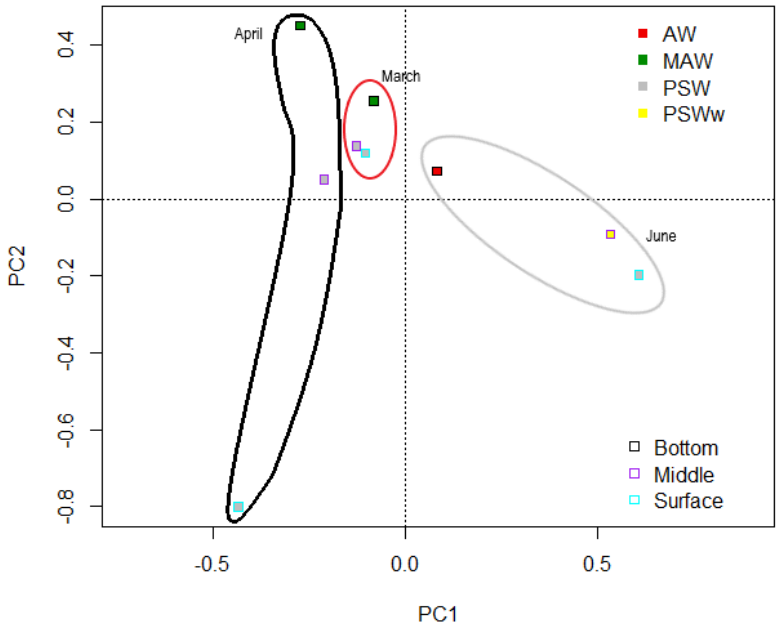

Figure 17. PCA of 16S rRNA gene amplicon sequencing data from the NICE 2015 dataset, for rare prokaryotic data. Samples are illustrated according with water masses (squares colored in red for AW, green for MAW, grey for PSW and yellow for PSWw), sampling depth (square border in black for Bottom, purple for Middle and cyan for Surface) and different areas are highlighted according with the date of sampling (red line for March, black line for April and gray line for June).

As in the TC-DNA shotgun sequencing data of the NICE 2015 dataset, each prokaryotic OTU was labeled a type of rarity for the 16S rRNA gene amplicon sequencing data (Figure 18). Transient rarity was the most frequent type of rarity in all comparisons made but was approximately 10% more

frequent across time than across different depths, followed by permanent rarity and conditional rarity. As transient rare OTUs are always rare, the vast majority of rare OTUs are always rare with only a minority becoming abundant in changing conditions, a similar pattern found on the previous section 3.4.1. For illustrative purposes, two examples are highlighted in figure 17.

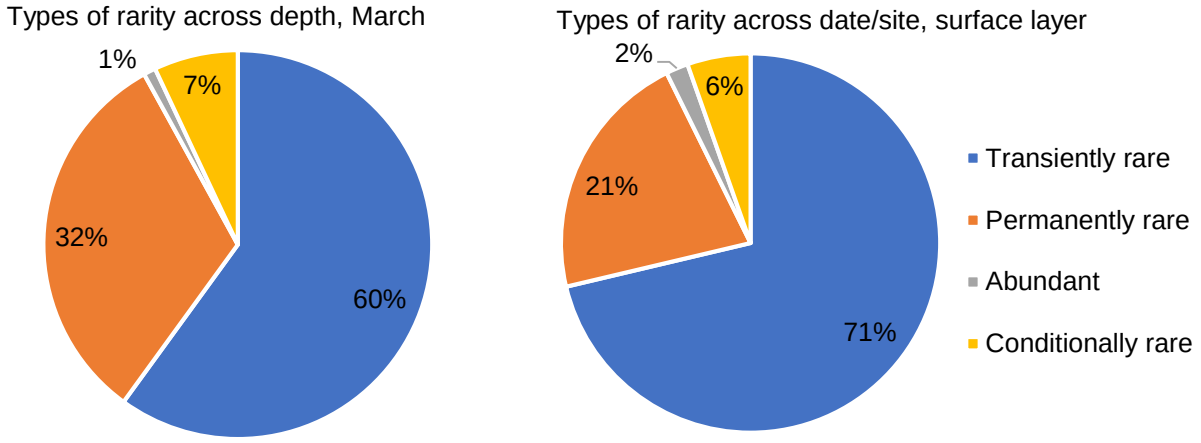

Figure 18. Percentage of prokaryotic rare OTUs (for each type of rarity) and abundant OTUs, identified in the 16S rRNA amplicon sequencing data from the NICE 2015 dataset. The types of rarity are calculated according with the variables compared: Across depth, for march on the left; across date/site, for surface on the right. The algorithm types.r was used according with Annex 3.

For the 16S rRNA gene amplicon sequencing, of the NICE 2015 dataset, the same circular visualization was made as before (section 3.4.1). In Figure 19, there were represented 30 different prokaryotic phyla, not counting unidentified bacterial and archaeal phyla. The inner heatmap, despite having low resolution, illustrated that most of the prokaryotic diversity is at low abundances, independently of the sample, as previously calculated (Table 10). Within more abundant prokaryotic phyla, such as Bacteroidetes and Proteobacteria there were a few classes with high abundance, but most of the other classes were rare. Also, from the heatmap it is immediately suggested that transient rarity was present, because there were many white spots (indicating absence) followed by light colored spots (indicating low abundance). A common pattern was that phyla with high numbers of transiently rare OTUs also had high numbers of permanently rare OTUs, phylogenetically close. CRT were mostly present in phyla such as Crenarchaeota, Euryarchaeota, Actinobacteria, Bacteroidetes, Cyanobacteria, Proteobacteria, SAR406 and Verrucomicrobia. CRT, in this dataset, seem to be phylogenetically closer to other PRT, despite sometimes being associated with abundant OTUs. Some phyla were represented mostly by one type of rarity, for example, Firmicutes is mostly transiently rare. Other phyla were very rich in all types of rarity, such as Proteobacteria. This pattern may derive from the number of different classes within each phylum, since phyla like Bacteroidetes possess classes displaying all types of rarity (e.g. Flavobacteria) while others display only one type (e.g. OTUs from the class Bacteroidia are transiently rare or absent).

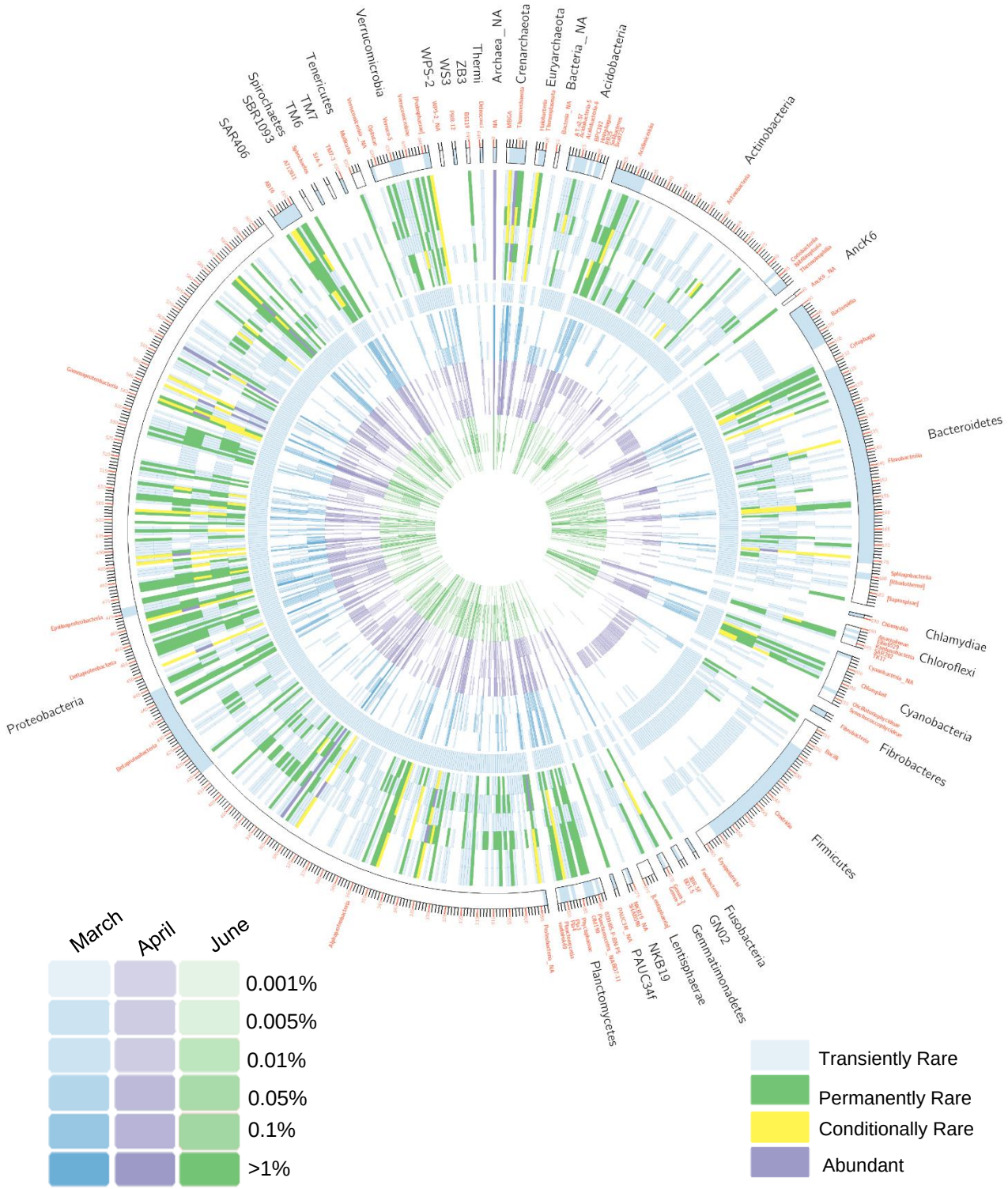

Figure 19. Circular visualization of the 16S rRNA gene amplicon sequencing data from the NICE 2015 dataset. Figure produced using Circos software. Read the Circos Figure clockwise, from outside to inside. All OTUs found at least in one sample are numbered from 1 to 697. OTUs are organized at phylum level by separated groups, with phylum labels in black. Within phylum, different classes have different colors, class labels are in red. The outer heatmaps represent the types of rarity of each OTU, if rare, or abundant or absent, with the color code on the bottom right. White spaces in the heatmaps represent absence. The color gradient for the types of rarity is highlighted in the superior right side of the Figure. The inner heatmaps represent abundance, where white spaces represent absence. The color gradient for abundance is on the bottom left side of the gradient. This Figure is better analyzed on the virtual version than the printed online.

3.4.3 Comparing TC-DNA shotgun sequencing data with 16S rRNA gene amplicon sequencing data for the rare prokaryotic diversity, from the NICE 2015 dataset

With data from TC-DNA shotgun sequencing and 16S rRNA gene amplicon sequencing for the same environmental samples (NICE 2015 dataset) and focusing on the prokaryotic data, it was possible to compare how these different approaches behave in the description of the prokaryotic rare diversity (Table 12 and Figure 20). The number of marker gene reads was much higher in the 16S rRNA gene amplicon data, due to the targeted amplification by PCR. However, that did not translate into more OTUs, neither into the OTUs abundance equilibrium (Figure 20). In fact, the number of total and rare prokaryotic OTUs were sometimes superior in the TC-DNA shotgun sequencing data than in the 16S rRNA gene amplicon data (Figure 20). Also, the TC-DNA shotgun sequencing data delivered higher Shannon diversity values for the total and rare biosphere. For the abundant OTUs, there were more OTUs in the 16S rRNA amplicon data and no significant differences in Shannon indices (Table 12).

Table 12. Significance values of the differences across the 16S rRNA gene amplicon sequencing vs TC-DNA shotgun sequencing data, from the NICE 2015 dataset. One-way ANOVA test comparing the samples from the NICE 2015 dataset, across both metagenomic strategies. Rare and abundant communities were divided according with the metagenomic strategy, for the TC-DNA shotgun sequencing, 42 reads per sample (section 3.1.3) and 1054 reads per sample (section 3.1.3) for the 16S rRNA gene amplicon sequencing data. P-values<0.05 are in **bold**.

Variables compared	Total	Abundant	Rare
Shannon index	**3.50E-06**	0.492	**6.50E-08**
Number of OTUs	0.832	0.0483	0.629
Number of reads	**5.30E-05**	**0.00011**	**3.60E-08**

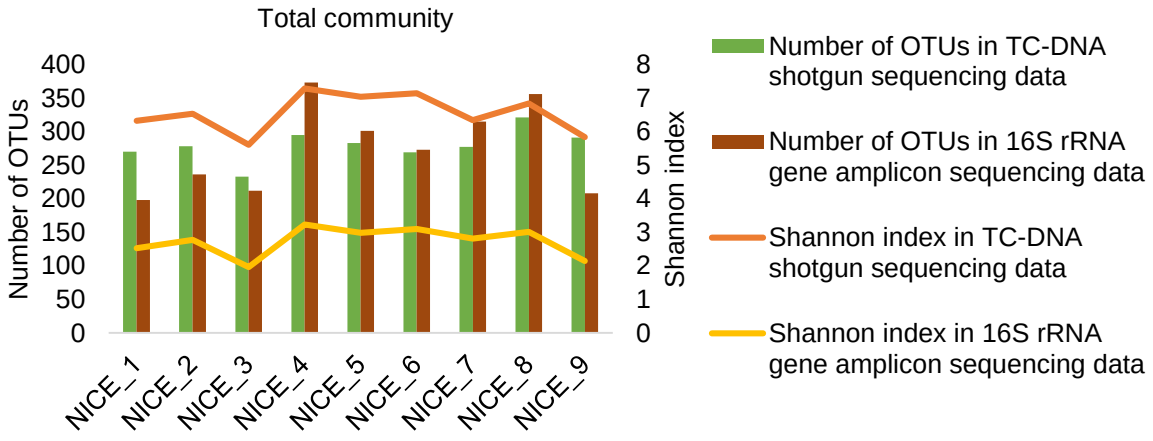

Figure 20 is continued on the next page.

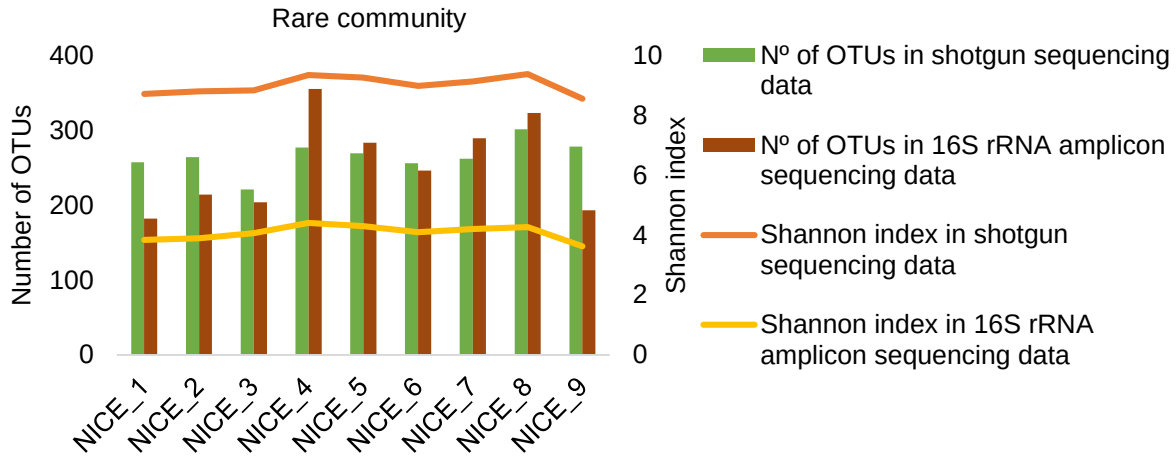

Figure 20. Comparison of the number of OTUs and Shannon index for TC-DNA shotgun sequencing and 16S rRNA gene amplicon sequencing data, from the NICE 2015 dataset. Bar plots illustrate the number of different OTUs, and the line plots illustrate the Shannon index value. The top figure is relative to the entire community, whereas the bottom figure is relative to the rare communities identified in the 16S rRNA gene amplicon data (using the 1054 reads per sample threshold, from section 3.1.3) and the TC-DNA shotgun sequencing data (using the 42 reads per sample, from section 3.1.3).

4. Discussion

The microbial rare biosphere is a cryptic community, elusive to traditional methods, and its real dimension was discovered due to the advances of HTS. Since the pioneer work by Sogin et al. (6), the number of studies on this topic increased year by year, resulting in the recognition of the relevance of the microbial rare biosphere in ecosystem functioning (12), modeling (13) and on the understanding of metabolism at low abundances (9).

4.1 Definition of the microbial rare biosphere

By analyzing the literature cited in this work, focusing on the prokaryotic rare biosphere as assessed by 16S rDNA amplicon sequencing methods, it is evident that there is no coherence in the definition of the concept of microbial rarity. For instance, 67% of the studies cited used relative abundance thresholds, 3% used absolute abundance and 4% used both relative abundance and the absolute abundance equivalent thresholds to delineate rare communities. But approximately one third of the studies did not specify the definition used. Within the studies using relative abundances per sample, the most used threshold was 0.1% (9,11,22,24,29,34,35,68,77,80,83,84,100–102), followed by 0.01% (7,37,38,41,103–107) and 1% (14,36,128) and 0.001% (75,176). Other thresholds have also been used, including 0.005% (177), 0.004% (17), 0.02% (18) and 0.2% (15). Other studies have used whole dataset thresholds, for example, the usage of 0.01% per sample and 0.001% for the whole dataset (16,32,126) or 0.1% per sample and 0.006% for the whole dataset (178). The lowest thresholds, for example 0.001%, are sometimes equivalent to the singletons present in the dataset, depending on the sequencing depth of the marker gene used for taxonomic assignment. The overall range of thresholds is from 0.001% to 1%, thus a difference of up to three orders of magnitude. The lack of coherence in the statistical

delineation - and consequently theoretical definition - of the microbial rare biosphere is evident. Furthermore, there is no biological basis to use one threshold or the other (13). Some authors indicate that the rare biosphere is the long tail of the RAC, e.g. Pedrós-Alió (3), however without presenting methods to decide where the long tail begins.

The lack of coherence indicates the need for a set of guidelines to define microbial rarity, and the use of arbitrary thresholds reflects the need for a biological interpretation of the concept itself. The first attempt to solve the latter problem was based on the MultiCoLA algorithm (108), once adapted to define microbial rarity in a non-arbitrary way (13). The method is based on the comparison of different thresholds, each one corresponding to a truncated community, with the original (not truncated) community. In theory, if the rare community is distinct from the abundant community, then it will be very dissimilar to the original community. In this thesis, MultiCoLA did respond differently to different sequencing approaches (16S rRNA genes from TC-DNA shotgun sequencing data, 16S rRNA gene amplicon sequencing, with and without library sizing, 18S rRNA gene amplicon sequencing) in the EMOSE 2017 dataset, in terms of absolute value thresholds per sample. Those values, when converted to their relative abundance per sample equivalents, were all close to 0.1% (Table 3). Except for the 16S rRNA gene amplicon sequencing data with library sizing for larger sequences (MetaB 16S large), because that subset of data was expected to reflect incorrectly the prokaryotic diversity. Also, across independent datasets such as the *Spongia officinalis* 2014 and the NICE 2015 datasets, which employed TC-DNA shotgun sequencing to characterize microbial communities, the values obtained for MultiCoLA were similar (Figures 6 and 8). In this regard, MultiCoLA can be useful, because it delivers a specific definition, adjusted to each sequencing power of the marker gene used to assign taxonomy, in a consistent way. However, it did not prevent the usage of a “choice step”, in this study, because the correlation values behaved in a monotonous way, with no drastic changes (Figures 4,6 and 8). Even though correlations went down with more stringent thresholds, reflecting different community structures, the point where the change begins, or the point where the rare community is defined, is dependent on a subjective choice, with no objective criteria. In practice, the MultiCoLA output is a set of possible thresholds and not a specific one, meaning that different researchers using the same output could select different thresholds. Thus, the MultiCoLA based thresholds that are obtained remain arbitrary. Also, when comparing how the algorithm worked within datasets, by separating different types of samples, in the *Spongia officinalis* 2014 dataset, the correlation values obtained varied in unpredictable ways, meaning that separating samples from the analysis influences the MultiCoLA results (Figure 7). When joining all *Spongia officinalis* 2014 dataset samples the results are as expected, the reason why this happens is unknown. In an unpublished study (personal communication from Xiu Jia, 2019), the same limitations in using MultiCoLA were highlighted for soil datasets, using 16S rRNA gene amplicon sequencing data.

Despite the problems listed above regarding the utilization of the MultiCoLA algorithm, for most datasets tested, the definitions obtained are not far from those found in the literature. Thus, our results showed that, for practical purposes, the rare communities defined with MultiCoLA can be used for subsequent analysis. There were few exceptions. For 16S rRNA gene amplicon sequencing with library

sizing, from the EMOSE dataset, the sizing was done because the set of primers selected can amplify 18S regions, but 18S amplicons are larger than 16S amplicons. Thus, by separating the sequences by size, after library preparation, sizing is expected to improve specificity in prokaryotic diversity and reduce eukaryotic diversity in small-sized libraries. When applying MultiCoLA with both prokaryotes and eukaryotes (non-sized libraries), the definition will be influenced and have a different meaning in the prokaryotic and eukaryotic community. In fact, by using the library sizing approach, rare prokaryotic diversity was augmented in relation with rare eukaryotic diversity for the small library size, whereas the opposite happened for the large library size, as expected (Figure 5). By applying the rarity thresholds obtained for the prokaryotes and eukaryotes in separate, there are more rare prokaryotes in the small library size and more eukaryotic rare OTUs in the large library size (if those OTUs are considered truly rare, which is not the case, as previously discussed). Another problem with defining rarity after library sizing is well reflected in the MetaB 16S large rarity threshold, because it is several orders of magnitude superior to the others (both for prokaryotes and eukaryotes). This can be because the eukaryotic diversity overshadowed the prokaryotic diversity, as expected from the primer set (153). But library sizing was used to improve prokaryotic diversity and for that objective it worked, by selecting the small size and discarding the large size. The small library size threshold for rarity, when applied to prokaryotic sequences, is around 0.13% and for eukaryotes, is around 1%, meaning it can give a meaningful value for the rare prokaryotic biosphere and not for the rare eukaryotic biosphere, intended to be removed in the first place. The other exception was for the TC-DNA shotgun sequencing data from the NICE 2015 dataset, because the average relative abundance threshold, per sample, was 1.25%, superior to those found on the literature and is a result from the lower number of 16S rRNA gene sequences available. In this case, despite providing a threshold different from those in the literature, it remains coherent given the abundance values.

The ideal definition of the microbial rare biosphere should have consistent results across different datasets. Meaning, it should represent the same biological reality across independent datasets. For that, the sequencing power, technology and strategy should be taken in consideration. The fact that there are so many ways to define rarity is a reflection of the different methodologies and differences in sequencing power applied to study the microbial rare biosphere (13). One cannot assume *a priori* that a specific threshold will universally fit the rare biosphere, because the meaning of a specific threshold is dependent on community composition and different datasets have different diversity values. Despite that, the RAC figures always have the same pattern, namely, the long tail (3,51). The universal RAC shape is most probably the only safe assumption across any microbial community dataset, thus, the answer to define rarity may be on the development of a method to accurately calculate the beginning of the long tail, in a reproducible way. Furthermore, factors such as the utilization of a sample by sample threshold or for the whole dataset should be explored, as well as the use of relative abundance and absolute abundance thresholds. Besides the statistical definition of the rare biosphere, the methodological efficiency of the recovery of the taxa present in the environment is relevant. For example, if there are missing taxa, the total community described is not truly representative of the real diversity, thus influencing the estimation of rarity. Also, the recovery process should not be biased towards certain taxa, as it can have dramatic changes in the measured relative abundance (49).

4.2 Methods for microbial rare biosphere recovery

High sequencing power is necessary to fully grasp the rare biosphere diversity (3,13,83), thus favoring the usage of 16S rRNA gene amplicon next generation sequencing technologies, for the prokaryotic rare biosphere. With the TC-DNA shotgun sequencing approaches, the sequence recovery of the 16S rRNA gene is much lower, but by increasing sequencing power and accuracy, the TC-DNA shotgun sequencing approach can present different advantages. For example, the Candidate Phyla Radiation (CPR) was partially absent in 16S rRNA gene amplicon datasets, but was identified in TC-DNA shotgun sequencing datasets from a previous study (175). Also, our TC-DNA shotgun sequencing data from the *Spongia officinalis* 2014 and NICE 2015 dataset, suggests that CPR taxa are mostly rare (Figures 13 and 16). Considering this from the point of view of the definition of rarity, specifically for rare prokaryotes, 16S rRNA gene amplicon sequencing is arguably better in representing the universe sampled, due to more sequencing power, thus more percentage of diversity collected. But the TC-DNA shotgun sequencing approach is better at enabling a more harmonized identification of OTUs since PCR bias in the rare biosphere cannot be ruled out in amplicon sequencing datasets (49,140,179,180). Also, with TC-DNA sequencing, it is possible to assess functional information, e.g. Karimi et al. (158). Another advantage of TC-DNA shotgun sequencing relies on the ability to further use MAGs to study the biology of specific rare taxa (148), although metagenomic binning of contigs into genomes is easier for low complexity communities (181), and less efficient for low abundance microorganisms. In this work, when using TC-DNA shotgun sequencing, in both *Spongia officinalis* 2014 and NICE 2015 datasets, CPR taxa were identified as members of the rare biosphere (Figures 13 and 16). For example, the Candidate Phyla Levibacteria in the *Spongia officinalis* 2014 dataset (Figure 13) and the Candidate Phyla Saccharibacteria in the NICE 2015 TC-DNA shotgun sequencing data (Figure 16). However, these taxa were not found in the NICE 2015 16S rRNA gene amplicon sequencing dataset (Figure 19). The reason for this is probably related with the identification of CPR taxa with self-splicing introns in the 16S rRNA gene, with varying size and positions (175,182). Another possible reason is the bias within databases towards the sequences of previously cultured prokaryotes (183,184), resulting in primer design bias and consequent non-annealing of the primers.

By direct comparison of 16S rRNA amplicon and TC-DNA shotgun sequencing in the same samples of the NICE 2015 dataset, despite the difference in the number of total and rare reads used for taxonomic identification, there was not an accentuated difference in the number of different rare prokaryotic OTUs identified in both approaches (Figure 20). In fact, there is probably a bias towards more dominant prokaryotes in the 16S rRNA gene amplicon sequencing data, as the Shannon index is lower in this case. The patterns of the rare prokaryotic community differ across both approaches. Specifically, the March and April rare communities are clustered together from the TC-DNA perspective (Figure 14) but represent different groups from the 16S rRNA gene amplicon sequencing perspective (Figure 17). This can be explained by the lower Shannon index in the 16S rRNA gene amplicon data, because it might be producing higher discrepancies of abundance across different taxa, thus leading to clustering of different groups. Alternatively, it might be a result of under sampling of the rare prokaryotic biosphere in the TC-DNA shotgun sequencing data. Missing taxa due to lower sequencing power of the

marker gene can lead to missing patterns. This is further supported by the fact that differences across the environmental variables are more significant from the perspective of the 16S rRNA gene amplicon data set (Tables 9 and 11). Missing taxa might also result in overestimation of transient rarity, if it is defined as rare OTUs that are absent in at least one sample, as is evident from comparing Figures 16 and 19. Considering our data and the above arguments, this work argues that to study the prokaryotic rare biosphere, it is important to complement 16S rRNA gene amplicon sequencing with TC-DNA sequencing. The amplicon sequencing approach is more useful to identify most of the rare taxa and the TC-DNA sequencing approach is important to complement missing taxa, provide functional information and solidify biological interpretation, for example, through extensive analysis of MAGs and community functional profiles, when possible.

Besides the sequencing strategy, the methodologies for environmental DNA collection will presumably influence the description of the rare prokaryotic biosphere. Little is known about the effect of different seawater sampling methodologies on the recovery of the marine prokaryotic rare biosphere. For instance, in a study which addressed how different DNA extraction kits influence diversity estimates, it was found that the rare prokaryotic biosphere is more sensible to different DNA extraction protocols than the abundant prokaryotic biosphere (130). From the literature cited, the influence of different seawater filtering methods and filtered volumes on the view of the marine prokaryotic rare biosphere is missing. When comparing the scope of seawater filtering methodologies, in marine microbial rare biosphere studies using 16S rRNA gene amplicon sequencing, pre-filtration prior to prokaryotic cells filtration is commonly used to lower eukaryotic “contamination” (15,27,64,68,83,84,101,104,126,128). Alternatively, whole water filtration is employed (6,18,29,34,35,176). Pre-filtration methodologies used in the literature include the usage of a mesh with pore size of 200µm (84,101,126,128), or membranes for pore sizes of 3µm (15,41,68,83,104) or 0.8µm (27). The bacterial cells, in the marine prokaryotic rare biosphere studies, independently of pre filtration or not, are mostly filtered with the Sterivex filter unit (15,27,29,41,68,83,84,101,104,128), but membrane filter units are also used (6,34,35,126,176), both with pore sizes of 0.22µm. Regarding volume, marine prokaryotic rare biosphere studies use: less than 1L (126,176), 1L (6,18,34), 2L (101), 5-7L (15,41,104), 20L (68) and 170L (35); Some studies use a range of volumes instead, such as 5/6L to 15L (83,128). From this overview, except for 170L (35), the seawater volumes filtered for the study of the marine prokaryotic rare biosphere range from less than 1L to up to 20L. Thus, regarding the seawater volume, it is relevant to understand if there are significant differences across the range of volumes used in the literature. It is also important to know if it is necessary to filter more seawater to have a better view of the marine prokaryotic rare biosphere.

The 16S rRNA gene amplicon sequencing data from the EMOSE 2017 campaign (without library sizing), was used to explore how the view on prokaryotic rare diversity changes with increasing water filtration volume. Despite not representing entirely the complete range of values used in the literature, since volumes filtered within the EMOSE project ranged from 2.5L to up to more than 500L, our results suggest that volume was not an important factor determining the rare prokaryotic diversity in seawater. For example, when using Sterivex filters with whole water filtration, from 2.5L to 10L, the diversity values (number of rare OTUs, number of rare reads and Shannon index) did not differ significantly (Table 5).

When comparing a broader range of volumes, from 10L to more than 500L, the most significant differences were on the number of reads, possibly because more cells are collected, but the diversity itself, as measured for species richness (equivalent to number of rare OTUs in this work) and species equilibrium (as measured by the Shannon index) did not change significantly (Table 5). To discard the possibility that the increase in number of reads is due to volume and not sequencing power, a rarefaction analysis should be added in future work. Also, from the point of view of community composition, as analyzed by ordination analysis (Figure 10), the rare communities from larger volumes overlap with the rare communities from smaller volumes. Meaning that more volume, with more cells, does not necessarily lead to a better representation of the rare biosphere from the sampled universe. Thus, this work supports the utilization of the current range of seawater filtered volumes in the literature, since volumes superior to 100L do not compensate for the extra time and costs. Regarding the utilization of membrane large filters or Sterivex filters, with pore size of 0.22μm and whole water filtering, results are similar (Table 5). This is expected, because the pore size is the same, meaning that the same rare communities are being filtered, without being significantly influenced by the physical properties of the filter. Conversely, by changing the pore size, it is expected that different communities are retrieved. For instance, different size fractions, sometimes equivalent to pre-filtrations in the literature (when pre filtration is around 20 μm), resulted in significantly different communities (Table 5). When analyzing the medium size fraction (pore size 3μm – size fraction between 20μm and 3μm) and large size fraction (pore size 20μm - size fraction > 20μm), it was verified that these fractions select for different prokaryotic rare communities. This is supported by both the values of diversity (number of OTUs, number of reads and Shannon index, Table 5) and by the ordination analysis, where there is a clear grouping according with the pore size, independently of the remaining sampling variables (Figure 10). Then, caution should be taken when comparing datasets with or without pre-filtration steps.

The most challenging aspect to explain the size fractioned diversity is not the existence of different communities, but why the medium and large size fractions have an excess of rare prokaryotic diversity compared with the small fraction. One hypothesis to explain such finding is that the high fractions also include rare host-associated prokaryotes, prokaryotes associated with rare hosts and/or particulate matter-associated prokaryotes. Furthermore, the marine rare biosphere is thought to be an important component of most host-associated microbiomes (e.g., Taylor et al. (94)), and host-associated microbiomes presumably include rare CPR (e.g. Wu et al., 2011 (184)), representing a significant component of biodiversity, as is well illustrated by Hug et al., 2016 (185). Thus, to have a full picture of the marine prokaryotic rare biosphere, it might be important to include eukaryotic cells. The presented data is not enough to accept or reject the hypothesis that the excess of rare prokaryotic diversity in the pre filtrate from EMOSE 2017 dataset is due to host-associated relationships, but the high unknown diversity of host associated microbes is in favor of such hypothesis.

This finding suggests that it is better to use whole water filtration without pre filtration, since pre filtration is sometimes used to remove eukaryotic ‘contamination’. Thus, the use of pre filtration steps can hide an underexplored component of the prokaryotic rare biosphere. The exception is when the objective relies solely on the free-living marine rare prokaryotes, in this case the pre filtration step might

be useful to rule out rare prokaryotes associated with eukaryotes. But in the latter case, the existence of an excess of prokaryotic rarity in the pre filtrate should be analyzed, as some apparently free-living, rare prokaryotes might be host-associated prokaryotes that got randomly dispatched from the original host, as e.g. sponge symbionts were previously found as rare prokaryotes in geographically distant areas (94).

4.3 Marine prokaryotic rare biosphere ecology

There are multiple questions related to the rare biosphere ecology that are still far from a complete answer, such as: (i) How does the prokaryotic rare biosphere behaves in the ocean? (ii) Are those rare communities dispersed everywhere or represent distinct communities in different sites? (iii) The abundance is permanent or varies over time? and under what processes? (iv) Is the rare biosphere ecology modeled by stochastic or deterministic mechanisms? The relevance of understanding these mechanisms relies on the ecological functions of the prokaryotic rare biosphere (12), which provides resistance and resilience to ecosystems by working as a seed bank (68), and/or through disproportionally high activity (22), and/or by transferring functional genes (24). In addition, as illustrated by the long tail of the RAC, most of the known biodiversity is rare, both for microbes and non-microbes (3,51), arguably representing an important pool of genetic diversity (3,6,8,40). In this work, across independent datasets (EMOSE 2017, NICE 2015 and *Spongia officinalis* 2014) and within different samples of each dataset, this pattern was universal, a minimal number of rare prokaryotic reads corresponds to most of the prokaryotic OTUs from the community. The Shannon index indicated that the rare communities have more homogenous abundances than the total and abundant communities, with similar values (Table 6, 8 and 10). Thus, when the Shannon index is used to study the equilibrium of diversity it fails to consider the rare biosphere, this is because the index weights the relative abundance of OTUs, since the abundant OTUs are much more abundant than the rare OTUs, the latter ones much down-weighted.

Soon after the first prokaryotic rare biosphere description (6), it was hypothesized that all microbes can be found everywhere, but at different abundances, bringing back to life the Baas Becking dictum “Everything is everywhere, but the environment selects” (21,59). Not because of the existence of the rare biosphere, but because it revealed that most of the microbial diversity was not collected with traditional methods, so one could not rule out the hypothesis that all OTUs are present in any environmental sample, but are not identified due to methodological limitations (6,59). Despite of the difficulty in proving or disproving the dictum, if it applies to the ocean, where prokaryotes are dependent on external factors for movement, it implies the existence of unlimited dispersal associated with stochastic patterns (5). Moreover, it would suggest the existence of a seed bank, able to respond to changing conditions (59). But as described in the introduction section of this work, despite the existence of a seed bank (68), there is evidence that everything is not distributed everywhere, as there are biogeographic distributions associated with deterministic patterns for both prokaryotes (41), and microeukaryotes (16). Notwithstanding, nowadays it is consensual that both deterministic and stochastic patterns model microbial community dynamics (126) (for a comprehensive review, see Zhou and Ning (5)). This reflects the existence of different types of rarity and respective mechanisms across different

variables (13). In this context, and in agreement with the findings of this thesis (see discussion below), the prokaryotic rare biosphere is now known to be influenced by water masses across different oceans (27,41,128). The marine prokaryotic rare biosphere distribution is also thought to be highly influenced by host-associated interactions (78).

The NICE 2015 dataset allows to describe the behavior of the prokaryotic rare biosphere through spatiotemporal variations, because sampling was performed with a vessel fixed on ice, drifting along time, with different space coordinates (160). Other studies in the Arctic ocean indicate that there are biogeographical patterns of the prokaryotic rare biosphere, due to different water masses (27,41) and it was suggested that the relative abundance of the prokaryotic rare OTUs behaves in the same way through time (41). When comparing different depths (surface, middle and bottom) and different spatiotemporal coordinates (from March to June, in drifting ice), there are not significant differences in the prokaryotic rare biosphere diversity (Table 8 and 10), in accordance with Vergin et al. (41). But, when comparing different water masses, there are significant differences in the values of diversity. A shift in community composition was observed in June (Figures 14 and 17), because of the inflow of warmer Atlantic waters (AW) to the deeper layers. This inflow of AW works as a source of heat to the Arctic ocean, leading to the melting of ice, responsible for the warm Polar Surface Water (wPSW), thus forming three layers of water: PSW, PSWw and AW (161). On the other side, during March and April, the surface and middle layers of water correspond to PSW, thus explaining why the prokaryotic rare communities of March and April are closer to each other (Figures 14 and 17), except for the deeper layer, corresponding to MAW, resulting of the mixing with AW in the previous seasons (160,161). The main variable influencing the marine prokaryotic rare biosphere in the Arctic ocean is probably the different water masses at play, corroborating existing literature (27,41).

A deeper look into the behavior of the prokaryotic rare biosphere, in the Arctic ocean, requires the distinction of the different types of rarity. From insights gained by previous studies of the prokaryotic rare biosphere in the same ocean, it is expected that most rare prokaryotic OTUs remain always rare (41), with a minor group of CRT (7,27). In this work, most of the prokaryotic rarity in the Arctic ocean was associated with transient rarity, followed by permanent rarity and a minor fraction of CRT (Figures 15 and 18). Despite that, the methodology used in this study to distinguish the different types of rarity (types.r, in Annex III) is possibly overestimating transient rarity. Also, transient rarity might be a result of under-sampling due to lower sequencing depth. But the concept of transient rarity is within the concept of permanent rarity, the difference is that permanently rare OTUs remain viable through changing conditions, whereas transiently rare OTUs grow and eventually decay into extinction, with changing variables (8,13). Furthermore, in the previous studies of the prokaryotic rare biosphere, at different sites and through different time periods, at the same ocean, transient rarity would probably be included in the permanently rare group (27,41). It is less relevant to identify each type of rarity than it is to identify and explore the mechanisms explaining the rarity patterns.

Dispersal limitation, in the NICE 2015 dataset, can explain the behavior of the abundance of the prokaryotic rare biosphere, because water masses were identified as relevant environmental drivers of the communities composition (27,41). Since prokaryotic cells movement is dependent of the water

current, if there are no water currents determining the biogeography of the cells, they will be diluted everywhere, by unlimited dispersal. However, if there are water masses at play, then they work as physical barriers separating different zones and, thus, they are equivalent to dispersal limitation. Galand et al. (41) considered water masses deterministic selective pressures, in principle contrasting with dispersal limitation that is considered stochastic. But as explained previously, if the cause of the effect is known (as it is the case of water masses within the NICE 2015 dataset), then it is considered deterministic (5). There is one remaining problem in this framework, because transient rarity is assumed to be stochastic and permanent rarity is assumed to be deterministic (13). This is because one cannot assume why each transiently rare OTU disappears across changing conditions, whereas it can be assumed that, by maintaining the same selective pressures over time, all permanently rare OTUs remain rare. Thus, it is possible to assume a constant selective pressure (homogenous selection in Jia et al. (13) terminology), favoring permanent rarity in the Arctic ocean. Combining these results with the community assembly theory, applied to the microbial rare biosphere (13), it is possible to extrapolate some important mechanisms in the Arctic ocean. One hypothesis, from our data and partially supported by previous results (27,41), is that the water masses randomly transport different cells. Cells that are not well adapted to the new environment eventually die and the water current will have the same effect on the distribution of the cells within the same water mass (because it is just a transportation effect). That would be the stochastic component. The deterministic component would be on the group of conditions that remain approximately constant through the spatiotemporal variation, within the water mass. This homogenous selection would promote permanent rarity. As highlighted before, transient rarity is a concept encompassed by the permanent rarity concept, because a transiently rare OTU is always rare, the difference is that it eventually disappears. Thus, because the significant majority of OTUs were identified as permanently rare and transiently rare, the following hypothesis might hold: stochastic mechanisms distribute the rare prokaryotes and deterministic mechanisms determine which ones subsist at low abundances and which ones disappear. To test this hypothesis thoroughly, it would be necessary to improve the method of identification of the types of rarity, to avoid overestimation of transient rarity. Then, it would be important to compare different methodologies to define each type of rarity. For instance, the method proposed by Jia et al. (13) is based on community assembly theory, which decides the types of rarity from the perspective of the community. Contrarily, in this work the algorithm types.r (Annex III) decides the type of rarity from the perspective of each individual OTU, and that might present caveats when modeling assembly mechanisms. The community assembly-based approach by Jia et al. (13), relied on ecological modeling perspectives (186–191). Consequently, it provides a more deductive and solidified method to disentangle the mechanisms debated on here.

It is noteworthy that there are other sources of movement of cells not identifiable in the NICE 2015 dataset. For example, the transportation of rare prokaryotes by hosts that travel long distances, or the recruitment or sinking of microbial diversity in sessile hosts (78). The latter example is stochastic from a predictive point of view, because, despite that the cause of the movement is known, what cells will be transported to, and where, will appear as a random process. But considering the active recruitment of prokaryotes due to host-microbiome mechanisms preserved by natural selection (192), deterministic processes also play important roles for the recruitment of rare species.

The sustaining of the rare hypothesis (78) explains the distribution of rare prokaryotes from the perspective of the inherent diversity of host-associated microbiomes. This framework is theoretically coherent, highlighting the relevance of the marine hosts microbiomes in the study of diversity, but is hard to fully prove with the current methodologies. For example, most studies on host associated rare prokaryotes focus on sessile animals, e.g. corals (43,96) and sponges (75,79,94,100). And most studies usually focus on the process of symbionts transportation to the host, rather than the opposite view, meaning, how hosts help the maintenance of a widespread and diverse rare biosphere (78).

There are insights into how sponges (as a model example of sessile host organism, with a diverse microbiome) can contribute for rare prokaryotic community assembly. Seawater works as a reservoir of sponge symbionts that remain viable and rare outside the host until being filtrated by the sponge (94,100). This is further supported by the finding that abundant sponge microbial symbionts are essentially either generalists or specialists (192) and by the finding that some rare microbial symbionts are species-specific (75). Furthermore, the sustaining of rare hypothesis (78), regarding the contribution of sessile macro organisms, fits well with the recognition that sponges, for example, represent an important reservoir of diversity (192). In fact, there were 44 prokaryotic phyla alone in the *Spongia officinalis* 2014 dataset analyzed in this work (158). Similarly, other studies have identified 32 different phyla in sponge-associated microbiomes (75) and a later study with higher sampling effort, found 41 different phyla (192). In this study, using *Spongia officinalis* 2014 dataset, virtually all prokaryotic phyla host rare OTUs, similar to Reveillaud et al. (75), where 20 phyla out of 32 have rare OTUs. In the latter study, it was also highlighted that most phyla with rare OTUs are CPR (75). This stresses the importance of adding the host-associated rare microbial diversity into the picture of the global rare biodiversity.

Marine sediments are an understudied component in the tradeoff of prokaryotic OTUs and sessile hosts (158). It was suggested that sponge cellular shedding could work as a source of sponge associated microbial OTUs to the sediment environment (158). If specialized (host-associated) OTUs sink in the sediment, then it is expected that they become rare in the sediment. It was also suggested that particle intake by sponge (193) can justify the existence of shared OTUs between sediment and sponge associated communities (158). From a cyclical point of view, specialized OTUs in the sponge microbiome can sink in the sediment, live as particle-associated rare OTUs, and re-enter the sponge environment through particle intake of the sponge. The analysis of the rare biosphere from this work, that use the same dataset as Karimi et al. (158), but with a more recent version of the bioinformatics analysis (136), indicated that seawater and sponge tissue have more similar rare community compositions with each other than with sediment (Figure 12). This might reflect sponge filtration of seawater, the influx of water brings randomly selected cells for the sponge microbiome, once there, they can adapt (becoming abundant or remaining rare) or disappear. This explanation is supported by the findings of Hardoim et al. (79), where they found a stable community over time, for the dominant symbionts, but a rare community characterized by many transiently rare OTUs and some permanently rare (symbiotic) OTUs. The latter ones were hypothesized to be functionally redundant relatively to their abundant neighbors. The distinct rare communities found between sediment and other samples (water and sponge), and within sediment replicates (Table 6, figures 12 and 13), might be a consequence of highly diverse sediment samples. Sediments were previously shown to be more diverse than seawater

and sponge tissue, for the overall community (158). If the sediment diversity results from particle and cell sinking, then it is the result of stochastic mechanisms and the sponge specific OTUs sinking there, from sponge cell shedding, would be predicted to be transiently rare taxa. This work did not use the types.r algorithm in the sponge dataset, to confirm the proportion of transient and permanently rare taxa, because the algorithm is probably overestimating transient rarity and is better adapted to the NICE 2015 dataset. Despite that, from the circular visualization (Figure 13), there is evidence for the transfer of transiently rare taxa from the sediment and seawater to the sponge. The same figure also reveals that most candidate phyla are present in sediments and are always rare, except for the Candidatus Poribacteria, abundant in sponge tissue samples (194). Another pattern is that some shared OTUs are phylum-specific, for example, the phylum Bacteroidetes is essentially shared across sediment and seawater, with a few rare OTUs sporadically occurring in the sponge samples (Figure 13). Important insights could also be gained from the analysis of Venn diagrams (Figure 11), where the number of total, rare and abundant prokaryotic OTUs are counted for shared and specific OTUs across samples: the majority of the assigned prokaryotic OTUs are sediment-specific, in agreement with the higher diversity found in sediment samples (158). Whereas most seawater prokaryotic OTUs are shared with sediment (Figure 11), possibly because of the random sinking of cells in the sediment layer. When looking at the rare component, the number of specific prokaryotic OTUs increases in both sediment, seawater and sponge tissue. From an analytical point of view, it is noteworthy that if one OTU is abundant in sediment, for example, and rare in the sponge tissue, then it will be classified as sponge tissue specific in the rare biosphere subset. Thus, OTUs that are shared in the total community and specific in the rare community are inferred to be CRT.

Integrating both the results from this work and the cited literature in the framework of the rare community assembly mechanism: Influx of seawater and sponge tissue cells shedding into sediments randomly transports prokaryotes across different types of environment (sponge tissue, sediment and seawater). This stochastic component explains the high numbers of transient rarity in the sponge tissue. The deterministic component is within each environment, where a group of conditions are maintained through time, resulting in a constant selective pressure, that allows some of the randomly distributed cells to persist. For example, permanently rare symbionts in the sponge tissue, that are viable and with possible functional redundancy (79). Regarding CRT, they result from deterministic mechanisms, in this context they can remain viable in the surrounding, non-optimal environment, and wait to (randomly) get in the optimal environment, where they are able to grow. Thus, CRT, in the host-associated landscape, can be considered opportunistic, whereas dominant symbionts are generalists and specialists (192). Thus, as in the water masses from the Arctic ocean, stochastic mechanisms distribute prokaryotic cells and deterministic mechanisms decide which ones remain rare, grow abundant or disappear.

5. Conclusion and future perspectives

The research performed in this study allowed to understand the phylogenetic diversity of the marine prokaryotic rare biosphere from the perspective of both stochastic and deterministic mechanisms. It is important to stress that different definitions of rarity and different methodologies to assess the marine prokaryotic rare biosphere will originate different types of rarity. Thus, it is relevant to understand the

concept of rarity, to define it in a biologically meaningful way and to know the best methods to recover the microbial rare biosphere.

MultiCoLA was proposed to solve the problem of the rarity definition (13), but in this work it was shown to fail in providing a non-arbitrary definition. Despite that, it can adjust to different sequencing approaches. A recently proposed method is based on the calculation of the beginning of the RAC “long tail” (personal communication with Xiu Jia, 2019). By testing that method, as well as other possible methods, combined with an understanding of the biological meaning of each rarity definition already used in the literature, it will be possible to establish guidelines for a comprehensive, coherent and biologically meaningful definition of the rare biosphere.

In this study, the missing gaps regarding methodology were assessed, specifically, the differences across TC-DNA shotgun sequencing and 16S rRNA gene amplicon sequencing and between different seawater sampling methodologies. This is relevant, because the definition of rarity – regardless the method employed - will only work if the sampled community is representative of the universe and collected in a non-biased way. Regarding the sequencing strategy, the SSU rRNA gene amplicon-based approach is the best for a good representativeness of the sampled universe, but the TC-DNA shotgun sequencing is better to have a non-biased view of diversity. Besides that, by crossing the data acquired with both strategies, it is possible to gain a better view of the rare microbial diversity in cause and answer functional questions. This work also provides some guidelines to seawater sampling, in the context of the rare biosphere: (i) Increasing volume results in more collected cells, but the representativeness of rarity does not change enough to justify the extra time and cost associated with volumes superior to 100L; (ii) The usage of large membrane or Sterivex type filters with the same pore size, is indifferent; (iii) The usage of pre-filtration steps might omit an important component of the rare prokaryotic biosphere, specifically the one that is host-associated, thought to represent a heavy component of the rare biosphere biodiversity; (iv) Seawater pre-filtration steps should be used only in studies focusing in the free-living marine prokaryotes, but the excess of rare prokaryotic diversity in the pre filtrate should be analyzed, to identify possible false free-living prokaryotes. In the future, by comparing these results with the mock communities from the same dataset (EMOSE 2017), it will be possible to quantify the bias in relative abundance produced by the different methodological steps, by using the model proposed by McLaren et al. (49).

When analyzing the prokaryotic rare biosphere ecology, in this work, it became evident that the Shannon index acutely down weights the rare diversity when applied for the total microbial community, due to the drastic discrepancies in abundance from abundant to rare taxa. Thus, other indexes should be explored, also in studies of the total biosphere, since the true diversity is masked by the dominant OTUs. To understand the ecological mechanisms of rarity, this work further suggests that it is necessary to classify each type of rarity. For that objective, the algorithm types.r (Annex III) was developed in here, but it overestimated transient rarity. Instead, it is suggested the use of the model based on community assembly by Jia et al. (13).

This work corroborated previous community assembly findings, interpreting how stochastic and deterministic mechanisms can simultaneously explain different components of the marine prokaryotic

rare biosphere, across different variables. For the Arctic ocean, water masses stochastically distribute cells. Those cells are under homogeneous selection due to the characteristics of each water mass, which in turn (deterministically) selects for a permanently rare biosphere, that is viable. The remaining rare OTUs that are not well adapted randomly disappear, thus explaining the transiently rare biosphere component. To confirm this point of view, cross analysis of data from different methodologies, sites and time periods will be necessary. Comprehensive comparisons of different approaches to define the types of rarity are also lacking in the literature. Once the RAC-based approach to define rarity is solidified, it will be possible to proceed with, and test, such analysis and have a complete spatiotemporal picture of the Arctic ocean, with relevance in the context of climatic change. For the host-associated interactions, again, stochastic mechanisms randomly distribute cells across environments (in this work, sponge tissue, seawater and/or sediment). Those taxa which are not well adapted to the new environment disappear, becoming transiently rare. Whereas other taxa can persist as permanently rare or grow abundant, thus having opportunistic strategies. To integrate the sustaining the rare hypothesis (78) with the community assembly theory (13), applied to the marine prokaryotic rare biosphere, it will be necessary to gather data on mobile macro organisms, besides the data on sessile-host associated rare prokaryotes. In the future, the analysis of the host associated prokaryotic rare biosphere, the *Spongia officinalis* host in here, should be complemented with the information regarding the types of rarity, in the framework of the community assembly theory (13), to test the mechanisms predicted in this study.

6. References

1. Caron D, Countway P. Hypotheses on the role of the protistan rare biosphere in a changing world. Aquat Microb Ecol [Internet]. 2009 Nov 24;57(3):227–38. Available from: http://www.int-res.com/abstracts/ame/v57/n3/p227-238/

2. Coveley S, Elshahed MS, Youssef NH. Response of the rare biosphere to environmental stressors in a highly diverse ecosystem (Zodletone spring, OK, USA). PeerJ [Internet]. 2015 Aug 20;3:e1182. Available from: https://peerj.com/articles/1182

3. Pedrós-Alió C. The Rare Bacterial Biosphere. Ann Rev Mar Sci [Internet]. 2012 Jan 15;4(1):449–66. Available from: http://www.annualreviews.org/doi/10.1146/annurev-marine-120710-100948

4. Reid A, Buckley M. The Rare Biosphere. Am Acad Microbiol Washingt [Internet]. 2011; Available from: https://pdfs.semanticscholar.org/0459/a32e85b50ee84a922efb757790adb272bd5d.pdf

5. Zhou J, Ning D. Stochastic Community Assembly: Does It Matter in Microbial Ecology? Microbiol Mol Biol Rev [Internet]. 2017 Dec 11;81(4):e00002-17. Available from: http://mmbr.asm.org/lookup/doi/10.1128/MMBR.00002-17

6. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored “rare biosphere.” Proc Natl Acad Sci [Internet]. 2006 Aug 8;103(32):12115–20. Available from: http://doi.wiley.com/10.1002/9781118010549.ch24

7. Shade A, Jones SE, Caporaso JG, Handelsman J, Knight R, Fierer N, et al. Conditionally Rare Taxa Disproportionately Contribute to Temporal Changes in Microbial Diversity. MBio [Internet]. 2014 Aug 29;5(4):1–9. Available from: http://mbio.asm.org/cgi/doi/10.1128/mBio.01371-14

8. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol [Internet]. 2015 Apr 2;13(4):217–29. Available from: http://www.nature.com/articles/nrmicro3400

9. Hausmann B, Pelikan C, Rattei T, Loy A, Pester M. Long-Term Transcriptional Activity at Zero Growth of a Cosmopolitan Rare Biosphere Member. Bailey MJ, editor. MBio [Internet]. 2019 Feb 12;10(1):1–16. Available from: http://mbio.asm.org/lookup/doi/10.1128/mBio.02189-18

10. Jones SE, Lennon JT. Dormancy contributes to the maintenance of microbial diversity. Proc Natl Acad Sci [Internet]. 2010 Mar 30;107(13):5881–6. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0912765107

11. Debroas D, Hugoni M, Domaizon I. Evidence for an active rare biosphere within freshwater protists community. Mol Ecol [Internet]. 2015 Mar;24(6):1236–47. Available from: http://doi.wiley.com/10.1111/mec.13116

12. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J [Internet]. 2017 Apr 10;11(4):853–62. Available from: http://www.nature.com/articles/ismej2016174

13. Jia X, Dini-Andreote F, Salles JF. Community Assembly Processes of the Microbial Rare Biosphere. Trends Microbiol [Internet]. 2018 Sep;26(9):738–47. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0966842X18300477

14. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. Activity of abundant and rare bacteria in a coastal ocean. Proc Natl Acad Sci [Internet]. 2011 Aug 2;108(31):12776–81. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1101405108

15. Hugoni M, Taib N, Debroas D, Domaizon I, Jouan Dufournel I, Bronner G, et al. Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. Proc Natl Acad Sci [Internet]. 2013 Apr 9;110(15):6004–9. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1216863110

16. Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, et al. Patterns of Rare and Abundant Marine Microbial Eukaryotes. Curr Biol [Internet]. 2014 Apr;24(8):813–21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0960982214002188

17. Youssef NH, Couger MB, Elshahed MS. Fine-Scale Bacterial Beta Diversity within a Complex Ecosystem (Zodletone Spring, OK, USA): The Role of the Rare Biosphere. Ahmed N, editor. PLoS One [Internet]. 2010 Aug 26;5(8):e12414. Available from: https://dx.plos.org/10.1371/journal.pone.0012414

18. Bowen JL, Morrison HG, Hobbie JE, Sogin ML. Salt marsh sediment diversity: a test of the variability of the rare biosphere among environmental replicates. ISME J [Internet]. 2012 Nov 28;6(11):2014–23. Available from: http://dx.doi.org/10.1038/ismej.2012.47

19. Logares R, Lindström ES, Langenheder S, Logue JB, Paterson H, Laybourn-Parry J, et al. Biogeography of bacterial communities exposed to progressive long-term environmental change. ISME J [Internet]. 2013 May 20;7(5):937–48. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3635229&tool=pmcentrez&rendertype=abstract%5Cnhttp://www.nature.com/doifinder/10.1038/ismej.2012.168

20. Fuentes S, Barra B, Caporaso JG, Seeger M. From Rare to Dominant: a Fine-Tuned Soil Bacterial Bloom during Petroleum Hydrocarbon Bioremediation. Löffler FE, editor. Appl Environ Microbiol [Internet]. 2016 Feb 1;82(3):888–96. Available from: http://aem.asm.org/lookup/doi/10.1128/AEM.02625-15

21. Pedrós-Alió C. Marine microbial diversity: can it be determined? Trends Microbiol [Internet]. 2006 Jun;14(6):257–63. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0966842X06000989

22. Pester M, Bittner N, Deevong P, Wagner M, Loy A. A 'rare biosphere' microorganism contributes to sulfate reduction in a peatland. ISME J [Internet]. 2010 Dec 10;4(12):1591–602. Available from: http://dx.doi.org/10.1038/ismej.2010.75

23. Kalenitchenko D, Le Bris N, Peru E, Galand PE. Ultrarare marine microbes contribute to key sulphur-related ecosystem functions. Mol Ecol [Internet]. 2018 Mar;27(6):1494–504. Available from: http://doi.wiley.com/10.1111/mec.14513

24. Wang Y, Hatt JK, Tsementzi D, Rodriguez-R LM, Ruiz-Pérez CA, Weigand MR, et al. Quantifying the Importance of the Rare Biosphere for Microbial Community Response to Organic Pollutants in a Freshwater Ecosystem. Stams AJM, editor. Appl Environ Microbiol [Internet]. 2017 Apr 15;83(8):e03321-16. Available from: http://aem.asm.org/lookup/doi/10.1128/AEM.03321-16

25. Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol Oceanogr [Internet]. 2000;45(6):1320–8. Available from: https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.4319/lo.2000.45.6.1320

26. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, et al. Lytic to temperate switching of viral communities. Nature [Internet]. 2016 Mar 16;531(7595):466–70. Available from: http://dx.doi.org/10.1038/nature17193

27. Kirchman DL, Cottrell MT, Lovejoy C. The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. Environ Microbiol [Internet]. 2010 Feb 3;12(5):1132–43. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2010.02154.x

28. Gobet A, Böer SI, Huse SM, van Beusekom JEE, Quince C, Sogin ML, et al. Diversity and dynamics of rare

and of resident bacterial populations in coastal sands. ISME J [Internet]. 2012 Mar 6;6(3):542–53. Available from: http://www.nature.com/doifinder/10.1038/ismej.2011.132

29. Vergin K, Done B, Carlson C, Giovannoni S. Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean. Aquat Microb Ecol [Internet]. 2013 Nov 15;71(1):1–13. Available from: http://www.int-res.com/abstracts/ame/v71/n1/p1-13/

30. Barber DG, Hop H, Mundy CJ, Else B, Dmitrenko IA, Tremblay J-E, et al. Selected physical, biological and biogeochemical implications of a rapidly changing Arctic Marginal Ice Zone. Prog Oceanogr [Internet]. 2015 Dec;139:122–50. Available from: http://dx.doi.org/10.1016/j.pocean.2015.09.003

31. Sauret C, Séverin T, Vétion G, Guigue C, Goutx M, Pujo-Pay M, et al. 'Rare biosphere' bacteria as key phenanthrene degraders in coastal seawaters. Environ Pollut [Internet]. 2014 Nov;194:246–53. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0269749114003169

32. Liu L, Yang J, Yu Z, Wilkinson DM. The biogeography of abundant and rare bacterioplankton in the lakes and reservoirs of China. ISME J [Internet]. 2015 Sep 6;9(9):2068–77. Available from: http://dx.doi.org/10.1038/ismej.2015.29

33. Aanderud ZT, Jones SE, Fierer N, Lennon JT. Resuscitation of the rare biosphere contributes to pulses of ecosystem activity. Front Microbiol [Internet]. 2015 Jan 30;6(JAN):1–11. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00024/abstract

34. Baltar F, Palovaara J, Vila-Costa M, Salazar G, Calvo E, Pelejero C, et al. Response of rare, common and abundant bacterioplankton to anthropogenic perturbations in a Mediterranean coastal site. FEMS Microbiol Ecol [Internet]. 2015 Jun 1;91(6):1–12. Available from: https://academic.oup.com/femsec/article-lookup/doi/10.1093/femsec/fiv058

35. Anderson RE, Sogin ML, Baross JA. Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents. FEMS Microbiol Ecol [Internet]. 2015 Jan 1;91(1):1–11. Available from: https://academic.oup.com/femsec/article-lookup/doi/10.1093/femsec/fiu016

36. Fernandez-Gonzalez N, Huber JA, Vallino JJ. Microbial Communities Are Well Adapted to Disturbances in Energy Input. Chu H, editor. mSystems [Internet]. 2016 Oct 25;1(5):1–15. Available from: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00117-16

37. Kaminsky R, Morales SE. Conditionally rare taxa contribute but do not account for changes in soil prokaryotic community structure. Front Microbiol. 2018;9(APR):1–6.

38. Zhang Y, Wu G, Jiang H, Yang J, She W, Khan I, et al. Abundant and Rare Microbial Biospheres Respond Differently to Environmental and Spatial Factors in Tibetan Hot Springs Yanmin. Front Microbiol. 2018;9(SEP):1–16.

39. Szabó KÉ, Itor POB, Bertilsson S, Tranvik L, Eiler A. Importance of rare and abundant populations for the

structure and functional potential of freshwater bacterial communities. Aquat Microb Ecol. 2007;47:1–10.

40. Elshahed MS, Youssef NH, Spain AM, Sheik C, Najar FZ, Sukharnikov LO, et al. Novelty and Uniqueness Patterns of Rare Members of the Soil Biosphere. Appl Environ Microbiol [Internet]. 2008 Sep 1;74(17):5422–8. Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.00410-08

41. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. Ecology of the rare microbial biosphere of the Arctic Ocean. Proc Natl Acad Sci [Internet]. 2009 Dec 29;106(52):22427–32. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0908284106

42. Weisse T. Ciliates and the Rare Biosphere-Community Ecology and Population Dynamics. J Eukaryot Microbiol [Internet]. 2014 Jul;61(4):419–33. Available from: http://doi.wiley.com/10.1111/jeu.12123

43. Ziegler M, Eguíluz VM, Duarte CM, Voolstra CR. Rare symbionts may contribute to the resilience of coral-algal assemblages. ISME J. 2018;12(1):161–72.

44. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner HW, et al. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol. 2010;19(Suppl. 1):21–31.

45. Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A, et al. Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. BMC Biol. 2009;7(i):72.

46. Aguilar M, Richardson E, Tan BF, Walker G, Dunfield PF, Bass D, et al. Next-Generation Sequencing Assessment of Eukaryotic Diversity in Oil Sands Tailings Ponds Sediments and Surface Water. J Eukaryot Microbiol [Internet]. 2016 Nov;63(6):732–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27062087

47. Yang S, Winkel M, Wagner D, Liebner S. Community structure of rare methanogenic archaea: Insight from a single functional group. FEMS Microbiol Ecol. 2017;93(11):1–10.

48. Shade A, Hogan CS, Klimowicz AK, Linske M, McManus PS, Handelsman J. Culturing captures members of the soil rare biosphere. Environ Microbiol [Internet]. 2012 Sep;14(9):2247–52. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2012.02817.x

49. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Elife. 2019;8:1–31.

50. Darwin C. The Origin of Species [Internet]. The Pennsylvania State University. Amsterdam: Amsterdam University Press; 1859. 448 p. Available from: http://www.f.waseda.jp/sidoli/Darwin_Origin_Of_Species.pdf%0A

51. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, et al. Species abundance distributions:

moving beyond single prediction theories to integration within an ecological framework. Ecol Lett [Internet]. 2007 Oct;10(10):995–1015. Available from: http://doi.wiley.com/10.1111/j.1461-0248.2007.01094.x

52. Corbet AS. The distribution of butterflies in the Malay Peninsula (Lepid.). Proc R Entomol Soc London Ser A, Gen Entomol [Internet]. 2009 Apr 2;16(10–12):101–16. Available from: http://doi.wiley.com/10.1111/j.1365-3032.1941.tb00970.x

53. Fisher RA, Corbet AS, Williams CB. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. J Anim Ecol [Internet]. 1943 May;12(1):42. Available from: http://www.jstor.org/stable/1411?origin=crossref

54. Preston FW. The commonness and rarity of species. Ecology [Internet]. 1948;29(3):254–83. Available from: http://www.jstor.org/stable/1930989

55. MacArthur RH. On the relative abundance of bird species. Proc Natl Acad Sci U S A [Internet]. 1957 Mar 15;43(3):293–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16590018

56. Magurran AE, Henderson PA. Commonness and rarity. In: Biological Diversity. 2011. p. 97–104.

57. H Rabinowitz RA. The Biological Aspects of Rare Plant Conservation Edited by Hugh Synge Seven forms of rarity. In: The Biological Aspects of Rare Plant Conservation [Internet]. 1981. p. 205–17. Available from: https://www.esf.edu/efb/parry/Invert_Cons_14_Readings/Rabinowitz_1981.pdf

58. Kunin WE, Gaston KJ. The biology of rarity: Patterns, causes and consequences. Trends Ecol Evol [Internet]. 1993 Aug;8(8):298–301. Available from: http://linkinghub.elsevier.com/retrieve/pii/016953479390259R

59. Pedrós-Alió C. Dipping into the Rare Biosphere. Science (80-) [Internet]. 2007 Jan;315(5809):192–3. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1135933

60. Kysela DT, Palacios C, Sogin ML. Serial analysis of V6 ribosomal sequence tags (SARST-V6): a method for efficient, high-throughput analysis of microbial community composition. Environ Microbiol [Internet]. 2005 Mar;7(3):356–64. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2004.00712.x

61. Neufeld JD, Li J, Mohn WW. Scratching the surface of the rare biosphere with ribosomal sequence tag primers. FEMS Microbiol Lett [Internet]. 2008 Apr 21;283(2):146–53. Available from: https://academic.oup.com/femsle/article-lookup/doi/10.1111/j.1574-6968.2008.01124.x

62. de Wit R, Bouvier T. “Everything is everywhere, but, the environment selects”; what did Baas Becking and Beijerinck really say? Environ Microbiol [Internet]. 2006 Apr;8(4):755–8. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2006.01017.x

63. Musat N, Halm H, Winterholler B, Hoppe P, Peduzzi S, Hillion F, et al. A single-cell view on the ecophysiology of anaerobic phototrophic bacteria. Proc Natl Acad Sci [Internet]. 2008 Nov

18;105(46):17861–6. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0809329105

64. Galand PE, Casamayor EO, Kirchman DL, Potvin M, Lovejoy C. Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. ISME J [Internet]. 2009 Jul 26;3(7):860–9. Available from: http://dx.doi.org/10.1038/ismej.2009.23

65. Shade A, Gilbert JA. Temporal patterns of rarity provide a more complete view of microbial diversity. Trends Microbiol [Internet]. 2015 Jun;23(6):335–40. Available from: http://dx.doi.org/10.1016/j.tim.2015.01.007

66. Steven B, Hesse C, Soghigian J, Gallegos-Graves LV, Dunbar J. Simulated rRNA/DNA Ratios Show Potential To Misclassify Active Populations as Dormant. Löffler FE, editor. Appl Environ Microbiol [Internet]. 2017 Jun 1;83(11):1–11. Available from: http://aem.asm.org/lookup/doi/10.1128/AEM.00696-17

67. Steiner PA, De Corte D, Geijo J, Mena C, Yokokawa T, Rattei T, et al. Highly variable mRNA half-life time within marine bacterial taxa and functional genes. Environ Microbiol. 2019;21(10):3873–84.

68. Sjöstedt J, Koch-Schmidt P, Pontarp M, Canbäck B, Tunlid A, Lundberg P, et al. Recruitment of Members from the Rare Biosphere of Marine Bacterioplankton Communities after an Environmental Disturbance. Appl Environ Microbiol [Internet]. 2012 Mar 1;78(5):1361–9. Available from: http://aem.asm.org/lookup/doi/10.1128/AEM.05542-11

69. Hubbell SP. Neutral theory and the evolution of ecological equivalence. Ecology [Internet]. 2006;87(6):1387–98. Available from: papers2://publication/uuid/64C5D248-28D8-4E33-AF88-B8AF83EB015C

70. Hutchinson GE. Concluding Remarks. Cold Spring Harb Symp Quant Biol [Internet]. 1957 Jan 1;22:415–27. Available from: http://link.springer.com/10.1007/s00726-011-1022-z

71. Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA. The Western English Channel contains a persistent microbial seed bank. ISME J [Internet]. 2012 Jun 10;6(6):1089–93. Available from: http://dx.doi.org/10.1038/ismej.2011.162

72. Ai D, Chu C, Ellwood MDF, Hou R, Wang G. Migration and niche partitioning simultaneously increase species richness and rarity. Ecol Modell [Internet]. 2013 Jun;258:33–9. Available from: http://dx.doi.org/10.1016/j.ecolmodel.2013.03.001

73. Nolte V, Pandey RV, Jost S, Medinger R, Ottenwalder B, Boenigk J, et al. Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol Ecol [Internet]. 2010 Jul 1;19(14):2908–15. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2010.04669.x

74. Wei ST-S, Wu Y-W, Lee T-H, Huang Y-S, Yang C-Y, Chen Y-L, et al. Microbial Functional Responses to Cholesterol Catabolism in Denitrifying Sludge. mSystems [Internet]. 2018;3(5):1–19. Available from: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00113-18

75. Reveillaud J, Maignien L, Eren AM, Huber JA, Apprill A, Sogin ML, et al. Host-specificity among abundant and rare taxa in the sponge microbiome. ISME J [Internet]. 2014 Jun 9;8(6):1198–209. Available from: http://www.nature.com/articles/ismej2013227

76. Hol WHG, de Boer W, de Hollander M, Kuramae EE, Meisner A, van der Putten WH. Context dependency and saturating effects of loss of rare soil microbes on plant productivity. Front Plant Sci [Internet]. 2015 Jun 30;6:1–10. Available from: http://journal.frontiersin.org/Article/10.3389/fpls.2015.00485/abstract

77. Dawson W, Hör J, Egert M, van Kleunen M, Pester M. A Small Number of Low-abundance Bacteria Dominate Plant Species-specific Responses during Rhizosphere Colonization. Front Microbiol [Internet]. 2017 May 29;8:1–13. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00975/full

78. Troussellier M, Escalas A, Bouvier T, Mouillot D. Sustaining rare marine microorganisms: Macroorganisms as repositories and dispersal agents of microbial diversity. Front Microbiol. 2017;8(MAY):1–17.

79. Hardoim CCP, Costa R. Temporal dynamics of prokaryotic communities in the marine sponge *Sarcotragus spinosulus*. Mol Ecol [Internet]. 2014 Jun;23(12):3097–112. Available from: http://doi.wiley.com/10.1111/mec.12789

80. Hausmann B, Knorr K-HH, Schreck K, Tringe SG, Glavina del Rio T, Loy A, et al. Consortia of low-abundance bacteria drive sulfate reduction-dependent degradation of fermentation products in peat soil microcosms. ISME J [Internet]. 2016 Oct 25;10(10):2365–75. Available from: http://dx.doi.org/10.1038/ismej.2016.42

81. Bodelier P LE, Meima-Franke M, Hordijk CA, Steenbergh AK, Hefting MM, Bodrossy L, et al. Microbial minorities modulate methane consumption through niche partitioning. ISME J [Internet]. 2013 Nov 20;7(11):2214–28. Available from: http://dx.doi.org/10.1038/ismej.2013.99

82. Hua Z-SS, Han Y-JJ, Chen L-XX, Liu J, Hu M, Li S-JJ, et al. Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. ISME J [Internet]. 2015 Jun 7;9(6):1280–94. Available from: http://dx.doi.org/10.1038/ismej.2014.212

83. Crespo BG, Wallhead PJ, Logares R, Pedrós-Alió C. Probing the Rare Biosphere of the North-West Mediterranean Sea: An Experiment with High Sequencing Effort. Martinez-Abarca F, editor. PLoS One [Internet]. 2016 Jul 21;11(7):e0159195. Available from: http://dx.doi.org/10.1371/journal.pone.0159195

84. Hamasaki K, Taniguchi A, Tada Y, Kaneko R, Miki T. Active populations of rare microbes in oceanic environments as revealed by bromodeoxyuridine incorporation and 454 tag sequencing. Gene [Internet]. 2016 Feb;576(2):650–6. Available from: http://dx.doi.org/10.1016/j.gene.2015.10.016

85. Steger D, Wentrup C, Braunegger C, Deevong P, Hofer M, Richter A, et al. Microorganisms with Novel Dissimilatory (Bi)Sulfite Reductase Genes Are Widespread and Part of the Core Microbiota in Low-Sulfate Peatlands. Appl Environ Microbiol [Internet]. 2011 Feb 15;77(4):1231–42. Available from:

http://aem.asm.org/lookup/doi/10.1128/AEM.01352-10

86. Griffiths BS, Kuan HL, Ritz K, Glover LA, McCaig AE, Fenwick C. The Relationship between Microbial Community Structure and Functional Stability, Tested Experimentally in an Upland Pasture Soil. Microb Ecol [Internet]. 2004 Jan 1;47(1):104–13. Available from: http://link.springer.com/10.1007/s00248-002-2043-7

87. Philippot L, Spor A, Hénault C, Bru D, Bizouard F, Jones CM, et al. Loss in microbial diversity affects nitrogen cycling in soil. ISME J [Internet]. 2013 Aug 7;7(8):1609–19. Available from: http://www.nature.com/articles/ismej201334

88. Giebler J, Wick LY, Chatzinotas A, Harms H. Alkane-degrading bacteria at the soil-litter interface: comparing isolates with T-RFLP-based community profiles. FEMS Microbiol Ecol [Internet]. 2013 Oct;86(1):45–58. Available from: https://academic.oup.com/femsec/article-lookup/doi/10.1111/1574-6941.12097

89. Dell'Anno A, Beolchini F, Rocchetti L, Luna GM, Danovaro R. High bacterial biodiversity increases degradation performance of hydrocarbons during bioremediation of contaminated harbor marine sediments. Environ Pollut [Internet]. 2012 Aug;167:85–92. Available from: http://dx.doi.org/10.1016/j.envpol.2012.03.043

90. Sanchez MA, Vasquez M, Gonzalez B. A previously unexposed forest soil microbial community degrades high levels of the pollutant 2,4,6-trichlorophenol. Appl Environ Microbiol. 2004;70(12):7567–70.

91. Hernandez-Raquet G, Durand E, Braun F, Cravo-Laureau C, Godon J-J. Impact of microbial diversity depletion on xenobiotic degradation by sewage-activated sludge. Environ Microbiol Rep [Internet]. 2013 Aug;5(4):588–94. Available from: http://doi.wiley.com/10.1111/1758-2229.12053

92. Walke JB, Becker MH, Loftus SC, House LL, Cormier G, Jensen R V., et al. Amphibian skin may select for rare environmental microbes. ISME J [Internet]. 2014;8(11):2207–17. Available from: http://dx.doi.org/10.1038/ismej.2014.77

93. Webster NS, Thomas T. The Sponge Hologenome. MBio [Internet]. 2016 May 4;7(2):e00135-16. Available from: http://mbio.asm.org/lookup/doi/10.1128/mBio.00135-16

94. Taylor MW, Tsai P, Simister RL, Deines P, Botte E, Ericson G, et al. 'Sponge-specific' bacteria are widespread (but rare) in diverse marine environments. ISME J [Internet]. 2013 Feb 4;7(2):438–43. Available from: http://www.nature.com/articles/ismej2012111

95. Hol WHG, de Boer W, Termorshuizen AJ, Meyer KM, Schneider JHM, van Dam NM, et al. Reduction of rare soil microbes modifies plant-herbivore interactions. Ecol Lett [Internet]. 2010 Mar;13(3):292–301. Available from: http://doi.wiley.com/10.1111/j.1461-0248.2009.01424.x

96. Gaidos E, Rusch A, Ilardo M. Ribosomal tag pyrosequencing of DNA and RNA from benthic coral reef

microbiota: community spatial structure, rare members and nitrogen-cycling guilds. Environ Microbiol [Internet]. 2011 May;13(5):1138–52. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2010.02392.x

97. Horz H-P. Archaeal Lineages within the Human Microbiome: Absent, Rare or Elusive? Life [Internet]. 2015 May 5;5(2):1333–45. Available from: http://www.mdpi.com/2075-1729/5/2/1333/

98. van der Gast CJ, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW, et al. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. ISME J [Internet]. 2011 May 9;5(5):780–91. Available from: http://dx.doi.org/10.1038/ismej.2010.175

99. Hajishengallis G, Liang S, Payne MA, Hashim A, Jotwani R, Eskan MA, et al. A Low-Abundance Biofilm Species Orchestrates Inflammatory Periodontal Disease through the Commensal Microbiota and Complement. Cell Host Microbe [Internet]. 2011 Nov;10(5):497–506. Available from: https://linkinghub.elsevier.com/retrieve/pii/S193131281100299X

100. Webster NS, Taylor MW, Behnam F, Lücker S, Rattei T, Whalan S, et al. Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. Environ Microbiol [Internet]. 2009 Oct;12(8):2070–82. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2009.02065.x

101. Quero GM, Luna GM. Diversity of rare and abundant bacteria in surface waters of the Southern Adriatic Sea. Mar Genomics [Internet]. 2014 Oct;17:9–15. Available from: http://dx.doi.org/10.1016/j.margen.2014.04.002

102. Idris H, Goodfellow M, Sanderson R, Asenjo JA, Bull AT. Actinobacterial Rare Biospheres and Dark Matter Revealed in Habitats of the Chilean Atacama Desert. Sci Rep [Internet]. 2017;7(1):1–11. Available from: http://dx.doi.org/10.1038/s41598-017-08937-4

103. Ashby MN, Rine J, Mongodin EF, Nelson KE, Dimster-Denk D. Serial Analysis of rRNA Genes and the Unexpected Dominance of Rare Members of Microbial Communities. Appl Environ Microbiol [Internet]. 2007 Jul 15;73(14):4532–42. Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.02956-06

104. Alonso-Sáez L, Zeder M, Harding T, Pernthaler J, Lovejoy C, Bertilsson S, et al. Winter bloom of a rare betaproteobacterium in the Arctic Ocean. Front Microbiol. 2014;5(AUG):1–9.

105. Xue Y, Chen H, Yang JR, Liu M, Huang B, Yang J. Distinct patterns and processes of abundant and rare eukaryotic plankton communities following a reservoir cyanobacterial bloom. ISME J [Internet]. 2018;12(9):2263–77. Available from: http://dx.doi.org/10.1038/s41396-018-0159-0

106. Kurm V, van der Putten WH, Hol WHG. Cultivation-success of rare soil bacteria is not influenced by incubation time and growth medium. Smidt H, editor. PLoS One [Internet]. 2019 Jan 10;14(1):e0210073. Available from: http://dx.plos.org/10.1371/journal.pone.0210073

107. Kurm V, van der Putten WH, Weidner S, Geisen S, Snoek BL, Bakx T, et al. Competition and predation as possible causes of bacterial rarity. Environ Microbiol [Internet]. 2019 Apr 18;21(4):1356–68. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1462-2920.14569

108. Gobet A, Quince C, Ramette A. Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets. Nucleic Acids Res [Internet]. 2010 Aug;38(15):e155–e155. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq545

109. Kendall MG. A New Measure of Rank Correlation. Biometrika [Internet]. 1938 Jun;30(1/2):81. Available from: https://www.jstor.org/stable/2332226?origin=crossref

110. Gower JC. Generalized procrustes analysis. Psychmetrika [Internet]. 1975;40(1):33–51. Available from: https://link.springer.com/article/10.1007/BF02291478

111. Vellend M. Conceptual Synthesis in Community Ecology. Q Rev Biol [Internet]. 2010 Jun;85(2):183–206. Available from: https://www.journals.uchicago.edu/doi/10.1086/652373

112. Staley J, Konopka A. Measurement of In Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annu Rev Microbiol [Internet]. 1985 Jan 1;39(1):321–46. Available from: http://micro.annualreviews.org/cgi/doi/10.1146/annurev.micro.39.1.321

113. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. Neufeld JD, editor. mSystems [Internet]. 2018 Sep 25;3(5):1–12. Available from: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00055-18

114. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci [Internet]. 1977 Nov 1;74(11):5088–90. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.74.11.5088

115. Ruddy KJ, Partridge AH. Adherence with adjuvant hormonal therapy for breast cancer. Ann Oncol [Internet]. 2008 Oct 7;20(3):401–2. Available from: https://academic.oup.com/annonc/article-lookup/doi/10.1093/annonc/mdp039

116. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. Genetic diversity in Sargasso Sea bacterioplankton. Nature [Internet]. 1990 Nov;345:60–2. Available from: http://linkinghub.elsevier.com/retrieve/pii/0021979780905019

117. Ward DM, Roland W, Bateson MM. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature [Internet]. 1990 Nov;345:63–5. Available from: http://linkinghub.elsevier.com/retrieve/pii/0021979780905019

118. Hong S-H, Bunge J, Jeon S-O, Epstein SS. Predicting microbial species richness. Proc Natl Acad Sci [Internet]. 2006 Jan 3;103(1):117–22. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0507245102

119. Vacher C, Tamaddoni-Nezhad A, Kamenova S, Peyrard N, Moalic Y, Sabbadin R, et al. Learning Ecological Networks from Next-Generation Sequencing Data. Lett To Nat [Internet]. 2016;430:1–39. Available from: https://doi.org/10.1016/bs.aecr.2015.10.004

120. Pommier T, Canback LR, Bostrom KH, Simu K, Lundberg P, Tunlid A, et al. Global patterns of diversity and community structure in marine bacterioplankton. Mol Ecol [Internet]. 2006 Dec 12;16(4):867–80. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2006.03189.x

121. Muyzer G, de Waal EC, Uitterlinden AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol [Internet]. 1993 Mar;59(3):695–700. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7683183

122. Fisher MM, Triplett EW. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol. 1999;65(10):4630–6.

123. Casamayor EO, Calderon-Paz JI, Pedros-Alio C. 5S rRNA fingerprints of marine bacteria, halaophilic archaea and natural prokaryotic assemblages. Fems Microbiol Ecol. 2000;34(3):113.

124. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature [Internet]. 2005 Sep;437(7057):376–80. Available from: http://www.nature.com/articles/nature03959

125. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J [Internet]. 2017 Dec 21;11(12):2639–43. Available from: http://dx.doi.org/10.1038/ismej.2017.119

126. Mo Y, Zhang W, Yang J, Lin Y, Yu Z, Lin S. Biogeographic patterns of abundant and rare bacterioplankton in three subtropical bays resulting from selective and neutral processes. ISME J [Internet]. 2018;12(9):2198–210. Available from: http://dx.doi.org/10.1038/s41396-018-0153-6

127. Zhang Y, Zhao Z, Dai M, Jiao N, Herndl GJ. Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea. Mol Ecol [Internet]. 2014;23(9):2260–74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4230472/

128. Ruiz-González C, Logares R, Sebastián M, Mestre M, Rodríguez-Martínez R, Galí M, et al. Higher contribution of globally rare bacterial taxa reflects environmental transitions across the surface ocean. Mol Ecol. 2019;28(8):1930–45.

129. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol [Internet]. 2007 Aug 15;73(16):5261–7. Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.00062-07

130. Liu M, Xue Y, Yang J. Rare Plankton Subcommunities Are Far More Affected by DNA Extraction Kits Than Abundant Plankton. Front Microbiol [Internet]. 2019;10(March):1–12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5632291/

131. Shen R, Fan J-B, Campbell D, Chang W, Chen J, Doucet D, et al. High-throughput SNP genotyping on universal bead arrays. Mutat Res Mol Mech Mutagen [Internet]. 2005 Jun;573(1–2):70–82. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0027510705000278

132. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics [Internet]. 2016;107(1):1–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26554401

133. Voelkerding K V., Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin Chem [Internet]. 2009 Apr 1;55(4):641–58. Available from: http://www.clinchem.org/cgi/doi/10.1373/clinchem.2008.112789

134. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al. Metagenomics: Tools an insights for analyzing Next-Generation Sequencing data derived from biodiversity studies. Bioinform Biol Insights. 2015;(9):75–88.

135. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol [Internet]. 2010 Mar 11;12(7):1889–98. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2010.02193.x

136. Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, et al. EBI Metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res [Internet]. 2018;46(D1):D726–35. Available from: https://academic.oup.com/nar/article-abstract/46/D1/D726/4561650

137. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res [Internet]. 2012 Nov 27;41(D1):D590–6. Available from: http://academic.oup.com/nar/article/41/D1/D590/1069277/The-SILVA-ribosomal-RNA-gene-database-project

138. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res [Internet]. 2018;47(November 2018):351–60. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky1100/5162469

139. Hamp TJ, Jones WJ, Fodor AA. Effects of Experimental Choices and Analysis Noise on Surveys of the "Rare Biosphere." Appl Environ Microbiol [Internet]. 2009 May 15;75(10):3263–70. Available from: http://aem.asm.org/cgi/doi/10.1128/AEM.01931-08

140. Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A. Amplification by PCR Artificially Reduces the Proportion of the Rare Biosphere in Microbial Communities. Gilbert JA, editor. PLoS One [Internet]. 2012 Jan

11;7(1):e29973. Available from: https://dx.plos.org/10.1371/journal.pone.0029973

141. Hardoim CCP, Cardinale M, Cúcio ACB, Esteves AIS, Berg G, Xavier JR, et al. Effects of sample handling and cultivation bias on the specificity of bacterial communities in keratose marine sponges. Front Microbiol. 2014;5(NOV):1–15.

142. Rego A, Raio F, Martins TP, Ribeiro H, Sousa AGG, Séneca J, et al. Actinobacteria and Cyanobacteria Diversity in Terrestrial Antarctic Microenvironments Evaluated by Culture-Dependent and Independent Methods. Front Microbiol [Internet]. 2019;10(May):1–19. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2019.01018/full

143. Zehavi T, Probst M, Mizrahi I. Insights into culturomics of the rumen microbiome. Front Microbiol [Internet]. 2018;9(AUG):1–10. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term=10.3389%2Ffmicb.2018.01999+Insights

144. Karimi E, Keller-Costa T, Slaby BM, Cox CJ, da Rocha UN, Hentschel U, et al. Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable Alphaproteobacteria. Sci Rep [Internet]. 2019;9(1):1–15. Available from: http://dx.doi.org/10.1038/s41598-019-38737-x

145. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, et al. Microbial culturomics: Paradigm shift in the human gut microbiome study. Clin Microbiol Infect [Internet]. 2012;18(12):1185–93. Available from: http://dx.doi.org/10.1111/1469-0691.12023

146. Jiang C-Y, Dong L, Zhao J-K, Hu X, Shen C, Qiao Y, et al. High-Throughput Single-Cell Cultivation on Microfluidic Streak Plates. Parales RE, editor. Appl Environ Microbiol [Internet]. 2016 Apr 1;82(7):2210–8. Available from: http://aem.asm.org/lookup/doi/10.1128/AEM.03588-15

147. Zhang QG, Buckling A, Godfray HCJ. Quantifying the relative importance of niches and neutrality for coexistence in a model microbial system. Funct Ecol [Internet]. 2009;23(6):1139–47. Available from: https://doi.org/10.1111/j.1365-2435.2009.01579.x

148. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31(6):533–8.

149. Karsenti E. The making of Tara Oceans: funding blue skies research for our Blue Planet. Mol Syst Biol [Internet]. 2015;11(5):811. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25999086%0A

150. Duarte CM. Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition. Limnol Oceanogr Bull. 2015;24(1):11–4.

151. Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K, et al. The ocean sampling day consortium. Gigascience [Internet]. 2015;4(1). Available from: http://dx.doi.org/10.1186/s13742-015-0066-5

152. Pesant S. Registry of all samples from the EuroMarine Inter-Comparison of Marine Plankton Metagenome Analysis Methods (EMOSE 2017 edition). 2017; Available from: https://doi.pangaea.de/10.1594/PANGAEA.879516

153. Parada AE, Needham DM, Fuhrman JA. Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol. 2016;18(5):1403–14.

154. Apprill A, Mcnally S, Parsons R, Weber L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb Ecol [Internet]. 2015 Jun 4;75(2):129–37. Available from: http://www.int-res.com/abstracts/ame/v75/n2/p129-137/

155. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J [Internet]. 2012 Aug 8;6(8):1621–4. Available from: http://dx.doi.org/10.1038/ismej.2012.8

156. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A [Internet]. 2011;108 Suppl:4516–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20534432%0A

157. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes. PLoS One. 2009;4(7):1–9.

158. Karimi E, Ramos M, Gonçalves JMS, Xavier JR, Reis MP, Costa R. Comparative Metagenomics Reveals the Distinctive Adaptive Features of the Spongia officinalis Endosymbiotic Consortium. Front Microbiol [Internet]. 2017 Dec 14;8(2499). Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2017.02499/full

159. Hardoim CCP, Esteves AIS, Pires FR, Gonçalves JMS, Cox CJ, Xavier JR, et al. Phylogenetically and Spatially Close Marine Sponges Harbour Divergent Bacterial Communities. Harder T, editor. PLoS One [Internet]. 2012 Dec 27;7(12):e53029. Available from: https://dx.plos.org/10.1371/journal.pone.0053029

160. Granskog M, Assmy P, Gerland S, Spreen G, Steen H, Smedsrud L. Arctic Research on Thin Ice: Consequences of Arctic Sea Ice Loss. Eos (Washington DC) [Internet]. 2016 Jan 26;97. Available from: https://eos.org/project-updates/arctic-research-on-thin-ice-consequences-of-arctic-sea-ice-loss

161. Meyer A, Sundfjord A, Fer I, Provost C, Villacieros Robineau N, Koenig Z, et al. Winter to summer oceanographic observations in the Arctic Ocean north of Svalbard. J Geophys Res Ocean [Internet]. 2017 Aug;122(8):6218–37. Available from: http://doi.wiley.com/10.1002/2016JC012391

162. de Sousa AGG, Tomasino MP, Duarte P, Fernández-Méndez M, Assmy P, Ribeiro H, et al. Diversity and Composition of Pelagic Prokaryotic and Protist Communities in a Thin Arctic Sea-Ice Regime. Microb Ecol. 2019;78(2):388–408.

163. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics [Internet]. 2014 Aug 1;30(15):2114–20. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170

164. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics [Internet]. 2013 Nov 15;29(22):2933–5. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt509

165. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46(D1):D335–42.

166. Matias Rodrigues JF, Schmidt TSB, Tackmann J, Von Mering C. MAPseq: Highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics. 2017;33(23):3808–10.

167. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput Vienna, Austria [Internet]. 2018; Available from: https://www.r-project.org/

168. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Watson M, editor. PLoS One [Internet]. 2013 Apr 22;8(4):e61217. Available from: https://dx.plos.org/10.1371/journal.pone.0061217

169. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J [Internet]. 1948 Jul;27(3):379–423. Available from: https://pure.mpg.de/rest/items/item_2383164/component/file_2383163/content

170. Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Community Ecology Package. R Package Version 2.5-3 [Internet]. 2018. Available from: http://cran.r-project.org/package=vegan

171. Hill MO, Gauch HG. Detrended Correspondence Analysis: An Improved Ordination Technique. Vegetatio [Internet]. 1980;42(1):47–58. Available from: http://www.jstor.org/stable/20145789

172. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol [Internet]. 1933;24(6):417–41. Available from: http://content.apa.org/journals/edu/24/6/417

173. Chen H. Generate High-Resolution Venn and Euler Plots. R package version 1.6.20. 2018; Available from: https://cran.r-project.org/package=VennDiagram

174. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Sep 1;19(9):1639–45. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.092759.109

175. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature [Internet]. 2015 Jul 15;523(7559):208–11. Available from: http://www.nature.com/articles/nature14486

176. Gonnella G, Böhnke S, Indenbirken D, Garbe-Schönberg D, Seifert R, Mertens C, et al. Endemic hydrothermal vent species identified in the open ocean seed bank. Nat Microbiol [Internet]. 2016 Aug 13;1(8):1–7. Available from: http://dx.doi.org/10.1038/nmicrobiol.2016.86

177. Sachdeva R, Campbell BJ, Heidelberg JF. Rare microbes from diverse Earth biomes dominate community activity. bioRxiv [Internet]. 2019;636373. Available from: https://www.biorxiv.org/content/10.1101/636373v1

178. Liao J, Cao X, Wang J, Zhao L, Sun J, Jiang D, et al. Similar community assembly mechanisms underlie similar biogeography of rare and abundant bacteria in lakes on Yungui Plateau, China. Limnol Oceanogr [Internet]. 2017;62(2):723–35. Available from: https://aslopubs.onlinelibrary.wiley.com/doi/full/10.1002/lno.10455

179. Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ, McDonald IR, et al. Groundtruthing Next-Gen Sequencing for Microbial Ecology-Biases and Errors in Community Structure Estimates from PCR Amplicon Pyrosequencing. PLoS One [Internet]. 2012;7(9):15–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22970184

180. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: Insights from comparison of two 16s rRNA clone libraries constructed from the same sample. Appl Environ Microbiol [Internet]. 2005;71(12):8966–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16332901

181. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature [Internet]. 2004 Mar 1;428(6978):37–43. Available from: http://www.nature.com/articles/nature02340

182. Salman V, Amann R, Shub DA, Schulz-Vogt HN. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. Proc Natl Acad Sci [Internet]. 2012;109(11):4203–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22371583

183. Baker BJ, Dick GJ. Omic Approaches in Microbial Ecology: Charting the Unknown. Microbe Mag [Internet]. 2013 Sep 1;8(9):353–60. Available from: http://www.asmscience.org/content/journal/microbe/10.1128/microbe.8.353.1

184. Wu D, Wu M, Halpern A, Rusch DB, Yooseph S, Frazier M, et al. Stalking the fourth domain in metagenomic data: Searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. PLoS One [Internet]. 2011;6(3). Available from: https://www.ncbi.nlm.nih.gov/pubmed/21437252

185. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. 2016;1(May):1–6. Available from: https://www.nature.com/articles/nmicrobiol201648

186. Chase JM, Kraft NJB, Smith KG, Vellend M, Inouye BD. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. Ecosphere [Internet]. 2011 Feb;2(2):art24. Available

from: http://doi.wiley.com/10.1890/ES10-00117.1

187. Dini-Andreote F, Stegen JC, van Elsas JD, Salles JF. Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. Proc Natl Acad Sci [Internet]. 2015 Mar 17;112(11):E1326–32. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1414261112

188. Stegen JC, Lin X, Fredrickson JK, Konopka AE. Estimating and mapping ecological processes influencing microbial community assembly. Front Microbiol [Internet]. 2015 May 1;6(MAY):1–15. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00370

189. Stegen JC, Lin X, Fredrickson JK, Chen X, Kennedy DW, Murray CJ, et al. Quantifying community assembly processes and identifying features that impose them. ISME J [Internet]. 2013 Nov 6;7(11):2069–79. Available from: http://dx.doi.org/10.1038/ismej.2013.93

190. Webb CO. Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. Am Nat [Internet]. 2000 Aug;156(2):145–55. Available from: http://www.journals.uchicago.edu/doi/10.1086/303378

191. Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. Phylogenies and Community Ecology. Annu Rev Ecol Syst [Internet]. 2002 Nov;33(1):475–505. Available from: http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.33.010802.150448

192. Thomas T, Moitinho-Silva L, Lurgi M, Björk JR, Easson C, Astudillo-García C, et al. Diversity, structure and convergent evolution of the global sponge microbiome. Nat Commun [Internet]. 2016 Dec 16;7(1):11870. Available from: http://www.nature.com/articles/ncomms11870

193. Schönberg CHL. Happy relationships between marine sponges and sediments – a review and some observations from Australia. J Mar Biol Assoc United Kingdom [Internet]. 2016 Mar 4;96(2):493–514. Available from: https://www.cambridge.org/core/product/identifier/S0025315415001411/type/journal_article

194. Hentschel U, Piel J, Degnan SM, Taylor MW. Genomic insights into the marine sponge microbiome. Nat Rev Microbiol [Internet]. 2012 Sep 30;10(9):641–54. Available from: http://www.nature.com/articles/nrmicro2839

7. Annexes

Annex I – MultiCoLA R script, applied to define microbial rarity

#Load functions necessary for MultiCoLA (108), available at:
#https://www.mpi-bremen.de/en/Software-4.html#section1550

#set working directory

setwd("/path_to_working_directory/")

#Load the original OTU table in text (.txt) format, with samples as columns and OTUs absolute #abundance values as rows, the first row indicates the column label and the first column indicates #the OTU label.

#The last column gives taxonomic information for each row, if available.

OTU_table <- read.table("OTU_table.txt",header=TRUE,row.names=1)

class(OTU_table) # to confirm if the table is a data frame

sapply(OTU_table,class) # to confirm if OTUs abundance is in numeric value and taxonomic #information as factor values

#Use the function taxa.pooler.1.4.r to transform the data frame in a list, where each line is a matrix #with each sample as rows (instead of columns) and taxonomic information as columns (instead of #rows).

source("taxa.pooler.1.4.r")

OTU_taxa<-taxa.pooler(OTU_table)

#The R prompt will ask:

#1. Number of samples? (e.g. 16)...

#2. Number of taxonomic levels? (e.g. phylum+class+order+family+genus=5)...

#3. Presence/absence tables as output? (y/n)...

#4. Output as text files? (y/n)...

class(OTU_taxa)

sapply(OTU_taxa,class) #Verify that the output is a list of matrices;

#To produce each truncated table with a specific cutoff, use COtables.1.4.r.

source("COtables.1.4.r") #Before using this function it is important to define:

#The number of taxonomic levels;

#(Type=): Application of cutoff based on the entire dataset "ADS" or by-sample "SAM";

#(typem=): Application of the cutoff to study the impact on "rare" or "dominant" fraction.

#The function is repeated for each taxonomic level selected, in this example, for all taxonomic levels #(from domain (level = 1) to species (level = 7));

emose_truncated.DS.1<-COtables(OTU_taxa[[1]],Type="SAM",typem="rare")

emose_truncated.DS.2<-COtables(OTU_taxa[[2]],Type="SAM",typem="rare")

emose_truncated.DS.3<-COtables(OTU_taxa[[3]],Type="SAM",typem="rare")

emose_truncated.DS.4<-COtables(OTU_taxa[[4]],Type="SAM",typem="rare")

emose_truncated.DS.5<-COtables(OTU_taxa[[5]],Type="SAM",typem="rare")

emose_truncated.DS.6<-COtables(OTU_taxa[[6]],Type="SAM",typem="rare")

emose_truncated.DS.7<-COtables(OTU_taxa[[7]],Type="SAM",typem="rare")

source("cutoff.impact.1.4.r") #Compares each truncated table with the original OTU table

#This function needs the following information:

#(Type=): cutoff used ("ADS" vs "SAM"). Use the same as before, in this example with "SAM";

#(corcoef): Correlation coefficients: "pearson", "spearman" and "kendal", in this example "spearman";

#(typem=): “dominant” vs “rare”, the same as before “rare“;

OTU_impact <- cutoff.impact (OTU_taxa, Type="SAM", corcoef="spearman", typem="rare")

#This function asks:

#Details of the NMDS calculations? (y/n)...;

#If SAM-based only, maximum cutoff value? (e.g. 208)... (it should correspond to “the lowest number

#of maximum OTU occurrences in all samples” (108))

#To produces figures and tables, use the function cutoff.impact.fig.1.4.r.

source(“cutoff.impact.fig.1.4.r”)

OTU_impact_out<-cutoff.impact.fig(OTU_impact)

#The function asks:

#Output as text files? (y/n)...

#Plot the results? (y/n)...

#The output in text file (.txt) allows further analysis;

#The threshold for rarity is decided by interpretation of data.

#chose a threshold in absolute value, n.

OTU_rare<-sapply(OTU_table,function(x) ifelse(x<n,x,0))

write.table(OTU_rare,"OTU_rare.txt")

Annex II – One-way ANOVA R script

#Groups to be compared are loaded as vectors

Group1 <- c() #values of group 1, separated by “,”;

Group2 <- c() #values of group 2, separated by “,”;

Groupn <- c() #values of group n, separated by “,”;

#Calculate p-values

Combined_Groups <- data.frame(cbind(Group1, Group2,Groupn))

Combined_Groups

summary(Combined_Groups)

Stacked_Groups <- stack(Combined_Groups)

Stacked_Groups

Anova_Results <- aov(values ~ ind, data = Stacked_Groups)

summary(Anova_Results)

Annex III – types.r function for defining different types of microbial rarity

```
types <- function(A,t){
#A is matrix of samples
#t is rarity threshold in number of reads per OTU per sample
M <- c() #empty matrix
for(i in 1:dim(A)[1]){
        if (sum(A[i,])==0){
                M[i] <- c("Absent")
        } else {
```

```
if (A[i,1] > t & A[i,2] > t & A[i,3]>t){
        M[i] <- c("Abundant")
} else {
if (A[i,1] == 0 | A[i,2] == 0| A [i,3] ==0){
        M[i] <- c("Transiently Rare")
}else{
if (A[i,1]>t | A[i,2]> t|A[i,3]>t){
        M[i] <- c("Conditionally Rare")
}else{
if (var(A[i,])>=10){
        M[i]<- c("Permanently Rare, with variation")
}else{
if (var(A[i,])<=10){
        M[i]<- c("Permanently Rare")
}}}}}}}M}
```