U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

PhD

PhD

3rd CYCLE

FCUP
2025

U.PORTO

Novel insights into the Ecology of the Microbial Rare Biosphere in the Arctic Ocean through the combination of Microbiome DNA sequencing and Machine Learning Approaches

Francisco Pascoal

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Novel insights into the Ecology of the Microbial Rare Biosphere in the Arctic Ocean through the combination of Microbiome DNA sequencing and Machine Learning Approaches

Francisco Daniel de Oliveira Pascoal
Biology
Department of Biology
Faculty of Sciences of the University of Porto
2025

PhD

PhD

Novel insights into the Ecology of the Microbial Rare Biosphere in the Arctic Ocean through the combination of Microbiome DNA sequencing and Machine Learning Approaches

Francisco Daniel de Oliveira Pascoal

Thesis carried out as part of the Doctor's Degree in Biology
Department of Biology
2025

Supervisor
Catarina Magalhães, Professor, Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Portugal.

Co-supervisor
Rodrigo Costa, Professor, Departamento de Bioengenharia, Instituto de Bioengenharia e Biociências (iBB), Instituto Superior Técnico, Universidade de Lisboa, Portugal.

Co-supervisor
Paula Branco, Professor, School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, Canada.

Co-supervisor
Luís Torgo, Professor, Faculty of Computer Science, Dalhousie University.

All the corrections determined by the jury, and only those, were made.

The Supervisor,

Porto, 30/05/2025

Dedicated to Professor António Manuel Nunes dos Santos

Acknowledgements

This PhD thesis is the result of work that was possible thanks to many people and institutions, whose gratitude I would like to express.

I acknowledge the support of the Fundação para a Ciência e Tecnologia (FCT) given through my PhD grant (2020.04453.BD), financed by the State Budget of the República Portuguesa and by the Programa Operacional Regional Do Norte 2020 (NORTE 2020). I also acknowledge FCT for financing several research projects that contributed directly to the research carried out in this thesis, as well as contributing to my participation in national and international scientific conferences, to present PhD results (2023.14123.PEX, UIDB/04423/2020, UIDP/04423/2020, UIDB/04565/2020, LA/P/0140/2020, NORTE-01–0145-FEDER-000040, NORTE-01-0145-FEDER-000064, Programa Operacional Regional de Lisboa - Project N. 007317). And I also acknowledge the FCT grants awarded to my supervisors, which played a crucial role in financing my PhD (PTDC/CTA-AMB/30997/2017, PTDC/CTA-AMB/4946/2020, 2022.02983.PTDC).

I acknowledge that the research carried out in the context of this PhD thesis benefited from additional funding sources. The Canada Research Chairs program to Professor Luís Torgo and Professor Paula Branco and Discovery Grants attributed to Professor Luís Torgo, both grants awarded by the Natural Sciences and Engineering Research Council of Canada. The ATLANTIDA project, which was partially financed by NORTE 2020, under the PORTUGAL 2020 Partnership Agreement and through the European Regional Development Fund. I also acknowledge financing and logistics support directly involved in the Arctic Ocean campaigns that resulted in datasets processed and analyzed in this PhD thesis, specifically, I acknowledge the Portuguese Polar Program, the Research Council of Norway (project no. 244646) and the Norwegian Ministries of Foreign Affairs and Climate and Environment (Arktis 2030).

A special acknowledgement to my supervisor (Professor Catarina Magalhães) and co-supervisors (Professor Rodrigo Costa, Professor Paula Branco and Professor Luís Torgo), because their support and scientific contribution was key for carrying out the work of this PhD thesis and for the development of my own scientific skills. I am grateful for the time and effort they invested in me and in my PhD thesis.

I want to express my gratitude to Professor Catarina Magalhães for being a continuous and consistent supporter of my PhD work, my scientific career, and for actively including me in the research carried out by the Microbiome Ecology and Biogeochemistry team. Professor Catarina Magalhães consistently helped me in the resolution of problems,

description of the Arctic Ocean datasets, general scientific guidance, and she also helped me find research opportunities.

I extend my gratitude to the continuous support of Professor Rodrigo Costa, who was my previous supervisor, during the period of my MSc thesis in Microbiology, and accepted to be my co-supervisor for this PhD thesis. The continuous help of Professor Rodrigo Costa was very important every step of the way, and our discussions were helpful in clarifying ideas, interpreting host-associated microbiomes and metagenomics data.

I am grateful for the contributions of Professor Luís Torgo, which was my co-supervisor during the first half of my PhD thesis, before unfortunately passing away. Professor Luís Torgo introduced me to the data science field and machine learning, he incentivized me to build an R package, guided me through the initial stage of its development, and shared many useful R learning materials with me, which were important in developing my R coding skills.

I am grateful to Professor Paula Branco, because she accepted to step in as my co-supervisor after Professor Luís Torgo, and her help was very important in building an R package, but also in applying and interpreting both unsupervised and rule-based machine learning approaches, which were extensively used in this PhD thesis. Additionally, Professor Paula Branco was always available to help me with questions regarding machine learning, statistics, and programming.

I acknowledge the support of Departamento de Biologia, Faculdade de Ciências (FCUP), University of Porto. This acknowledgment extends to the support given by the FCUP administrative services.

I acknowledge the support of the Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR), of the University of Porto. CIIMAR was the research center where the experimental work of this thesis was carried out. Besides the scientific and institutional support given by CIIMAR, I always felt welcome and very well integrated at CIIMAR. Within CIIMAR, I would like to acknowledge the support of the administrative services, the ECOBIOTEC lab and the Microbiome Ecology and Biogeochemistry research group. A special thanks to my ECOBIOTEC colleagues, which integrated me in CIIMAR and helped me whenever necessary: Rafaela Mendes, Diogo Alexandrino, Joana Fernandes, Mariana Oliveira, Mariana Girão, Maria Boto, Eva Gomes, Leonor Pizarro, and Ana Margarida Pereira.

I also acknowledge CIIMAR for the continued investment in the generation of long-term data from the Arctic Ocean, which was the main data used in this thesis. And I acknowledge CIIMAR for the support given to the European Open Science Exploration Initiative (EMOSE), which resulted in useful data that was also used in this PhD thesis.

I acknowledge the support of the Faculty of Computer Science of Dalhousie University, Halifax, Canada, where I spent three months of my PhD, as a visiting student. Dalhousie University was a welcoming university, with a campus culture that valued human values and interchange of knowledge. The interaction with several graduate students at Dalhousie working on data science provided valuable perspectives for my PhD, and the multi-cultural environment was personally enriching.

I acknowledge the institutional support of Departamento de Bioengenharia, Instituto de Bioengenharia e Biociências (iBB) of Instituto Superior Técnico, which was mainly provided through my co-supervisor, Professor Rodrigo Costa. I also acknowledge iBB for providing datasets of host-associated data, which were very useful for this PhD thesis.

I acknowledge the institutional support of School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, Canada, which was mainly provided through my co-supervisor, Professor Paula Branco.

I acknowledge the support of the Norwegian Polar Institute (NPI), which was necessary for the Arctic Ocean expeditions that culminated in several datasets used in this PhD thesis. I also acknowledge NPI for making available environmental metadata, which was necessary for the Arctic Ocean analysis carried out in this thesis.

I acknowledge the Norwegian Young Sea Ice Expedition (N-ICE, 2015) and its scientific team. I acknowledge the Environmental Monitoring of Svalbard and Jan Mayen (MOSJ), a Norwegian government monitoring program, which was necessary for several Arctic Ocean expeditions, culminating in a large dataset that was processed, organized and analyzed in this thesis. I acknowledge the EMOSE team, their expedition to the Mediterranean Sea, and the Service d'Observation du Laboratoire Arago, Banyuls-sur-Mer, for the creation of the EMOSE dataset in combination with CIIMAR.

I acknowledge the original authors of additional publicly available datasets that were used in this thesis: Richard Condit, *et al.* (Barro Colorado Island tree count); Xavier Arnan, *et al*. (Ants dataset); Tina Keller-Costa *et al*. (coral microbiome).

I acknowledge the precious help of Roberta Piredda, Grazia Marina Quero and Maria Paola Tomasino in the analysis of protist results in the EMOSE dataset. I also thank Roberta Piredda for curating the protist taxonomy of the EMOSE dataset.

I acknowledge the continuous support of Pedro Duarte and Philipp Assmy in the interpretation of Arctic Ocean ecology, as well as several explanations and clarifications regarding Arctic Ocean data, hydrography and general environmental context.

I acknowledge the support of David Needham and Jed A. Fuhrman for support in interpretation of bioinformatics results from *DADA2* and amplicon sequencing, the support of Alex Mitchell in the interpretation of the bioinformatics used by the MGnify platform, and the support of Julie Poulain in the clarification of the molecular biology protocol employed in the EMOSE dataset.

I acknowledge the contribution of Ana Gomes, because she provided the graphical design of the figure used to represent the EMOSE experimental design (Figure 3.1).

Even though not directly related to experimental results of this thesis, other peers have given general scientific support, including in scientific publications outside the scope of this PhD thesis, which indirectly supported this thesis with perspective and technical knowledge. In that sense, I extend my scientific acknowledgments to Charle Lee and Craig Cary, who gave me important insights into polar ecology, metagenomic data interpretation and scientific writing; and Inês Areosa, who introduced me to association rule mining. On a more general note, I also acknowledge the interchange of ideas and technical knowledge with Miguel Semedo, Adriana Rego, Joana Costa, Maria Paola Tomasino, Mariana Oliveira and Mafalda Baptista.

Besides my supervisor and co-supervisors, I also acknowledge all co-authors that contributed to the research papers published in the context of this PhD thesis: Pedro Duarte, Phillipp Assmy, Maria Paola Tomasino, Roberta Piredda, Grazia Marina Quero, Luís Torgo, Julie Poulain, Pierre E. Galand, Jed A. Fuhrman, Alex Mitchell, Tinkara Tinta, Timotej Turk Dermastia, Antonio Fernandez-Guerra, Alessandro Vezzi, Ramiro Logares, Francesca Malfatti, Hisashi Endo, Anna Maria Dąbrowska, Fabio De Pascale, Pablo Sánchez, Nicolas Henry, Bruno Fosso, Bryan Wilson, Stephan Toshchakov, Gregory Kevin Ferrant, Ivo Grigorov, Fabio Rocha Jimenez Vieira, and Stephane Pesant.

I acknowledge my former colleagues from Associação Nacional de Estudantes de Biologia, especially Beatriz Silva, Ana Cristina Figueiredo, and Diogo Silva, because of our shared experiences, academic growth and for opening my academic horizons.

I acknowledge José Xavier and José Seco for giving me a recommendation to work with Professor Catarina Magalhães at CIIMAR, the starting point of a path that would culminate in this PhD thesis.

I wish to thank my psychologist, Ana Teixeira, for supporting my mental wellbeing.

I wish to thank my high school teachers, Paulo Sá, Carla Loureiro, and Ana Paula Henriques for incentivizing me to become a scientist, and for believing in my potential.

I'm grateful to all my friends and family, who always supported me during my PhD path. Special thanks for the emotional support and close friendship with Ana Macedo, Daniel Freitas, Luís Monteiro, João Letras, Mariana Frazão, Mário Jorge Ferro, Nuno Vassalo e Silva, Paulo Gomes, and Quixan Zhao. I also thank the life lessons obtained in conversing with my friends Isabel Batista and Christopher Damien Auretta. Special thanks to my parents and to my brother.

Finally, I wish to thank António Manuel Nunes dos Santos for teaching me the importance of understanding the "history of ideas" and for inspiring me to become a good scientist.

Fundação para a Ciência e a Tecnologia

PROPOLAR Programa Polar Português

NORTE2020 PROGRAMA OPERACIONAL REGIONAL DO NORTE

PORTUGAL 2020

List of publications

Peer-reviewed articles published in the framework of this PhD thesis:

1. **Pascoal, F.**, Costa, R. and Magalhães, C. (2021) "The microbial rare biosphere: current concepts, methods and ecological principles," *FEMS Microbiology Ecology*, 97(fiaa227), pp. 1-15. Available at: https://doi.org/10.1093/femsec/fiaa227.
2. **Pascoal, F.** *et al.* (2022a) "Exploration of the Types of Rarity in the Arctic Ocean from the Perspective of Multiple Methodologies," *Microbial Ecology*, 84(1), pp. 59–72. Available at: https://doi.org/10.1007/s00248-021-01821-9.
3. **Pascoal, F.** *et al.* (2023) "Inter-comparison of marine microbiome sampling protocols," *ISME Communications*, 3(1), pp. 1-16. Available at: https://doi.org/10.1038/s43705-023-00278-w.
4. **Pascoal, F.** *et al.* (2024) "Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities," *Annals of Microbiology*, 74(29), pp. 1-12. Available at: https://doi.org/10.1186/s13213-024-01767-6.

Articles in *peer-review* at time of thesis submission:

5. **Pascoal F.**, *et al.*, (2025a) "Definition of the microbial rare biosphere through unsupervised machine learning", *Communications Biology*.
6. **Pascoal F.**, *et al.*, (2025b) "*ulrb*: an R package to automatically define rare species", *Ecological Informatics*.

Datasets published in the framework of this PhD thesis:

7. Arctic microbiome along Svalbard Cross Shelf transects (2017-2020). Raw reads from V4V5 16S rRNA gene amplicon sequencing. MGnify and European Nucleotide Archive (ENA). Project ID: PRJEB72025, PRJEB72030. PRJEB60815, and PRJEB72034.

Software published in the framework of this PhD thesis:

8. **Pascoal F.** *et al.*, (2025) "*ulrb*: Unsupervised Learning Based Definition of Microbial Rare Biosphere", R Package, CRAN (v0.1.5), GitHub (v0.1.6).

Copyright disclaimer

Elsevier

The following article was under peer-review in an Elsevier journal at time of thesis submission:

- **Pascoal F.**, *et al.*, (2025b) "*ulrb*: an R package to automatically define rare species", *Ecological Informatics*. Article in press. Available at: https://doi.org/10.1016/j.ecoinf.2025.103229.

Elsevier grants permission for Authors to include their articles in full or in part in a thesis or dissertation for non-commercial purposes.

Springer Nature

The following articles were published or were under peer-review at time of thesis submission in a Springer Nature journal:

- **Pascoal, F.** *et al.* (2023) "Inter-comparison of marine microbiome sampling protocols," *ISME Communications*[*], 3(1), p. 84. Available at: https://doi.org/10.1038/s43705-023-00278-w. [Open access under Creative Commons License 4]
- **Pascoal, F.** *et al.* (2022a) "Exploration of the Types of Rarity in the Arctic Ocean from the Perspective of Multiple Methodologies," *Microbial Ecology*, 84(1), pp. 59–72. Available at: https://doi.org/10.1007/s00248-021-01821-9.
- **Pascoal, F.** *et al.* (2024) "Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities," *Annals of Microbiology*, 74(1), p. 29. Available at: https://doi.org/10.1186/s13213-024-01767-6. [Open access under Creative Commons License 4]
- **Pascoal, F.** *et al.* (2025) "Definition of the microbial rare biosphere through unsupervised machine learning," *Communications Biology*, 8(1), p. 544. Available at: https://doi.org/10.1038/s42003-025-07912-4.

Springer Nature grants permission for Authors to include their articles in full or in part in a thesis or dissertation for non-commercial purposes. Additionally, some of the articles

[*] *ISME Communications* is not a member of Springer Nature since January 2024. Currently, *ISME Communications* is a member of Oxford Academics, but at time of publishing Pascoal *et al*. (2023), *ISME Communications* was still a member of Springer Nature.

were published under Creative Commons License 4, which grants permission for reutilization under appropriate attribution to the original.

Oxford Academic

The following articles were published in an Oxford Academic journal:

- **Pascoal, F.**, Costa, R. and Magalhães, C. (2021) "The microbial rare biosphere: current concepts, methods and ecological principles," *FEMS Microbiology Ecology*, 97(1). Available at: https://doi.org/10.1093/femsec/fiaa227.

Oxford Academic grants permission for Authors to include their articles in full or in part in a thesis or dissertation for non-commercial purposes.

Resumo

A biosfera rara microbiana corresponde ao conjunto de microrganismos de baixa abundância numa comunidade de microrganismos. A biosfera rara microbiana é importante para a resistência e resiliência dos ecossistemas, uma vez que funciona como um reservatório de genes, que têm o potencial de serem expressos em resposta à presença de pressões ambientais. Apesar desta relevância, historicamente, os estudos de ecologia microbiana não se têm focado nos microrganismos raros, porque estes ainda são geralmente considerados menos importantes para o funcionamento dos ecossistemas do que os microrganismos abundantes. De facto, embora alguns estudos tenham reconhecido a relevância ecológica da biosfera rara microbiana, ainda são poucos os estudos que abordam o possível papel e função desta no contexto das alterações climáticas. Em parte, esta falta de estudos deve-se aos desafios metodológicos no estudo da biosfera rara microbiana, especialmente se se tiver em conta ambientes naturais inóspitos, como é o caso do Oceano Ártico. Um dos principais desafios no estudo da biosfera rara microbiana prende-se com a falta de dados obtidos de forma sistemática e durante vários anos, pois tais dados são necessários para se poder comparar e avaliar a resposta da biosfera rara microbiana a alterações ambientais. Além da aquisição dos dados, a análise desses dados também é um desafio, uma vez que não é consensual, entre os pares, qual é a melhor abordagem para distinguir os microrganismos raros dos restantes e não é trivial analisar a relação entre a diversidade microbiana e combinações complexas de variáveis ambientais. Assim, com o objetivo de ultrapassar estes desafios, esta tese comparou sistematicamente variações metodológicas usadas na aquisição de dados relativos a comunidades microbianas; processou e publicou novos dados; e desenvolveu novas estratégias de análise da biosfera rara microbiana, que incluíram a incorporação de diferentes estratégias de aprendizagem automática.

A primeira parte desta tese refere-se ao estado da arte da biosfera rara microbiana. Na análise do estado da arte, utilizámos uma perspetiva histórica para compreender como surgiu o conceito de raridade em microbiologia, o que evidenciou que o estudo da biosfera rara microbiana foi inteiramente dependente do surgimento de tecnologias de sequenciação de elevado rendimento. Na análise do estado da arte também resumimos as principais descobertas da ecologia da biosfera rara relativamente aos diferentes grupos microbianos. No entanto, como evidenciado pela análise da literatura, a maioria das descobertas ecológicas da biosfera rara limitaram-se a procariotas e protistas, com

muito menos estudos relativamente a fungos e vírus. Finalmente, a análise do estado da arte resumiu os principais avanços metodológicos no estudo da biosfera rara microbiana dos últimos anos e ainda organizou sistematicamente os diferentes conceitos utilizados na literatura para explicar a ecologia da biosfera rara microbiana.

O trabalho experimental desta tese dividiu-se em três partes. As partes II e III focaram-se principalmente em questões metodológicas, e a parte IV integrou os avanços metodológicos das partes anteriores para uma análise mais aprofundada da biosfera rara microbiana.

Na parte II, focámo-nos na comparação de metodologias utilizadas para descrever comunidades microbianas do meio marinho, bem como a biosfera rara microbiana. Para o efeito, focámo-nos, primeiro, na comparação de protocolos de filtragem de água marinha com amostras do Mar Mediterrâneo (Capítulo 3). Os protocolos selecionados são semelhantes aos protocolos utilizados em campanhas internacionais de amostragem de água (*e.g.*, *Tara Oceans, Ocean Sampling Day*). Os resultados do Capítulo 3 mostraram que filtrar diferentes volumes de água marinha ou utilizar diferentes tipos de filtros não altera significativamente a diversidade das comunidades procariotas e protistas. No entanto, utilizar consecutivamente vários filtros com poros de diferentes tamanhos modificou significativamente a composição das comunidades de protistas. No Capítulo 4, o nosso foco passou para a avaliação do impacto da utilização de diferentes estratégias de sequenciação do DNA do microbioma, bem como da classificação taxonómica dessas sequências através de diferentes bases de dados taxonómicas, utilizando dados do Oceano Ártico. Os resultados do Capítulo 4 mostraram que a sequenciação completa do gene de 16S RNA ribossomal (rRNA), em vez da sequenciação de uma região pequena desse mesmo gene, permite uma classificação taxonómica mais eficiente. Mostrámos, também, que a seleção de base de dados usada como referência para a atribuição de taxonomia é tão relevante quanto a seleção de estratégia de sequenciação. É importante realçar que, para níveis taxonómicos elevados (filo e classe), qualquer combinação de estratégia de sequenciação e de base de dados taxonómica seria válida, uma vez que as diferenças mais importantes entre estas metodologias se identificaram principalmente ao nível do género e da espécie. No Capítulo 5, avaliámos o impacto da utilização de diferentes abordagens bioinformáticas no processamento dos resultados brutos de sequenciação, resultantes da amplificação da região V4V5 do gene para o 16S rRNA, mas também resultantes da sequenciação de metagenomas, com amostras oriundas do Oceano Ártico. No Capítulo 5 mostrámos que a utilização de variantes exatas das sequências e a utilização de unidades

taxonómicas operacionais são abordagens bioinformáticas válidas para o estudo da biosfera rara procariótica. Também mostrámos, no Capítulo 5, que a incorporação de metagenomas no estudo da biosfera rara procariótica tem o potencial de revelar grupos taxonómicos que ainda não foram cultivados em laboratório. Independentemente da abordagem metodológica, os padrões da biosfera rara procariótica no Oceano Ártico foram consistentes com os reportados em estudos anteriores. Nomeadamente, as massas de água foram identificadas como fatores determinantes para a composição da biosfera rara procariótica no Oceano Ártico.

Com base nas dificuldades que enfrentámos para conseguir distinguir *taxa* raros dos restantes, propusemos uma nova abordagem para a definição da biosfera rara microbiana na Parte III. Assim, em vez de aplicarmos um valor de abundância relativa (por exemplo, 0,1% de abundância relativa, por amostra) para distinguir *taxa* raros de abundantes, propusemos a aplicação de aprendizagem automática não-supervisionada para o agrupamento automático de *taxa* presentes numa comunidade, com base apenas na abundância, ou seja, sem utilizar um valor pré-definido (Capítulos 6 e 7). Para aplicar a nossa metodologia, desenvolvemos um pacote de R, *ulrb*, que significa "*Unsupervised Learning based Definition of the Rare Biosphere*". Demonstrámos, no Capítulo 6, que o método *ulrb* é aplicável a dados representativos de diferentes metodologias usadas para descrever microbiomas (*e.g.*, sequenciação da região V4V5 do gene do 16S rRNA ou sequenciação do gene inteiro), também demonstrámos que o método *ulrb* funciona igualmente para dados de comunidades não-microbianas e, por fim, mostrámos que *ulrb* obtém informação ecologicamente útil. Por exemplo, o método *ulrb* foi capaz de identificar corretamente em que situações certos simbiontes procariotas de hospedeiros marinhos (*e.g.*, corais) eram abundantes ou raros.

Finalmente, produzimos dados de comunidades procarióticas recolhidas sistematicamente do Oceano Ártico ao longo de vários anos (2016-2020) e aplicámos o método *ulrb* a esses dados para descrever a biosfera rara procariótica (Parte IV). Numa abordagem baseada em análise estatística clássica, identificámos que as variáveis mais determinantes para a diversidade procariótica foram a temperatura, a concentração de silicatos, as camadas pelágicas e as massas de água. É importante salientar que os nossos dados revelaram um padrão significativo de decréscimo de diversidade procariótica com o aumento da temperatura, que se deveu quase inteiramente à perda de microrganismos raros. No entanto, a perda de diversidade em função da temperatura observou-se ao longo da coluna da água, mas não ao longo dos anos (2016-2020). De seguida, com o objetivo de extrair associações entre combinações complexas de

variáveis ambientais e *taxa* raros, indeterminados e abundantes, aplicámos uma estratégia de aprendizagem automática com base em regras. Com efeito, aplicámos mineração de regras de associação (representam associações entre variáveis) e desenvolvemos um algoritmo para a seleção de regras de associação relevantes e não-redundantes entre si. Através da abordagem de aprendizagem automática com base em regras, demonstrámos que a biosfera rara procariótica se associa a combinações complexas de variáveis ambientais que são diferentes das combinações que se associam aos grupos de *taxa* indeterminados e abundantes, o que indica que a classificação obtida previamente com aprendizagem automática não-supervisionada foi, de facto, capaz de criar classificações com significado ecológico. Adicionalmente, a mineração de regras de associação foi mais útil para a análise da taxonomia dos procariotas identificados em conjunto com as variáveis ambientais do que as abordagens tradicionais de análise de dados usadas em ecologia microbiana. De facto, através da mineração de regras de associação, foi possível identificar associações complexas entre variáveis ambientais e *taxa* em qualquer nível taxonómico simultaneamente, o que é extremamente difícil com as abordagens clássicas de análise de dados taxonómicos.

Em suma, esta tese melhorou a nossa capacidade de produzir dados de comunidades microbianas com elevada qualidade e gerados de maneira sistemática, o que é um passo fundamental para se conseguir entender a resposta dos microbiomas às alterações climáticas no Oceano Ártico. Os nossos dados permitiram a identificação das variáveis ambientais mais determinantes da diversidade procariótica no Oceano Ártico, com um foco particular na biosfera rara. Esta tese combinou a utilização de aprendizagem automática não-supervisionada com a baseada em regras para obter novo conhecimento relativo à biosfera rara procariótica.

Palavras-chave: biosfera rara, Oceano Ártico, aprendizagem não-supervisionada, mineração de regras de associação, aprendizagem automática com base em regras, ecologia microbiana, gene do 16S rRNA, amostragem de água marinha, microbiologia marinha.

Abstract

The microbial rare biosphere represents the low abundance microorganisms within a given microbial community. The microbial rare biosphere is important for ecosystem resistance and resilience, because it works as a genetic reservoir with the potential to respond to environmental stressors. However, most microbial ecology studies still overlook the low abundance microorganisms, because they are usually considered to be less important for the overall community functioning. Despite the growing number of studies about the microbial rare biosphere in several environments, little is known about its potential role and functions in a climate change scenario. Several methodological challenges still hinder the advancement of knowledge regarding the microbial rare biosphere, especially regarding underexplored environments of great importance such as the Arctic Ocean. Some of the most important challenges in the study of the microbial rare biosphere are related with the availability of high-quality data that is inter-comparable across multiple years, which is a key variable to understand the ecology of the microbial rare biosphere under environmental change. Besides the acquisition and processing of the data, the data analysis itself presents non-trivial challenges, such as the methodology to delineate rare from non-rare taxa, but also the ability to analyze complex combinations of environmental variables. To advance solutions for these challenges and obtain novel ecological insights into the Arctic Ocean, this thesis systematically compared methodologies used to produce microbial community datasets, produced novel and standardized datasets, and advanced data interpretation of the microbial rare biosphere by fitting and analyzing machine learning models.

In the first part of this thesis, we covered the state of the art of the microbial rare biosphere, adopting an historical perspective to understand how the concept of rarity in microbiology emerged, showing that the study of the microbial rare biosphere was dependent on high-throughput sequencing strategies. The state-of-the-art also summarized the main ecological findings regarding the major microbial groups, which showed that most studies on the microbial rare biosphere focus on prokaryotes and protists. Furthermore, this review described some of the most important methodological advances in the study of the rare biosphere and showed that the different concepts used to explain microbial rare biosphere patterns can be integrated into a single conceptual framework.

The experimental work was divided into three distinct sections. Parts II and III primarily addressed methodological questions, while part IV integrated the methodological

advances from the earlier sections to enhance the ecological analysis of the microbial rare biosphere.

In part II, we focused on methodologies to describe marine microbial communities and their rare biosphere. To do so, we first focused on seawater sampling protocols (Chapter 3), which mirrored the protocols used in major sampling campaigns (*e.g.*, Tara Oceans, Ocean Sampling Day), using data from the Mediterranean Sea. Our results showed that the sampling volume and the types of filters used were not significantly changing the diversity outcome of either prokaryotic or protistic communities. However, using size-fractionated filtration significantly changed the community composition of protists. Then, we focused on evaluating the impact of using different microbiome DNA sequencing strategies and taxonomy databases in the description of Arctic prokaryotic communities (Chapter 4). Our results showed that using full-length 16S ribosomal RNA (rRNA) gene amplicon sequencing was more efficient for taxonomy attribution than using a small region of the same gene, but we also showed that the taxonomy database selection was as important as the sequencing strategy. Importantly, we showed that using short regions was still a valid approach for studies that focused primarily on high taxonomic levels (phylum and class), because there were no differences between sequencing strategies and taxonomy databases at those levels. Next, we evaluated different bioinformatics processing of raw reads from V4V5 16S rRNA gene and metagenome sequencing, and we also compared two different approaches to delineate rare from non-rare taxa, using an Arctic Ocean dataset. We showed that both Amplicon Sequence Variants (ASV/ASVs) and Operational Taxonomic Units (OTU/OTUs) were both valid approaches to analyze the prokaryotic rare biosphere. We also showed that introducing metagenomic data could improve such analyzes, by improving the identification of uncultivated taxonomic groups. Across the multiple bioinformatics variations used in Chapter 5, the prokaryotic rare biosphere patterns in the Arctic Ocean were consistent with previous literature from that region, showing that most rare prokaryotes tend to remain rare and that prokaryotic composition changes across water masses.

Based on the difficulties faced in the step of delineating rare from non-rare taxa, we proposed a new approach for the microbial rare biosphere definition in Part III. Instead of using a relative abundance threshold (*e.g.,* 0.1% relative abundance per sample) to distinguish between rare and abundant taxa, we proposed the utilization of unsupervised machine learning to automatically cluster taxa based on their abundance scores (Chapters 6 and 7). To do so, we built the *ulrb* R package, standing for “Unsupervised Learning based Definition of the Rare Biosphere”. We showed that *ulrb* was applicable

to datasets representing several methodological variations currently used in microbial ecology (*e.g.*, V4V5 and full-length 16S rRNA gene sequencing), and we also showed that *ulrb* can be extended to non-microbial data and can obtain useful ecological information. For example, *ulrb* was able to correctly identify taxa known to be abundant in association with marine hosts (*e.g.*, corals), but rare in the surrounding seawater or in necrotic host tissue.

Finally, we produced a standardized, multi-year (2016-2020) prokaryotic dataset of the Arctic Ocean and then used *ulrb* to define the rare biosphere (Part IV). By means of a classical statistical approach, we found that temperature, silicates, pelagic layers and water masses were potentially the most important variables shaping prokaryotic composition. Importantly, we showed that increasing temperatures lead to a sharp decrease in prokaryotic diversity, which was almost entirely relative to loss of rare prokaryotes, while the diversity of abundant prokaryotes remained constant. The loss of rare prokaryotes as a function of temperature was identified across the vertical water column but was not identified across multiple years (from 2016 to 2020). To be able to extract complex associations between environmental variables and rare, undetermined and abundant taxa, we employed rule-based machine learning. For that purpose, we employed association rule mining and developed an algorithm to identify informative and non-redundant association rules (association between multiple variables). The rule-based machine learning results showed that the prokaryotic rare biosphere was associated with a different set of environmental variables than undetermined and abundant taxa, which indicates that *ulrb* provides an ecologically meaningful definition of the rare biosphere. Furthermore, rule-based machine learning proved more useful to analyze the associations between prokaryotic taxa and environmental variables than the standard approaches currently used in microbial ecology. For instance, rule-based machine learning enabled us to analyze the association between complex combinations of environmental variables and taxa at any taxonomic level at once, which is extremely difficult with traditional methods.

In summary, this thesis improved our ability to produce high-quality and standardized datasets, which is a key step in understanding the microbial response to climate change in the Arctic Ocean. Such data allowed us to identify the most important environmental variables in shaping prokaryotic diversity in the Arctic Ocean, which was further enabled by improvements in data analysis. This thesis effectively combined unsupervised and rule-based machine learning to derive new knowledge on the prokaryotic rare biosphere.

Keywords: rare biosphere, Arctic Ocean, unsupervised learning, association rule mining, rule-based machine learning, machine learning, microbial ecology, 16S rRNA gene, seawater sampling, marine microbiology

Table of Contents

List of Tables ..... **xxvi**

List of Supplementary Tables ..... **xxvii**

List of Supplementary Data ..... **xxx**

List of Figures ..... **xxxi**

List of Supplementary Figures ..... **xxxiv**

List of Equations ..... **xxxix**

List of Abbreviations ..... **xl**

General Introduction **1**

1.1. Thesis overview 2

1.2. What is the microbial rare biosphere and why it is relevant 4

1.3. Methodological challenges in the study of marine microbial ecology 5

1.4. Challenges in the definition of the microbial rare biosphere 6

1.5. The Arctic Ocean as a case study 7

1.6. Objectives 8

1.7. General experimental design, statistical approach and data 13

Part I. State of the art. **21**

Chapter 2. 22

2. The microbial rare biosphere: current concepts, methods and ecological principles 23

2.1. Introduction 24

2.2. An historical perspective on the concept of biological rarity 27

2.3. The prokaryotic and microeukaryotic rare biosphere 30

2.4. Viral control of rare microbial populations 32

2.5. Rare biosphere and host associations 34

2.6. Methodological assessment of the microbial rare biosphere 37

2.7. Fundamental ecological principles and dynamics of the microbial rare biosphere 46

2.8. Conclusion and open questions 51
2.9. Declarations 52
Part II. Advances in marine microbial ecology methods: from seawater sampling to bioinformatics. 53
Chapter 3 54
3. Inter-comparison of marine microbiome sampling protocols 55
3.1. Introduction 57
3.2. Methods 60
3.2.1. Seawater sampling 60
3.2.2. DNA extraction, amplification, and sequencing 63
3.2.3. Bioinformatics processing of raw sequences 64
3.2.4. Abundance table pre-processing 65
3.2.5. Statistical analysis 66
3.3. Results 68
3.3.1. Environmental communities 68
3.3.2. The effect of seawater filtered volume in marine microbial diversity, accounting for filter pore size 69
3.3.2.1. Prokaryotes 69
3.3.2.2. Protists 75
3.4. Discussion 81
3.4.1. The effect of methodological variation in common seawater sampling protocols 81
2.4.2. Consistency between sequencing strategies 83
3.4.3. Strengths and limitations of the EMOSE dataset 84
3.5. Conclusion 86
3.6. Declarations 87
3.7. Supplementary Data 89
Chapter 4 90

4. Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities 91

4.1. Introduction ... 93

4.2. Methods ... 95

4.3. Sampling campaign ... 95

4.2. DNA extraction, amplification, and sequencing ... 97

4.2.3. Bioinformatics processing of reads ... 97

4.2.4. Taxonomic classification ... 97

4.2.5. Statistical analysis and data visualization ... 98

4.3. Results ... 99

4.3.1 Overview of sequencing results ... 99

4.3.2. Effectiveness of taxonomic classification ... 99

4.3.3. Alpha and beta diversity ... 101

4.4. Discussion ... 106

4.5. Conclusions ... 110

4.6. Declarations ... 111

Chapter 5 ... 112

5. Exploration of the types of rarity in the Arctic Ocean from the perspective of multiple methodologies ... 113

5.1. Introduction ... 114

5.2. Methods ... 117

5.2.1. Environmental context ... 117

5.2.2. Seawater sampling and processing for microbiome analysis ... 119

4.2.3. Bioinformatics processing of 16S rRNA gene and total community DNA sequences ... 119

5.2.4. Alpha and Beta diversity calculation ... 120

5.2.5. Rare Biosphere definition and calculation of types of rarity ... 120

5.3. Results ... 122

5.3.1. Rare prokaryotic alpha and beta diversity 122
5.3.2. Different types of prokaryotic rarity 127
4.3.3. Rare prokaryotic taxonomy 129
5.4. Discussion 134
5.4.1. The prokaryotic rare biosphere of the Arctic Ocean is sensitive to environmental heterogeneity 134
5.4.2. Methodological strategies and trade-offs 136
5.5. Conclusion 139
5.6. Declarations 140
Part III. Development of an unsupervised machine learning approach to define the microbial rare biosphere. **142**
Chapter 6 143
6. Definition of the microbial rare biosphere through unsupervised machine learning 144
6.1. Introduction 145
6.2. Methods 148
6.2.1. The *ulrb* algorithm 148
6.2.1.1. *ulrb* algorithm description 148
6.2.1.2. *ulrb* R package construction and utilization 149
6.2.1.3. Unsupervised learning statistics 151
6.2.1.3.1. Silhouette score 151
6.2.2. Datasets 152
6.2.2.1. N-ICE dataset 156
6.2.2.2. MOSJ 2019 dataset 156
6.2.2.3. Coral microbiome dataset 156
6.2.2.4. Ants dataset 157
6.2.2.5. BCI dataset 157
6.2.2.6. MOSJ 2016-2020 dataset 157
6.2.2.6.1. Sampling and sequencing details 157

6.2.2.6.2. Processing of V4V5 16S rRNA gene amplicons 158
6.2.2.6.3. Combining sequencing data from different years (2016-2020) into a single dataset 158
6.2.3. Types of rarity 159
6.2.4. Statistical analysis, visualization and reproducibility 159
6.3. Results 160
6.3.1. Testing *ulrb* for different kinds of phylogenetic units 160
6.3.2. Testing *ulrb* for different amplicon sequencing strategies 163
6.3.3. Verifying robustness of *ulrb* against sample size, sequencing depth and number of taxa 166
6.3.4. Validating *ulrb* for non-microbial datasets 169
6.3.5. Using *ulrb* to establish types of rarity in a coral microbiome 170
6.4. Discussion 172
6.5. Conclusion 177
6.6. Declarations 178
6.7. Supplementary Methods 180
6.7.1. Unsupervised Learning Statistics 180
6.7.1.1. Calinski-Harabasz index 180
6.7.1.2. Davies-Bouldin index 180
6.7.2 Datasets 181
6.7.2.1. N-ICE 181
6.7.2.2. MOSJ 2019 dataset 183
6.7.2.3. Coral microbiome dataset 184
Chapter 7 185
7. *ulrb*: an R Package to Automatically Define Rare Taxa 186
7.1. Introduction 187
7.2. The *ulrb* R package 189
7.2.1. Description of the rare biosphere definition problem in machine learning terms 189

7.2.2. Unsupervised machine learning algorithm and features selection 189
7.2.3. Number of clusters (k) used 190
7.2.4. *ulrb* architecture 191
7.2.4.1. Dependencies 192
7.2.4.2. Functions' syntax and parameters 192
7.2.4.2. Data preparation 197
7.2.4.3. Classification task 198
7.2.4.4. Automatic selection of k 199
7.2.4.5. Unsupervised clustering statistics 199
7.2.4.6. Visualization 200
7.2.5. Implementation of pam to any number of samples by define_rb() function 200
7.2.6. Explanation of evaluation metrics 201
7.2.6.1. Silhouette score and average Silhouette score 201
7.2.6.2. Calinski-Harabasz index 202
7.2.6.3. Davies-Bouldin index 202
7.2.7. Data available in *ulrb* R package 203
7.3. Results 204
7.3.1. Standard example of *ulrb* application 204
7.3.2 Inspecting alternative number of classifications 206
7.4. Summary and conclusion 212
7.5. Declarations 213
Part IV. Ecology of the microbial rare biosphere in the Arctic Ocean. **214**
Chapter 8 215
8. Combining unsupervised and rule-based machine learning to obtain novel insights into the prokaryotic rare biosphere response to a changing Arctic Ocean 216
8.1. Introduction 218
8.2. Methods 222

8.2.1. Sampling campaign and available metadata 222
8.2.2. Seawater sampling, DNA extraction and amplicon sequencing 223
8.2.3. Bioinformatic processing of raw reads and taxonomic assignments 224
8.2.4. ASV abundance table pre-processing 224
8.2.5. Definition of the rare biosphere with unsupervised machine learning 225
8.2.6. Rule-based machine learning with association rule mining 226
8.2.7. Statistical evaluation of association rules 229
8.2.8. Statistical analysis and data visualization 231
8.3. Results 233
8.3.1. Overview of prokaryotic communities from 2016 to 2020, across pelagic layers 233
8.3.2. Diversity as a function of distinct environmental variables 238
8.3.3. Uncovering complex combinations between prokaryotes and multiple environmental variables 241
8.3.4. Fine-grained analysis of taxonomy with association rule mining 248
8.3.5. Monitoring targeted taxonomic groups by combining unsupervised and rule-based machine learning 250
8.4. Discussion 253
8.4.1. Prokaryotic composition follows the highly stratified water column 253
8.4.2. Rare, undetermined, and abundant taxa follow distinct environmental patterns 255
8.4.3. Targeted taxonomic analysis 257
8.4.4. Assessing types of rarity by monitoring targeted taxa 260
8.5. Conclusion and future work 262
8.6. Declarations 264
8.7. Supplementary Data 265
General Discussion and Conclusion 266
9.1. General remarks 267
9.2. Major findings and contributions 267

9.2.1. Insights from reviewing literature 268
9.2.2. Standardization enables the creation of modular, continuously growing multi-year datasets 269
9.2.3. Towards a solution to the microbial rare biosphere definition problems 270
9.2.4. Identification of environmental variables shaping the rare biosphere in the Arctic Ocean 272
9.2.5. Combining unsupervised and rule-based machine learning for in-depth analysis of microbial communities 272
9.3. Experimental limitations 273
9.3.1. High dimensionality and imbalanced datasets 273
9.3.2. Phylogenetic signal, community assembly theory, and classification of the types of rarity 275
9.3.3. Fast-paced sequencing advances 276
9.4. Open questions and future perspectives 278
9.4.1. Continued methodological improvements 278
9.4.2. Data science potential to boost microbial ecology studies 278
9.4.3. The Arctic Ocean as a laboratory for accessing climate change impacts 281
9.4.4. The microbial rare biosphere in the microbial ecology landscape 281
9.5. Final remarks 283
References 284
Attachments 338
10. Supplementary Figures 339
10.1. Supplementary Figures of Chapter 3 339
10.2. Supplementary Figures of Chapter 4 342
10.3. Supplementary Figures of Chapter 5 348
10.4. Supplementary Figures of Chapter 6 361
10.5. Supplementary Figures of Chapter 7 367
10.6. Supplementary Figures of Chapter 8 368

11. Supplementary Tables 378
11.1. Supplementary Tables of Chapter 3 378
11.2. Supplementary Tables of Chapter 4 399
11.3. Supplementary Tables of Chapter 5 408
11.4. Supplementary Tables of Chapter 6 411
11.5. Supplementary Tables of Chapter 8 423

List of Tables

Table 1.1. Summary of datasets used through this thesis and the main objective within each chapter. 19
Table 2.1. Recent publications applying advanced methodologies to access microbial rare biosphere diversity. 39
Table 2.2. Recent publications applying advanced methodologies to access microbial rare biosphere functions. 40
Table 4.1. Metadata of MOSJ 2019 samples that were used. 96
Table 5.1. Spatial and temporal characteristics of the nine sampling points, according to de Sousa *et al.*, (2019) and using the same sample IDs. 118
Table 5.2. Selected thresholds of rarity obtained from the MultiCoLA algorithm. 121
Table 5.3. List of the presence or absence of candidate phyla from within and outside the candidate phyla radiation(2016). 129
Table 6.1. Summary of datasets used in this study. 154
Table 8.1. Reference values for the discretization of quantitative variables, relative to nutrient concentrations. 228
Table 8.2. Reference values for the discretization of temperature (ºC) quantitative values. 229

List of Supplementary Tables

Supplementary Table 11.1. Details of sampling scheme of EMOSE dataset. 378
Supplementary Table 11.2. List of samples from the EMOSE dataset that were discarded in the rarefaction step. 382
Supplementary Table 11.3. Summary of the sequencing results of the EMOSE dataset. 384
Supplementary Table 11.4. Number of prokaryotic taxonomic lineages obtained in each sample (species richness) of the EMOSE dataset. 385
Supplementary Table 11.5. Details on the statistical tests for the prokaryotic species richness 390
Supplementary Table 11.6 – Details on the PERMANOVA analysis of prokaryotic communities and betadisper permutations for verification of homogeneity of variance prerequisite. 391
Supplementary Table 11.7. Details of the Tukey test for significance between differences in prokaryotic communities' distance to centroids of size fractions, calculated by betadisper. 392
Supplementary Table 11.8. Number of protist taxonomic lineages obtained in each sample (Species richness). 392
Supplementary Table 11.9. Details on the statistical tests for the protists species richness 396
Supplementary Table 11.10. Details on the PERMANOVA analysis of protist communities and betadisper permutations for verification of variance prerequisite. .. 397
Supplementary Table 11.11. Details of the Tukey test for significance between differences in protist communities' distance to centroids of size fractions, calculated by betadisper. 398
Supplementary Table 11.12. List of samples removed (< 10 000 reads) of MOSJ 2019. 399
Supplementary Table 11.13. Summary statistics of sequencing results of MOSJ 2019. 399
Supplementary Table 11.14. Details of Mann-Whitney U test , used to compare the distribution of two independent variables. 400
Supplementary Table 11.15. Details on linear regressions, including p-values and R-squared values. 402
Supplementary Table 11.16. Species with binomial name with more than 1 ASV, identified with GTDB and full-length 16S rRNA gene sequencing. 406

Supplementary Table 11.17. Details of PERMANOVA tests. 407
Supplementary Table 11.18. Summary of alpha diversity values for the ASVs, OTUs and mOTUs datasets, for the total and rare communities, with and without rarefaction, for the rarity definitions derived from the MultiCoLA algorithm and from the 0.1%, per sample, definition 408
Supplementary Table 11.19. Summary statistics of sequencing results from V4V5 16S rRNA gene amplicon (ASV and OTU) and shotgun metagenomic (mOTU) sequencing of the N-ICE 2015 samples. 411
Supplementary Table 11.20. Summary statistics of sequencing results from V4V5 and full-length 16S rRNA gene amplicon sequencing of the MOSJ 2019 samples. 411
Supplementary Table 11.21. Sampling metadata of the MOSJ 2016-2020 samples that were used for analysis. 412
Supplementary Table 11.22. Sequencing statistics from V4V5 16S rRNA gene amplicon sequencing of the MOSJ samples from 2016 to 2020 416
Supplementary Table 11.23. List of approaches used to define the rare biosphere in literature, from 2006 to 2024 (July). 416
Supplementary Table 11.24. Statistics on the number of reads before and after major processing steps, and taxonomic groups obtained (MOSJ 2016-2020). 423
Supplementary Table 11.25. Summary of two-way ANOVA results and Levene test, comparing the number of taxa across year and pelagic layers. 423
Supplementary Table 11.26. *Post-hoc* test for two-way ANOVA, using pairwise t-test, comparing the number of taxa between each combination of year and of pelagic layers. 424
Supplementary Table 11.27. Shapiro test results for number of taxa within each combination of year and pelagic layers 425
Supplementary Table 11.28. Linear least squares model results for number of taxa as a function of years, accounting for pelagic layers. 425
Supplementary Table 11.29. Poison model results for number of taxa as a function of years, accounting for pelagic layers. 426
Supplementary Table 11.30. Pearson and Kendal correlations between the number of taxa and years variable, accounting for pelagic layers. 426
Supplementary Table 11.31. Relative abundance of phylum and class level, across the whole MOSJ 2016-2020 dataset (n = 117). 428
Supplementary Table 11.32. PERMANOVA test results, comparing community composition between years and pelagic layers. 431

Supplementary Table 11.33. Linear least squared model results for number of taxa as a function of temperature and depth, accounting for pelagic layers, water masses and region. 431
Supplementary Table 11.34. Poisson model results for number of taxa as a function of temperature and depth, accounting for pelagic layers, water masses and region..... 432
Supplementary Table 11.35. Pearson and Kendal correlations between the number of taxa and temperature and depth, accounting for pelagic layers..... 433
Supplementary Table 11.36. Linear least squared model results for number of taxa as a function of silicate concentration (µM). 434
Supplementary Table 11.37. Poisson model results for number of taxa as a function of silicate concentration (µM). 434
Supplementary Table 11.38. Pearson and Kendall correlation between the number of taxa and silicate concentration (µM). 435
Supplementary Table 11.39. Top 100 association rules, ordered by descending Lift. 435
Supplementary Table 11.40. Dependent rules, with Lift above 15, ordered by Conviction. 441
Supplementary Table 11.41. Non-redundant and dependent rules, independent of Lift. 442
Supplementary Table 11.42. Top 100 rules with Conviction above 5, Confidence above 90%, and Lift above 1..... 444
Supplementary Table 11.43. Top100 rules, with Conviction above 5, Confidence above 90%, Lift above 1, and selected non-redundant rules based on Complexity. 449
Supplementary Table 11.44. Non-redundant and dependent rules that are taxonomically relevant. 461

List of Supplementary Data

Supplementary Data 3.1. Detailed information regarding the sampling campaign. 89
Supplementary Data 3.2. Report of the number of reads after each step of the MGnify platform for each sample and links for quality reports and public repository. 89
Supplementary Data 8.1. Sample metadata, adapted from Pascoal *et al*. (2025a). Nutrient measurements from Wold *et al*., (2024). 265
Supplementary Data 8.2. Dependent rules, for any Lift value, ordered by Conviction. 265

List of Figures

Figure 1.1. Schematic representation of the connection between the chapters of this thesis and their outcomes. 4
Figure 2.1. Conceptual illustration of a rank abundance curve. 25
Figure 2.2. Representation of the species abundance distribution curve 28
Figure 2.3. Schematic representation of how different ecological theories and concepts are intertwined. 48
Figure 3.1. Schematic representation of sampling campaign. 62
Figure 3.2. Overview of the prokaryotic species richness obtained. 70
Figure 3.3. Detailed comparison of the prokaryotic species richness across methodological variables. 71
Figure 3.4. Prokaryotic community composition overview. n 73
Figure 3.5. Prokaryotic species richness per taxonomic group, from MetaB16SV4V5. 74
Figure 3.6. Detailed analysis of *Candidatus* Marinimicrobia, from MetaB16SV4V5. 75
Figure 3.7. Overview of protist species richness obtained. 76
Figure 3.8. Detailed comparison of the protist species richness across methodological variables..... 77
Figure 3.9. Protist community composition overview. 79
Figure 3.10. Protist species richness per taxonomic group, from MetaB18SV9. 80
Figure 4.1. Map of sampling campaign. 95
Figure 4.2. The percentage of ASVs classified at each taxonomic level. 100
Figure 4.3. Relative abundance of ASVs classified at species level. 101
Figure 4.4. Alpha diversity scores across depth. 102
Figure 4.5. ASV richness at phylum, order, genus and species taxonomic levels across methodological combinations. 103
Figure 4.6. Beta diversity at phylum and species level. 105
Figure 5.1. Alpha diversity for the total community (non-rarefied) and for the rare biosphere, as defined by the MultiCoLA algorithm, across ASVs, OTUs and mOTUs123
Figure 5.2. Hierarchical clustering (a-c) and ordination analysis (d-f) for the rare biosphere as defined by MultiCoLA and using non-rarefied data. 125
Figure 5.3. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. 128
Figure 5.4. Circular visualizations of the ASVs taxonomy, abundance and types of rarity. 131

Figure 5.5. Circular visualizations of the OTUs taxonomy, abundance and types of rarity. 132
Figure 5.6. Circular visualizations of the mOTUs taxonomy, abundance and types of rarity. 133
Figure 6.1. Schematic representation of k-medoids. 149
Figure 6.2. Schematic representation of the information that the Silhouette scores can provide. 152
Figure 6.3. RAC and Silhouette score plots for N-ICE dataset. 160
Figure 6.4. Comparison of number of rare, undetermined (if applicable) and abundant phylogenetic units from ASVs, OTUs, and mOTUs. 162
Figure 6.5. Comparison of ASVs derived from V4V5 and full-length 16S rRNA gene. 163
Figure 6.6. Comparison of number of rare, undetermined (if applicable) and abundant ASVs from V4V5 and full-length 16S rRNA gene sequencing. 165
Figure 6.7. Quality of *ulrb* clustering measured by the average Silhouette score as a function of number of samples, ASVs, and sequencing depth. 167
Figure 6.8. Analysis of *ulrb* applicability to Ants and BCI datasets. 169
Figure 6.9. Monitoring of the abundance classifications of selected OTUs. 171
Figure 7.1. Schematic representation of *ulrb* architecture. 192
Figure 7.2. Schematic representation of wide and long table format in the context of abundance tables 197
Figure 7.3. Default plot from plot_*ulrb*() function. 205
Figure 7.4. Example of plot_*ulrb*() output for one single sample. 206
Figure 7.5. Standard plot_*ulrb*() output for user-defined classification 207
Figure 7.6. Manual evaluation of k using the function evaluate_k(). 209
Figure 7.7. OTU abundance score, on a Log10 scale, ordered from the most to the least abundant. 211
Figure 8.1. Map of the position of the sampling stations between the Fram Strait and Svalbard archipelago. 223
Figure 8.2. Boxplots of the number of taxa present per sample (alpha diversity estimate) as a function of sampling year. 233
Figure 8.3. Mean ± standard deviation (sd) of relative abundance (%) in Log10 scale of phyla across year and pelagic layers. 235
Figure 8.4. Beta diversity analysis of taxonomic groups, with nonmetric multidimensional scaling (nMDS) represented in an ordination plot. 236

Figure 8.5. Number of rare, undetermined and abundant taxa, across years and pelagic layers. 237
Figure 8.6. Number of taxa present in each sample (alpha diversity estimate) as a function of temperature (ºC). 238
Figure 8.7. Number of taxa present in each sample (alpha diversity estimate) as a function of silicate concentration (µM). 240
Figure 8.8. Confidence as a function of Conviction of rules with antecedents statistically dependent on their consequences. 243
Figure 8.9. Visual representation of association rules (n = 593 rules). 246
Figure 8.10. Abundance classification of target taxonomic groups. 252
Figure 9.1. Schematic representation of MOSJ 2016-2020 modularity. 270
Figure 9.2. Schematic representation of symmetry between unsupervised and rule-based machine learning approaches taken. 273

List of Supplementary Figures

Supplementary Figure 10.1. Pairwise Pearson correlation between alpha metrics.... 339
Supplementary Figure 10.2. Rarefaction curves by size fraction. 340
Supplementary Figure 10.3. Rarefaction curves by filtered volume. 340
Supplementary Figure 10.4. Prokaryotic species richness per taxonomic group, from metagenomes. 341
Supplementary Figure 10.5. Protist species richness per taxonomic group, from MetaB18SV9..... 341
Supplementary Figure 10.6. Rarefaction curves for all samples from V4V5 (left) and full-length (right) 16S rRNA gene amplicon sequencing. 342
Supplementary Figure 10.7. Rarefied version of Figure 4.2, with percentage of ASVs classified at each taxonomic level. 342
Supplementary Figure 10.8. Rarefied version of Figure 4.4, with alpha diversity scores across depth. 343
Supplementary Figure 10.9. Alpha diversity scores across depth. 344
Supplementary Figure 10.10. Rarefied version of Supplementary Figure 4.4, with alpha diversity scores across depth. 345
Supplementary Figure 10.11. Rarefied version of Figure 4.5, with ASV richness at phylum, order, genus and species taxonomic levels across methodological combinations..... 346
Supplementary Figure 10.12. Beta diversity at order and genus level..... 347
Supplementary Figure 10.13. Quality scores profile for the forward 16S rRNA gene amplicon sequencing reads of N-ICE 2015. 348
Supplementary Figure 10.14. Quality scores profile for the reverse 16S rRNA gene amplicon sequencing reads of N-ICE 2015. 349
Supplementary Figure 10.15. MultiCoLA results of Procrustes and boxplot with the relative abundance equivalent of the selected cutoff, for non-rarefied data 350
Supplementary Figure 10.16. MultiCoLA results of Procrustes for rarefied data..... 351
Supplementary Figure 10.17 – Rarefaction curves for ASVs, OTUs, and mOTUs of N-ICE 2015. 351
Supplementary Figure 10.18. Alpha diversity for the total community (rarefied) and for the rare biosphere, defined by the MultiCoLA algorithm, across ASVs, OTUs and mOTUs..... 352

Supplementary Figure 10.19. Alpha diversity for the total community (non-rarefied) and for the rare biosphere, defined by the 0.1% relative abundance threshold, per sample, across ASVs, OTUs and mOTUs. 353
Supplementary Figure 10.20. Alpha diversity for the total community (rarefied) and for the rare biosphere, defined by the 0.1% relative abundance threshold, per sample, across ASVs, OTUs and mOTUs. 354
Supplementary Figure 10.21. Dendrograms and multivariate analysis for the rare biosphere, defined by MultiCoLA (rarefied). 355
Supplementary Figure 10.22. Dendrograms and multivariate analysis for the rare biosphere, defined by 0.1% relative abundance threshold, per sample (non-rarefied) 356
Supplementary Figure 10.23. Dendrograms and multivariate analysis for the rare biosphere, defined by 0.1% relative abundance threshold, per sample (rarefied)..... 357
Supplementary Figure 10.24. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. 358
Supplementary Figure 10.25. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. Rare biosphere was defined by the 0.1% relative abundance threshold, per sample, from non-rarefied data. 359
Supplementary Figure 10.26. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. Rare biosphere was defined by the 0.1% relative abundance threshold, per sample, from rarefied data. 360
Supplementary Figure 10.27. – Schematic representation of RAC and of the problem with threshold-based definitions of the rare biosphere. 361
Supplementary Figure 10.28. Aggregate quality profiles of FASTQ files (Forward and Reverse) with V4V5 16S rRNA gene amplicon sequencing results from MOSJ 2016-2020. 362
Supplementary Figure 10.29. Rarefaction curves for the entire MOSJ dataset, grouped by years (from 2016 to 2020, totaling 117 samples). 362
Supplementary Figure 10.30. Evaluation of clusters obtained for ASVs, OTUs, and mOTUs, based on Silhouette scores. 363
Supplementary Figure 10.31. *FuzzyQ* results for different phylogenetic units (ASV, mOTU and OTU). 363
Supplementary Figure 10.32. Evaluation of clusters obtained for ASVs derived from V4V5 and full-length 16S rRNA gene, based on Silhouette scores. 364
Supplementary Figure 10.33. FuzzyQ results for V4V5 and full-length 16S rRNA gene sequencing. 364

Supplementary Figure 10.34. Illustration of RAC obtained by different number of randomly selected ASVs. 365
Supplementary Figure 10.35. Quality of *FuzzyQ* clustering measured by the average Silhouette score as a function of number of samples, ASVs, and sequencing depth. 365
Supplementary Figure 10.36. Schematic representation of the main differences between *ulrb* and *FuzzyQ*. 366
Supplementary Figure 10.37. Example of plot_*ulrb*() output for a single sample, using five classifications. 367
Supplementary Figure 10.38. Print screen of data frame obtained by setting with_plot = FALSE in the evaluate_k() function 367
Supplementary Figure 10.39. Relative abundance (%) of taxonomic groups, in each sample. 368
Supplementary Figure 10.40. Mean ± standard deviation (sd) of relative abundance (%) in Log10 scale of class across year and pelagic layers 369
Supplementary Figure 10.41. Number of taxa present in a given sample (alpha diversity estimate) as a function of temperature (ºC), on the left side, and depth (m) in Log10 scale, on the right side. 369
Supplementary Figure 10.42. Temperature (ºC) as a function of year (from 2016 to 2020), grouped by pelagic layer. 370
Supplementary Figure 10.43. Number of taxa as a function of temperature (ºC). 370
Supplementary Figure 10.44. Number of taxa as a function of temperature (ºC). 371
Supplementary Figure 10.45. Silicate concentration (µM) as a function of depth (m) in Log10 scale. 371
Supplementary Figure 10.46. Number of taxa present in a given sample (alpha diversity estimate) as a function of salinity concentration (PSU). 372
Supplementary Figure 10.47. Number of taxa present in a given sample (alpha diversity estimate) as a function of NO_2^- concentration (µM). 372
Supplementary Figure 10.48. Number of taxa present in a given sample (alpha diversity estimate) as a function of NO_3^- concentration (µM). 373
Supplementary Figure 10.49. Number of taxa present in a given sample (alpha diversity estimate) as a function of NH_4^+ concentration (µM). 373
Supplementary Figure 10.50. Number of taxa present in a given sample (alpha diversity estimate) as a function of chlorophyll *a* concentration (mg/m³). 374
Supplementary Figure 10.51. Confidence as a function of Conviction, for all rules obtained (n = 18 550 rules). 375
Supplementary Figure 10.52. Confidence as a function of Conviction for all rules. ... 375

Supplementary Figure 10.53. Confidence as a function of Conviction of all rules that had antecedents statistically dependent on their consequences (n = 593 rules). 376
Supplementary Figure 10.54. Boxplots with Lift values obtained for rules with each possible classification in the consequent: rare, undetermined or abundant. 376
Supplementary Figure 10.55. Abundance classifications of target taxa. 377

List of Supplementary Methods Figures

Supplementary Methods Figure 6.1. Rarefaction curves for N-ICE dataset samples 182
Supplementary Methods Figure 6.2. Rarefaction curves for all MOSJ 2019 dataset samples..... 183

List of Equations

Equation 6.1. Silhouette score formula. 151
Equation 6.2. Calinski-Harabasz index formula. 180
Equation 6.3. Davies and Bouldin index formula. 180
Equation 7.1. Silhouette score formula. 201
Equation 7.2. Calinski-Harabasz index (simplified). 202
Equation 7.3. Davies-Bouldin index. 202
Equation 8.1. Database of transactions. 226
Equation 8.2. Support of an itemset. 227
Equation 8.3. Support of an association rule. 227
Equation 8.4. Confidence of an association rule. 227
Equation 8.5. Conviction of an association rule. 229
Equation 8.6. Lift of an association rule. 230
Equation 9.1. Number of possible combinations of abundance classifications. 279

List of Abbreviations

ALESSANDRO VEZI	AV
ALEX MITCHELL	AM
AMMONIUM	NH_4^+
AMPLICON SEQUENCE VARIANTS	ASV/ASVs
ANA MARIA DĄBROWSKA	AMD
ANALYSIS OF VARIANCE	ANOVA
ANTONIO FERNANDEZ-GUERRA	AFG
ASSOCIATION RULE	AR
ASSOCIATION RULE 1	AR1
ASSOCIATION RULE 2	AR2
ASSOCIATION RULE 3	AR3
ASSOCIATION RULE 4	AR4
ASSOCIATION RULE 5	AR5
ASSOCIATION RULE 6	AR6
ASSOCIATION RULE 7	AR7
ASSOCIATION RULE 8	AR8
ASSOCIATION RULE 9	AR9
ASSOCIATION RULE 10	AR10
ASSOCIATION RULE 11	AR11
ASSOCIATION RULE MINING	ARM
ATLANTIC WATER	AW
BARRO COLORADO ISLAND TREE COUNTS	BCI
BROAD-RANGE	BR
BRYAN WILSON	BW
CALINSKI-HARABASZ INDEX	CH
CANDIDATE PHYLA RADIATION	CPR
CANONICAL CORRESPONDENCE ANALYSIS	CCA
CATALYZED REPORTER DEPOSITION-FLUORESCENCE IN SITU HYBRIDIZATION	CARD-FISH
CATARINA MAGALHÃES	CM
CENTRO INTERDISCIPLINAR DE INVESTIGAÇÃO MARINHA E AMBIENTAL	CIIMAR
CHLOROPHYLL A	Chl *a*
CIRCULAR CONSENSUS SEQUENCING	CCS
COMPLEMENTARY DNA	cDNA
CONDITIONALLY RARE TAXA	CRT
CORRESPONDENCE ANALYSIS	CA
DAVIES-BOULDING INDEX	DB
DEEP CHLOROPHYLL MAXIMUM	DCM
DEGREES OF FREEDOM	df/DF
DOUBLE STRANDED DNA	dsDNA
ENGINEERING RESEARCH COUNCIL OF CANADA	NSERC

ENVIRONMENTAL MONITORING OF SVALBARD AND JAN MAYEN	MOSJ
EUNICELLA GAZELLA	EG
EUNICELLA VERRUCOSA	EV
EUROPEAN MARINE OMICS BIODIVERSITY OBSERVATORY NETWORK	EMO BON
EUROPEAN NUCLEOTIDE ARCHIVE	ENA
EUROPEAN OPEN SCIENCE EXPLORATION INITIATIVE	EMOSE
EUROPEAN REGIONAL DEVELOPMENT FUND	ERDF
FABIO DE PASCALE	FDP
FABIO ROCHA JIMENEZ VIEIRA	FRJV
FACULTY OF SCIENCES OF THE UNIVERSITY OF PORTO	FCUP
FRANCESCA MALFATTI	FM
FRANCISCO PASCOAL	FP
FUNDAÇÃO PARA A CIÊNCIA E TECNOLOGIA	FCT
GENOME TAXONOMY DATABASE	GTDB
GRAZIA MARINA QUERO	GMQ
HIGH-SENSITIVITY	HS
HISASHI ENDO	HE
INSTITUTO DE BIOENGENHARIA E BIOCIÊNCIAS	iBB
INTEGRATED MICROBIOME RESOURCE	IMR
INTERMEDIATE WATER	IW
INTERNAL TRANSCRIBED SPACER	ITS
INTERNATIONAL CENSUS OF MARINE MICROBES	ICoMM
INTERQUARTILE RANGE	IQR
INTITUTUE FOR HEALTH AND BIOECONOMY	i4HB
JED A. FUHRMAN	JAF
JULIE POLAIN	JP
LEFT HAND SIDE	LHS
LEPTOGORGIA SARMENTOSA	LS
LOCAL WATER	LW
LUÍS TORGO	LT
MARIA PAOLA TOMASINO	MPT
METABARCODING, USING V4V5 16S RRNA GENE	MetaB16SV4V5
METABARCODING, USING V9 18S RRNA GENE	MetaB18SV9
METAGENOME	MetaG
METAGENOME ASSEMBLED GENOME	MAG/MAGs
METAGENOME DERIVED OTU	mOTU/mOTUs
METAGENOMICS FOR HUMAN INTESTINAL TRACT	MetaHIT
MODIFIED ATLANTIC WATER	MAW
MULTIVARIATE CUTOFF LEVEL ANALYSIS	MultiCoLA
NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION	NCBI
Next Generation Sequencing	NGS
NICOLAS HENRY	NH

NITRATE	NO_3^-
NITRITE	NO_2^-
NON-METRIC MULTIDIMENSIONAL SCALING	nMDS
NORWEGIAN POLAR INSTITUTE	NPI
NORWEGIAN YOUNG SEA ICE EXPEDITION	N-ICE
NOT APPLICABLE/NOT AVAILABLE	NA/Na
NOT SIGNIFICANT	ns/NS
OCEAN SAMPLING DAY	OSD
OCEANOLOGICAL OBSERVATORY OF BANYULS-SUR-MER	OOB
OPERATIONAL TAXONOMIC UNITS	OTU/OTUs
PABLO SÁNCHEZ	PS
PARTITIONING AROUND MEDOIDS	pam
PAULA BRANCO	PB
PEDRO DUARTE	PD
PERMANENTLY RARE TAXA	PRT
PERMUTATIONAL MULTIVARIATE ANALYSIS OF VARIANCE	PERMANOVA
PHILIPP ASSMY	PA
PHOSPHATE	PO_4^-
PIERRE E. GALAND	PEG
POLUAR SURFACE WATER	PSW
POLYMERASE CHAIN REACTION	PCR
PORTUGUESE POLAR PROGRAM	PROPOLAR
PRACTICAL SALINITY UNITS	PSU
PROGRAMA OPERACIONAL REGIONAL DO NORTE 2020	NORTE 2020
QUANTITATIVE PCR	qPCR
RAMIRO LOGARES	RL
RANK ABUNDANCE CURVE	RAC
REDUNDANCY ANALYSIS	RDA
RIBOSOMAL DATABASE PROJECT	RDP
RIBOSSOMAL RNA	rRNA
RIGHT HAND SIDE	RHS
RNA POLYMERASE SUBUNIT B GENE	*ropB*
ROBERTA PIREDDA	RP
RODRIGO COSTA	RC
SEAWATER	SW
SEDIMENT	SD
SERVICE D'OBSERVATION DU LABORATOIRE ARAGO	SOLA
SILICA	Si
SINGLE AMPLIFIED GENOMES	SAG/SAGs
SMALL SUBUNIT OF RIBOSOMAL RNA	SSU rRNA
STÉPHANE PESANT	SP
SURFACE WATER	SW
THE COMPREHENSIVE R ARCHIVE NETWORK	CRAN

TIMOTEJ TURK DERMASTIA	TTD
TINKARA TINTA	TT
TRANSFORMED ATLANTIC WATER	TAW
TRANSIENTLY RARE TAXA	TRT
UNIVERSITY OF PORTO	UP
UNSUPERVISED LEARNING BASED DEFINITION OF THE RARE BIOSPHERE	*ulrb*
UNWEIGHTED PAIR GROUP METHOD WITH ARITHMETIC MEAN	UPGMA
WARM POLAR SURFACE WATER	PSWw
WINTER COOLED WATER	WCW
β NEAREST TAXON INDEX	βNTI

General Introduction

1. General Introduction

1.1. Thesis overview

This thesis provides an in-depth analysis of methodological aspects of the microbial rare biosphere assessment and shows how methodological improvements allow novel ecological insights on marine microbiomes, using the Arctic Ocean as a case study.

Briefly, we start by reviewing the state of the art of the microbial rare biosphere subject (Part I). Then we focus specifically on marine microbial ecology methods, to produce high-quality data (Part II); then we focus on the problem of the definition of the rare biosphere, proposing our own methodology (Part III); and, finally, an in-depth analysis of the microbial rare biosphere ecology in the Arctic Ocean, using both traditional statistical approaches and machine learning (Part IV).

Part I includes a single chapter (Chapter 2) with a review of the rare biosphere, covering all levels of microbiology, *i.e.*, bacteria, archaea, fungi, protists, and viruses. In this review, we address the rare biosphere from an historical perspective, connecting the initial discovery of the microbial rare biosphere (Sogin *et al.*, 2006) with earlier studies in general ecology (Chapter 2). This historical perspective shows that the analysis of the microbial rare biosphere was dependent on Next Generation Sequencing (NGS) advances. Such methodological advances allowed for the exploration of the rare biosphere beyond prokaryotes (bacteria and archaea) and across a wide range of environments. In fact, the rare biosphere has been identified in multiple environments, including host-associated settings (Troussellier *et al.*, 2017), which exemplifies the importance of rarity in microbial ecosystems and its role in broader ecological and biogeochemical processes. Finally, in Chapter 2 we propose a conceptual combination of the theories used to explain the different ecological patterns identified in such a wide range of studies, following community assembly theory (Jia, Dini-Andreote and Falcão Salles, 2018). These insights pave the way for understanding microbial contributions to ecosystem stability and resilience. Overall, Part I provides a consistent framework for exploration of the rare biosphere in the following parts. In Part II, we focus on methods for assessment of marine microbial communities. Specifically, in seawater sampling of prokaryotes and protists (Chapter 3); and sequencing strategies and bioinformatics processing for prokaryotic data (Chapters 3 and 4). Therefore, we narrow the biological scope from microbes in general (Chapter 2), to prokaryotes and protists (Chapter 3), to just prokaryotes (Chapter 4 and 5). Such narrowing of biological scope was necessary,

because the more specific the methodological analysis, the more specific the target. Accordingly, seawater filtering methods can be assessed equally for prokaryotes and protists (Chapter 3), but that is not the case when we focus on methodological aspects of 16S rRNA gene amplicon sequencing (Chapters 3 and 4), which specifically target prokaryotes.

In Part II, the scope narrows down from the general microbial community towards the rare biosphere: Chapter 3 focuses only on the general community; Chapter 4 focuses on both the general community and the rare biosphere; and Chapter 5 is targeted at the rare biosphere. Again, the narrowness of the scope follows that of the underlying methodologies investigated and scope of each chapter.

In Part III, we focus on the proposal of a new perspective for the way the rare biosphere is defined (Chapter 6), followed by the creation of a new R package, *ulrb* (Unsupervised Learning based Definition of the Rare Biosphere) (Chapters 6 and 7). Essentially, we propose the utilization of an unsupervised machine learning approach to define the rare biosphere, a method that we show to work for microbial derived data and non-microbial data. Even though the thesis, at this point, focused primarily on prokaryotes, the tests applied to our new method clearly showed potential applicability to non-microbial data.

Finally, in Part IV, composed of a single chapter (Chapter 8), we advance the knowledge of the prokaryotic rare biosphere in the context of climate change in the Arctic Ocean, as a case study. To do so, all the previous chapters contribute to setting the theoretical framework (Chapter 2), producing a high-quality multi-year dataset (Chapters 3, 4 and 5), and analyzing the rare biosphere (Chapters 6 and 7). Additionally, Chapter 8 provides methodological advances in the analysis of microbial data in general, and particularly the prokaryotic rare biosphere, by contrasting traditional microbial ecology analysis tools against rule-based machine learning.

A high-level summary of the scientific connections between the chapters of this thesis, outlining their interrelationships and the overall outcomes, is presented in Figure 1.1. Below, we provide a general introduction to the central topics addressed in this thesis, providing a more detailed description and justification of the steps undertaken.

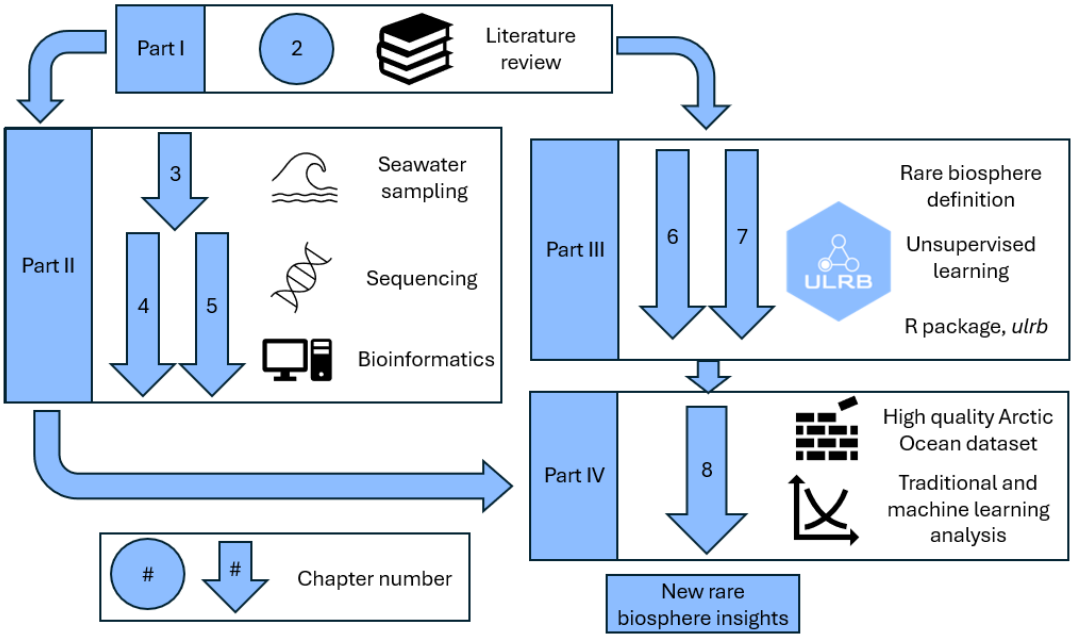


Figure 1.1. Schematic representation of the connection between the chapters of this thesis and their outcomes.

1.2. What is the microbial rare biosphere and why it is relevant

Microbial communities are composed of a few abundant taxa and many low abundance taxa, collectively referred to as the 'rare biosphere' (Sogin *et al.*, 2006; Pedrós-Alió, 2012; Lynch and Neufeld, 2015). The microbial rare biosphere represents a seed bank of high genetic diversity (Pedrós-Alió, 2012; Lynch and Neufeld, 2015), capable of providing responses to environmental stressors (Jousset *et al.*, 2017). Therefore, the genetic diversity of the rare biosphere makes it relevant for both the resistance and resilience of ecosystems (Jousset *et al.*, 2017), but also for bioprospecting novel natural products (Pascoal, Magalhães and Costa, 2020). Additionally, some rare taxa behave as 'conditionally rare taxa', which means that they can persist as dormant taxa in unfavorable conditions but grow and become abundant in more favorable conditions (Jones and Lennon, 2010; Shade *et al.*, 2014; Lynch and Neufeld, 2015). Such phenomena might explain, for example, how the global sponge microbiome shares so many symbionts, *i.e.*, some symbionts persist as rare, but viable, in the ocean, until they reach the host, where they thrive (Taylor *et al.*, 2013; Webster and Thomas, 2016). Not all rare taxa persist in a dormant state, some of them might be facing local extinction, might be disproportionally active (Pester *et al.*, 2010), or might be actively growing, but controlled by viral lysis (killing-the-winner theory (Thingstad, 2000)). In Chapter 2 we look

deeper into these ecological mechanisms and conceptually integrate them into a single framework.

1.3. Methodological challenges in the study of marine microbial ecology

In the context of climate change and environmental monitoring, one of the main challenges to study marine microbial ecology, relies on the creation of long-term data that is reliable, high quality and comparable across different years. For that to be possible, standardization is key. Several studies have launched sampling campaigns with standardized methodologies (Ocean Sampling Day, Tara Oceans, Malaspina) (Duarte, 2015; Kopf *et al.*, 2015; Pesant *et al.*, 2015; Sunagawa *et al.*, 2020), but such large-scale studies differ methodologically between each other. More specifically, the volume of seawater that is filtered varies a lot between studies; the filters used are composed of different materials; and some sampling campaigns divide the microbial component into different size fractions (Chapter 3). Seawater filtering and DNA extraction protocols can be used for either Metagenomic or Metabarcoding strategies. Metagenomic strategies consist of whole genome shotgun sequencing, whereby all the DNA is sequenced, providing both functional and taxonomic information; Metabarcoding, on the other hand, consists of amplicon sequencing of a target region. Usually, the target for metabarcoding is the 16S rRNA gene, for prokaryotes, or the 18S rRNA gene for unicellular eukaryotes. In Chapter 3, we used a Mediterranean Sea Metabarcoding and Metagenomic dataset, derived from seawater samples collected at the same site and time, to compare different seawater filtering protocols.

The comparisons made in Chapter 3 allowed us to understand to what extent data produced by protocols with slight variations are comparable. This was a key step in generating a multi-year dataset for Chapter 8.

To identify prokaryotes, the most common approach is to use Metabarcoding. Most studies have been using a short region of the 16S rRNA gene for that purpose, like the V4V5 region (Caporaso et al., 2011; 2012; Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016). The main advantage of this approach is the high throughput obtained. However, recent advances in sequencing techniques have allowed for high-throughput sequencing of larger sequences, enabling researchers to sequence the full 16S rRNA gene with remarkable accuracy (Callahan *et al.*, 2019; Tedersoo *et al.*, 2021). Therefore, we used replicated samples of an Arctic Ocean dataset, to directly compare how the two

Metabarcoding approaches performed for the same ecological samples (Chapter 4). Additionally, we also compared the most used taxonomy database, Silva (Pruesse *et al.*, 2007; Yilmaz *et al.*, 2014), against the Genome Taxonomy Database (GTDB) (Chaumeil *et al.*, 2020, 2022; Parks *et al.*, 2022), which is more recent. In Chapter 4 we cover the differences between databases in more detail.

Another factor to account for is the bioinformatics processing of raw sequences into comparable taxonomic units, or proxies of species. Initially, studies used OTUs, often with a 97% similarity threshold. However, recent studies have shown that it is more accurate to use ASVs, which also perform better for comparability across studies (Callahan *et al.*, 2016, 2019; Callahan, McMurdie and Holmes, 2017). In Chapter 5, we compared how OTUs and ASVs performed in the analysis of the prokaryotic rare biosphere. To do so, we used an Arctic Ocean dataset, consisting of Metabarcoding and Metagenomic data.

Besides Metabarcoding, Metagenomics is also a valid approach to survey microbial communities (both prokaryotes and eukaryotes), which is why we added Metagenomic data to Chapters 3 and 5.

Collectively, Chapters 3, 4 and 5 covered most aspects related to the methodological acquisition of a microbial community abundance table (information on taxa and their abundance in each sample) from seawater.

1.4. Challenges in the definition of the microbial rare biosphere

The microbial rare biosphere is generally defined as the group of low abundance taxa of a microbial community. Additionally, some studies might also consider rarity from the perspective of the frequency of occurrence, *i.e.,* how many times a given taxon is found in a dataset (Logares, Mangot and Massana, 2015), but microbial ecology studies usually refer to rarity as low abundance and account for frequency of occurrence separately. However, there is no agreement on what constitutes ‘low abundance’. Early studies on the microbial rare biosphere used to make a rank abundance curve, whereby all taxa are ordered from most to least abundant, and define the rare biosphere was the group of taxa in the so-called “long-tail” of the rank abundance curve (Sogin *et al.*, 2006). By doing so with data from 16S rRNA gene amplicon sequencing, early reviews suggested a threshold of 0.1% (Pedrós-Alió, 2012), which resulted in the widespread utilization of a

threshold of 0.1% relative abundance, per sample, in the literature (Chapter 2, 4, and 5). The main problem with using this threshold by default, or any other, is that it is arbitrary and provides no biological reasoning, and it is not comparable across studies that have used different methodologies. Previous studies have identified this problem and suggested the implementation of calculation-based methods to define the rare biosphere more objectively (Jia, Dini-Andreote and Falcão Salles, 2018). For example, it was proposed that, by adapting the Multivariate Cutoff Level Analysis (MultiCoLA) (Gobet, Quince and Ramette, 2010), it could be possible to delineate rare from abundant taxa in a biologically meaningful way (Jia, Dini-Andreote and Falcão Salles, 2018). Some studies apply their own strategies to define the rare biosphere (*e.g.*, (Espinosa-Asuar *et al.*, 2022)). Additionally, some studies do not explain clearly how they defined the rare biosphere (*e.g.*, (Rabelo-Ruiz *et al.*, 2021)). Therefore, there is a lack of a method that is easily applicable to different contexts, that facilitates inter-comparability of results, and that is ecologically meaningful.

In Chapter 5, we compared two different approaches to define the rare biosphere (MultiCoLA and a preselected threshold), highlighting the inconsistencies in available methods. To solve this problem, we proposed the utilization of unsupervised machine learning to cluster taxa based on their abundance within a given sample (Chapters 6 and 7), instead of pre-selecting any threshold of abundance. We termed this approach as *ulrb* and created an R package for its implementation. By using our approach, it is possible to automatically classify all taxa of a microbial community in a biologically meaningful way, without human bias.

1.5. The Arctic Ocean as a case study

To obtain relevant insights into the rare biosphere, we looked specifically at the Arctic Ocean. Climate change is changing the Arctic Ocean faster than the rest of the globe, due to Arctic amplification processes (Box *et al.*, 2019; Rantanen *et al.*, 2022). In fact, sea ice loss contributes to a positive feedback loop, whereby less ice results in less reflection of sunlight, which in turn results in higher heat absorption, further increasing ice loss (Box *et al.*, 2019; Rantanen *et al.*, 2022). Additionally, multiple climate phenomena result in a general increased rate of climate warming in this region. Furthermore, the loss of ice in the Arctic region is changing the landscape, which has implications for the ecosystem and also for humanity.

Environmental change in the Arctic Ocean also has a negative impact at the microbiological level, which is responsible for nutrient cycling and key for many host-

associations. Most studies on the ecology of prokaryotes in the Arctic Ocean focus mainly on the most abundant taxa, however, the rare biosphere holds the genetic potential to respond to environmental change. Therefore, an in-depth study of the rare biosphere in the context of the whole prokaryotic community is highly relevant.

Besides its relevance in the current climate change scenario, the Arctic Ocean is an interesting environment to test methodological questions, because it is an understudied region of the globe, with relatively less available information than temperate regions. Considering the underrepresentation of DNA sequences originating from the Arctic Ocean (Priest *et al.*, 2021), taxonomy database selection is a key step of microbial communities from such regions (Chapter 4).

We produced a multi-year microbiome dataset of high-quality in Chapter 8, taking advantage of the results and conclusions from Chapters 3, 4, and 5. After producing a large dataset (n > 100 samples) of Metabarcoding data, with environmental variables, we performed traditional microbial ecology analyses. Namely, we described the major taxa that were present, as well as alpha and beta diversity (sample-level and community composition, respectively). Furthermore, we used linear methods to compare alpha diversity against key environmental variables (temperature, salinity, nitrate, nitrite, ammonium, silicates, phosphate, chlorophyl *a*, depth and water mass). An overview of such environmental variables showed a remarkable impact of temperature on the diversity of the rare biosphere within the Arctic Ocean, whereby the cooler water masses were the most diverse, thanks to the rare biosphere (Chapter 8). This was supported by the initial rare biosphere analyses performed in Chapter 5, which were made using a smaller dataset, from samples collected at different sites.

One drawback of the traditional microbial ecology analysis framework is the difficulty in obtaining insights from complex interactions of variables. Therefore, to identify non-linear and complex patterns from the final dataset, we used rule-based machine learning in Chapter 8. Such an approach provides models that are human readable and informative, but with so-far limited utilization in microbial ecology studies (Giulia *et al*., 2021; Pascoal, *et al.*, 2022b). More specifically, we take the classification obtained by *ulrb* and try to explain them as a function of complex combinations of variables (Chapter 8).

1.6. Objectives

The main objective of this thesis is to advance the study of the microbial rare biosphere, with a particular focus on marine environments and the Arctic Ocean.

We start by covering the state of the art of the rare biosphere in all areas of microbial ecology (Part I), with the objective of understanding the current ecological concepts used to explain the rare biosphere, the different environmental contexts where the rare biosphere is playing important roles, and the methodological approaches being employed.

Then, the thesis focuses on methodological aspects of marine microbial ecology (Part II), with the objectives of: (1) understanding to what extent different marine microbiome sampling protocols are comparable, for the purpose of creating multi-year standardized microbiome datasets (Chapter 3); (2) compare advantages and disadvantages of using V4V5 or full-length 16S rRNA gene amplicon sequencing, in combination with different general purpose taxonomy databases (Silva and GTDB) to describe the microbial rare biosphere (Chapter 4); (3) compare the utilization of different bioinformatics strategies to process raw reads into phylogenetic units (Metabarcoding and Metagenomics), and compare different approaches to define the rare biosphere (Chapter 5).

The next objective was to propose a new method to define the microbial rare biosphere, including the construction of a new R package and a website with learning materials, to employ our proposed strategy (Part III). This part aimed to: (1) demonstrate that unsupervised machine learning can solve the problems with threshold-based definitions of the rare biosphere (Chapter 6) and build an open-source R package for the same purpose (Chapters 6 and 7).

Once the methodological objectives are fulfilled, the next main objectives are to produce a high-quality long term microbiome dataset of the Arctic Ocean to access the microbial rare biosphere in a climate change environmental context, and apply our own methodology to classify rare, undetermined, and abundant taxa (Chapter 8). Besides the identification of the rare biosphere, another important objective is to extract as much information as possible by using both a classical statistics approach and rule-based machine learning (Chapter 8). Rule-based machine learning has the potential to provide human interpretable models, and the ability to extract complex patterns from big datasets. Finally, such methodological advances both in the creation of standardized datasets and data analysis, contribute to the overarching goal of understanding the role of the rare biosphere in the environment in general, but more specifically its role in the Arctic Ocean. In this regard, this thesis contributes to closing the gap in the knowledge of the prokaryotic rare biosphere in the Arctic Ocean under climate change.

Below, we provide a summary of the specific objectives of each Part of the thesis, and outline the steps undertaken in each Chapter:

Part I To establish a theoretical framework for the thesis.

Chapter 2 Review literature on the rare biosphere topic, focusing on:

a) Historical perspective.
b) Major findings relative to prokaryotic and protists rare biosphere.
c) The impact of viruses on the rare biosphere, and the existence of rare viruses in the environment.
d) The role of the microbial rare biosphere in host associations.
e) Survey of methodologies used to study the microbial rare biosphere.
f) Analysis and organization of ecological concepts used to explain the rare biosphere, and ecological concepts that account for the rare biosphere impact.

Part II To compare methodological aspects of marine microbiome data that are relevant for the construction of high-quality, standardized datasets.

Chapter 3 Compare the impact of different seawater sampling protocols on the resulting prokaryotic and unicellular eukaryotic diversity estimates. More specifically,

a) Compare the following seawater sampling approaches:
   (1) Sampling volume (from as low as 1 L to up to 1000 L).
   (2) Filter whole water (0.22 μm) *vs* fractionated filtration in consecutive filters (0.22 μm, 3 μm, 0.22 μm).
   (3) Type of filter (cartridge filter or flat membrane).
b) The methods in (a) were compared in parallel for the following sequencing strategies:
   (1) V4V5 16S rRNA gene amplicon sequencing (target prokaryotes).
   (2) V8V9 18S rRNA gene amplicon sequencing (target eukaryotes).
   (3) Shotgun metagenomic sequencing (target prokaryotes and eukaryotes).
c) The impact of the methodological combinations between sampling protocols (a) and sequencing strategies (b) were evaluated in terms of:
   (1) Alpha diversity (sample-level diversity).
   (2) Beta diversity (community composition across samples).
   (3) Taxonomic composition.

Chapter 4 Compare the impact of using different 16S rRNA gene amplicon sequencing strategies in the context of the Arctic Ocean.

a) Compare the following sequencing strategies:
   (1) V4V5 16S rRNA gene amplicon sequencing (target prokaryotes).
   (2) Full-length 16S rRNA gene amplicon sequencing (target prokaryotes).
b) For each sequencing strategy in (a), the following databases were employed:
   (1) Silva database.
   (2) GTDB database.
c) The methodological combinations of (a) and (b) were evaluated in terms of:
   (1) Taxonomic classification efficiency.
   (2) Alpha diversity.
   (3) Beta diversity.
   (4) Species level taxonomic classification.
d) The impact of all the above combinations was tested with or without rarefaction.

Chapter 5 Compare the impact of different bioinformatics approaches to the study of the rare biosphere.

a) Utilization of different phylogenetic units:
   (1) OTUs, processed by the MGnify platform.
   (2) ASVs, processed with *DADA2* R package.
b) Applied to two different sequencing strategies:
   (1) V4V5 16S rRNA gene amplicon sequencing (for ASVs and OTUs)
   (2) Shotgun metagenomic sequencing (for metagenomic OTUs, mOTUs)
c) To identify the rare biosphere with the methodological combinations of (a) and (b), we used the following approaches:
   (1) 0.1% relative abundance, per sample, threshold.
   (2) MultiCoLA (Gobet, Quince and Ramette, 2010).
d) The impact of all the above combinations was tested with or without rarefaction.
e) To interpret the rare biosphere ecology in the Arctic Ocean, all the above variations were further used to estimate different types of rarities (*e.g.*, some taxa might be conditionally rare).

Part III Develop a methodology to delineate the rare biosphere.

Chapter 6 Present and test the method *ulrb.*

a) Verify *ulrb* applicability for microbial communities described by different bioinformatics approaches:
   (1) ASVs from V4V5 16S rRNA gene amplicon sequencing.
   (2) OTUs from V4V5 16S rRNA gene amplicon sequencing.

(3) mOTUs from shotgun metagenomic sequencing.

b) Verify *ulrb* applicability for microbial communities described by different sequencing strategies:
   (1) ASVs from V4V5 16S rRNA gene amplicon sequencing.
   (2) ASVs from full-length16S rRNA gene amplicon sequencing.

c) Verify *ulrb* statistical robustness against dataset size, in terms of:
   (1) Number of samples.
   (2) Sequencing depth.
   (3) Number of taxa.

d) Verify applicability of *ulrb* for non-microbial datasets, by applying *ulrb* to:
   (1) Dataset of ant species counts.
   (2) Dataset of tree counts.

e) Illustrate the identification of different types of rarity by applying *ulrb* to a coral microbiome, by:
   (1) Monitoring of target taxa, based on the coral host-association literature, and manually attribute types of rarity.

Chapter 7 Explain in detail the *ulrb* R package.

a) Explain the problem of the definition of the rare biosphere in machine learning terms.

b) Explain the architecture of *ulrb*, as well as all functions, their arguments and interactions.

c) Present illustrative examples of all the capabilities of *ulrb*, including R code.

Part IV Use a multi-year standardized microbiome dataset of the Arctic Ocean to explore in depth the microbial rare biosphere in its environmental context.

Chapter 8 Arctic Ocean case study objectives:

a) To provide a general description of the prokaryotic communities overall, from 2016 to 2020, across the epipelagic, mesopelagic and bathypelagic layer, including:
   (1) Diversity estimates (alpha and beta diversity).
   (2) Taxonomic composition of major phyla and class.

b) Exploratory analysis of diversity as a function of environmental variables:
   (1) Depth (m), temperature (ºC), PO_4^- (µM), NO_2^- (µM), NO_3^- (µM), Si (µM), NH_4^+ (µM), Chlorophyll *a* (Chl *a*) (mg/L), salinity (Practical Salinity Units - PSU), water mass, region (open ocean or fjord).

c) Support major patterns found in (b) with different statistics:
   (1) Linear Least Squares Model.

(2) Poisson Model.
(3) Pearson and Kendall correlations.

d) Apply rule-based machine learning as an alternative to the previous steps. This step is further divided into several objectives:
   (1) Create a transaction database from the full dataset, thus combining metadata, abundance data, and taxonomy data in a single database.
   (2) Apply Association Rule Mining (ARM) to transactions database, to obtain a list of association rules.
   (3) Develop a methodology to reduce noise and extract informative and non-redundant association rules.
   (4) Provide an ecological interpretation of the most important association rules identified.
   (5) Identify advantages and disadvantages of using ARM.
   (6) Compare the taxonomic groups highlighted by the rule-based machine learning approach, against the taxonomic groups usually addressed in the literature.

e) Explain the multi-year microbial ecology of the Arctic Ocean, accounting for all levels of abundance (rare, undetermined and abundant), considering the multiple analysis employed.

f) Assess whether the approach proposed here to define the rare biosphere, *ulrb*, was useful in the interpretation of the microbial ecology in the Arctic Ocean case study (as an example of a complex scenario, with multiple variables).

1.7. General experimental design, statistical approach and data

To fulfill the objectives described previously, we generally used publicly available data but also published new datasets. For the selection of datasets, priority was given to data produced by the research groups already involved with the PhD project, or datasets previously described in peer-reviewed scientific publications.

In general, most of the objectives in this thesis are dependent on the comparison of different methods. To identify differences in diversity results that are potentially derived from methodological differences, we generally relied on replicate samples. Whereby the same environmental site was assessed multiple times by different methods. For

methodological comparisons, we described the microbial communities using commonly used strategies and then compared the output. In other words, we verified if, by changing methodology, the same environmental sample would result in a different interpretation of the diversity results. Whenever statistical tests were necessary, we used common statistical tests, generally following this reasoning:

- To compare one numerical response variable between two independent groups:
  - Wilcoxon (non-parametric)[†].
  - t-student (parametric).
- To compare one numerical response variable between three, or more, independent groups:
  - Kruskall-Wallis (non-parametric).
  - One-way Analysis of Variance (ANOVA) (parametric).
- To compare one numerical response variable between multiple independent groups, accounting for a grouping variable:
  - Two-way ANOVA (parametric).
- To verify parametric requirements, we generally performed the following steps:
  - Verify the existence of extreme outliers.
  - Verify homogeneity of variance (using Levene test).
  - Verify normal distribution (using Shapiro test).

Notes relative to testing parametric assumptions: in real world datasets, parametric assumptions can be difficult to obtain, because the datasets tend to have many variables for relatively few samples (“high dimensionality” problem). Therefore, such assumptions should be considered and interpreted in a case-by-case scenario. However, in general, we did not apply data transformation to fit the data within the assumptions, instead opting for non-parametric tests.

- For numerical response variables represented as a function of another numerical variable:
  - Regression line, using linear least squares model.

Notes on regression lines: we used the linear least squares model for the general purpose of providing descriptive statistics, but did not intend to exhaustively curate a statistical model until the best fit was obtained. Instead, the regression line was used in its simplest form, because it is informative and easy to interpret.

[†] This test is also known as Mann-Whiney U test.

- For beta diversity:
    - Permutational Multivariate Analysis of Variance (PERMANOVA), usually as supporting statistics of multivariate analysis.

Some chapters used additional statistical tests, as needed. For instance, we are not summarizing here the statistical metrics specific to unsupervised and rule-based machine learning, as they are exhaustively explained in their respective chapters (Chapters 6, 7, and 8).

In terms of ecological analysis, we generally followed the classical approach used to describe microbial communities. We consider this approach to be quite useful in providing the reader with a comprehensive overview of the general characteristics of microbial communities. This general description is then used as a framework on top of which we can elaborate using more advanced machine learning methods.

In summary, our description of microbial communities generally consisted of:

- Calculating alpha diversity metrics:
    - "Species richness" – the number of species. Note that this traditional metric can be confusing in microbial ecology, because we often cannot classify taxa down to species level. Thus, this metric is sometimes named according to the specific methodology used, for example, ASV richness. In some cases, for simplicity, we use "number of ASV/OTU/taxa".
    - Alternative metrics, such as Shannon or Simpsons index.

Notes on alternative alpha diversity metrics: we generally focused on the number of taxa, because if we account for the rare biosphere, we are already incorporating into the diversity index the factor of abundance. Thus, adding metrics that simply introduce a factor of abundance in their calculations is redundant. However, in some cases, especially when focusing on general patterns (not accounting for the rare biosphere specifically), we also included alternative metrics (usually, the Shannon index). Another disadvantage to be considered when using multiple metrics of alpha diversity is the rapid increase in the number of results to interpret. For example, when we are comparing multiple combinations of methodologies, if we also account for an alternative alpha diversity metric, we effectively double the results, which often are relegated to supplementary materials. Finally, if we consider the additional effort for, oftentimes, limited differences in the result, adding several alpha diversity metrics can become counterproductive. In fact, in Chapter 3 we verified that distinct alpha diversity metrics

are very correlated with each other, meaning that they often do not provide new information, as was shown by (Swenson et al., 2014).

- Estimate beta diversity in terms of community composition, using multivariate level analysis, usually including:
    - Correspondence Analysis (CA) and/or non-metric Multidimensional Scaling (nMDS) – unconstrained.
    - Redundancy Analysis (RDA) and/or Canonical Correspondence Analysis (CCA) – constrained.
    - Others, as needed.
- Describe major taxonomic groups
    - Usually focus on phyla and class level.
    - For some specific objectives, we investigated lower taxonomic levels (*e.g.*, we examined species level in Chapter 4).

Regarding data visualization, we generally considered the following:

- For representation of numerical results, across independent groups:
    - For less than 5 samples, we preferred to represent the data in points.
    - Between 5 and 30 samples, we preferred to represent the data by boxplots.
    - For more than 30 samples, we preferred to represent the data by mean and standard deviations.
    - Notes:
        - In several situations, we considered it best to include both boxplots and the actual points, to aid the reader.
        - In several situations, it was not possible to follow these rules strictly, generally because of unequal sampling across environmental variables.

For representation of qualitative groups, we followed general data visualization good practices (Wilk, 2019).

For descriptive statistics, we usually relied on centrality metrics. Generally, when describing a small sample size (about 30 samples), we preferred to use the median value, followed by the interquartile range (IQR) and/or minimum and maximum values; when describing a larger sample size (more than 30 samples), we generally relied on mean and standard deviation. However, in several situations we considered that it would be more informative to include both.

Regarding taxonomic nomenclature, we generally attempted to follow the International Code of Nomenclature of Prokaryotes (ICSP) (Oren *et al.*, 2023). However, there were significant modifications of the nomenclature rules during the PhD period. Therefore, some of the chapters in this thesis were published before several nomenclature updates, while other chapters used databases that were not updated to reflect such changes at time of writing. For consistency, we adopted the practice of using the taxonomy nomenclature as provided by the taxonomy databases used. One of the major discrepancies between the new nomenclature of prokaryotic names and our results regards major phyla. For example, *Proteobacteria* instead of *Pseudomonadota* (Oren and Garrity, 2021). Note that the option to keep the original names given by the taxonomy databases is purely pragmatic and avoids the introduction of human error in manual updates of nomenclature but does not constitute any argument in favor or against the nomenclature updates enacted by the ICSP. Additionally, for consistency with the ICSP we applied italic text to prokaryotic nomenclature from phylum to species level, except for *Candidatus* taxa. Note, however, that in some figures and tables it was not possible to apply italic text to all taxonomic levels.

The work carried out in this thesis never included field work (sample collection), nor laboratory work (DNA extraction, quantification, or sequencing). Generally, the starting point of each experiment was the raw sequencing results, or the final abundance table, if previously published. Thus, the work of this thesis focused primarily on the bioinformatics processing of raw reads, and the data analysis of the abundance tables obtained after the bioinformatics processing. Below, we summarize the datasets and their original purpose:

- EuroMarine Open Science Exploration (EMOSE - 2017) (Pascoal *et al.*, 2023).
    - Original purpose: compare multiple seawater sampling protocols for the description of microbial communities.
    - Data available at: ENA (ERP090011), MGnify (MGYS00001935) and PANGAEA (https://doi.pangaea.de/10.1594/PANGAEA.879516).
- Norwegian Young Sea-Ice Expedition (N-ICE - 2015) (Granskog *et al.*, 2016; de Sousa *et al.*, 2019; Pascoal *et al.*, 2022a).
    - Original purpose: describe microbial communities in the Arctic Ocean (North of Svalbard) during a winter to summer transition, using a vessel fixed on drifting ice.
    - Data available at: ENA (PRJEB21950, PRJEB15043) and MGnify (MGYS00001922, MGYS00001869).

- Environmental Monitoring of Svalbard and Jan Mayen (MOSJ - 2016-2020) (Chapter 6) (Pascoal *et al.*, 2025a).
    - Original purpose: monitor microbial communities yearly, at fixed stations.
    - Data available at: MGnify (PRJEB24517, PRJEB72025, PRJEB72030, PRJEB60815, PRJEB72034).
- Coral microbiome (Keller-Costa *et al.*, 2021).
    - Original purpose: study host-associations between corals and their microbiomes.
    - Data available at: ENA (ERP014771) and MGnify (PRJEB13222).
- Other.
    - Non-microbial datasets:
        - Ants.
            - Data available at: (Arnan, Gaucherel and Andersen, 2011; Calatayud *et al.*, 2019; Balbuena *et al.*, 2021).
        - Barro Colorado Island tree counts (BCI).
            - Data available at: (Condit *et al.*, 2002, 2019).

Note that most datasets were named after the project or sampling campaigns they were associated with. In Table 1.1 we summarize the utilization of the datasets through the thesis, their objective for this thesis and original references.

Table 1.1. Summary of datasets used through this thesis and the main objective within each chapter. Some additional contextual information is given, including region and year. The number of samples presented here can change in the final ecological analyses, due to quality filtering or missing data. Some chapters included the bioinformatic raw read processing of some datasets, but in some cases some of the data was already processed (Mixed).

Dataset	Years	Part	Chapter	Objective	Region	Raw read processing	Total samples	Notes	Original reference
EMOSE	2017	II	3	Compare seawater sampling protocols	Mediterranean Sea	No	157		(Pascoal *et al.*, 2023) (Chapter 3)
MOSJ	2019	II	4	Compare V4V5 and full-length 16S rRNA gene amplicon sequencing Compare utilization of Silva and GTDB database	Arctic Ocean	Yes	36	Subset of MOSJ 2016-2020	(Pascoal *et al.*, 2024) (Chapter 4)
		III	6	Test *ulrb*					
N-ICE	2015	II	5	Compare V4V5 and shotgun metagenomic sequencing Compare ASVs, OTUs, and mOTUs Compare definitions of the rare biosphere Assess types of rarity	Arctic Ocean	Mixed	18		(Granskog *et al.*, 2018; de Sousa *et al.*, 2019)
		III	6	Test *ulrb*					
			7	Illustrate *ulrb* examples		Yes	9		
MOSJ	2016-2020	III	6	Test *ulrb*	Arctic Ocean	Yes	117		(Pascoal *et al.*, 2025a)(Chapter 6)
		IV	8	Multi-year analysis of prokaryotic communities in the Arctic Ocean					

Coral Microbiome	2021	III	6	Test *ulrb*	Atlantic Ocean	No	20		(Keller-Costa *et al.*, 2021)
Ants	2011	III	6	Test *ulrb*	Australian tropical savanna	No	99		(Arnan, Gaucherel and Andersen, 2011; Calatayud *et al.*, 2019; Balbuena *et al.*, 2021)
BCI	1985-2015	III	6	Test *ulrb*	Barro Colorado Island	No	18	Subset of long-term census data.	(Condit *et al.*, 2002, 2019)

End of Table 1.1.

Part I. State of the art.

Chapter 2. The microbial rare biosphere: current concepts, methods and ecological principles.

The materials from this chapter were previously published in:

Pascoal, F., Costa, R. and Magalhães, C. (2021) “The microbial rare biosphere: current concepts, methods and ecological principles,” *FEMS Microbiology Ecology*, 97(1). Available at: https://doi.org/10.1093/femsec/fiaa227.

See author contributions in section **2.9. Declarations.**

2. The microbial rare biosphere: current concepts, methods and ecological principles

Francisco Pascoal[1,2], Rodrigo Costa[3,4] and Catarina Magalhães[1,2]

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal; [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal; [3]Department of Bioengineering, Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico, University of Lisbon, Portugal; [4]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, University of Lisbon, Portugal.

Abstract

Our ability to describe the highly diverse pool of low abundance populations present in natural microbial communities is increasing at an unprecedented pace. Yet we currently lack an integrative view of the key taxa, functions and metabolic activity which make up this communal pool, usually referred to as the 'rare biosphere', across the domains of life. In this context, this review examines the microbial rare biosphere in its broader sense, providing an historical perspective on representative studies which enabled to bridge the concept from macroecology to microbial ecology. It then addresses our current knowledge of the prokaryotic rare biosphere, and covers emerging insights into the ecology, taxonomy and evolution of low abundance microeukaryotic, viral and host-associated communities. We also review recent methodological advances and provide a synthetic overview on how the rare biosphere fits into different conceptual models used to explain microbial community assembly mechanisms, composition and function.

Keywords: history of biological rarity; community assembly; rare biosphere functions; rare microeukaryotes, rare viruses; host-microbe interactions

2.1. Introduction

The term ‘rare biosphere’ was firstly coined to refer to the highly diverse pool of low relative abundance microbial populations reported in samples from deep water masses in the North Atlantic (Sogin *et al.*, 2006). Despite their low relative abundance, recent research suggests that rare microbial taxa may be more relevant to ecosystem functioning than previously anticipated. First, because many rare taxa can grow abundant under changing conditions (Shade *et al.*, 2014) and respond to perturbations - for example, by degrading polycyclic aromatic hydrocarbons (Sauret *et al.*, 2014). Second, the high diversity of low abundance taxa represents a large reservoir of genetic traits underpinning a wide spectrum of both known and potentially novel microbial functions (Elshahed *et al.*, 2008; Lynch and Neufeld, 2015). Third, members of the rare biosphere can be disproportionately active relative to their abundance for key functions (Pester *et al.*, 2010). Recently, these properties of the rare biosphere have been addressed in the context of future biotechnological applications in bioremediation and bioprospecting, which can be developed through the exploitation of the metabolism of low abundance populations (Pascoal, Magalhães and Costa, 2020). Furthermore, knowledge of the microbial rare biosphere and its properties is relevant to approaching microbial ecology at several levels – from community assembly theory (Jia, Dini-Andreote and Falcão Salles, 2018) to host-microbiome associations (Dawson *et al.*, 2017; Troussellier *et al.*, 2017), competitive interactions (Kurm, van der Putten and Hol, 2019) and microbial responses to climate change (Fernández-Gómez *et al.*, 2019; Liang *et al.*, 2020), to name a few key and recent examples.

From a statistical point of view, the microbial rare biosphere can be represented by the “long-tail” of a rank abundance curve (Figure 2.1), where all taxa are ordered from the most abundant to the least abundant ones. Because of the subjectivity associated with the concept of “rarity”, there is currently no clear consensus nor proper methodology to determining a specific threshold from which populations should be considered low abundant or “rare” within any given community (Jia, Dini-Andreote and Falcão Salles, 2018), *i.e.*, the point where the ‘long-tail’ starts (see question mark in Figure 2.1). Although efforts have been made to solve the subjective nature of rarity in microbial communities (Gobet, Quince and Ramette, 2010; Jia, Dini-Andreote and Falcão Salles, 2018), still most studies use the arbitrary threshold of 0.1% or 0.01% relative abundance per sample to label rare microbial populations. Furthermore, distinct methodologies used in the description of microbial communities will present different levels of resolution to

describe the rare biosphere (see arrows in Figure 2.1) and that should be considered when choosing a rarity threshold.

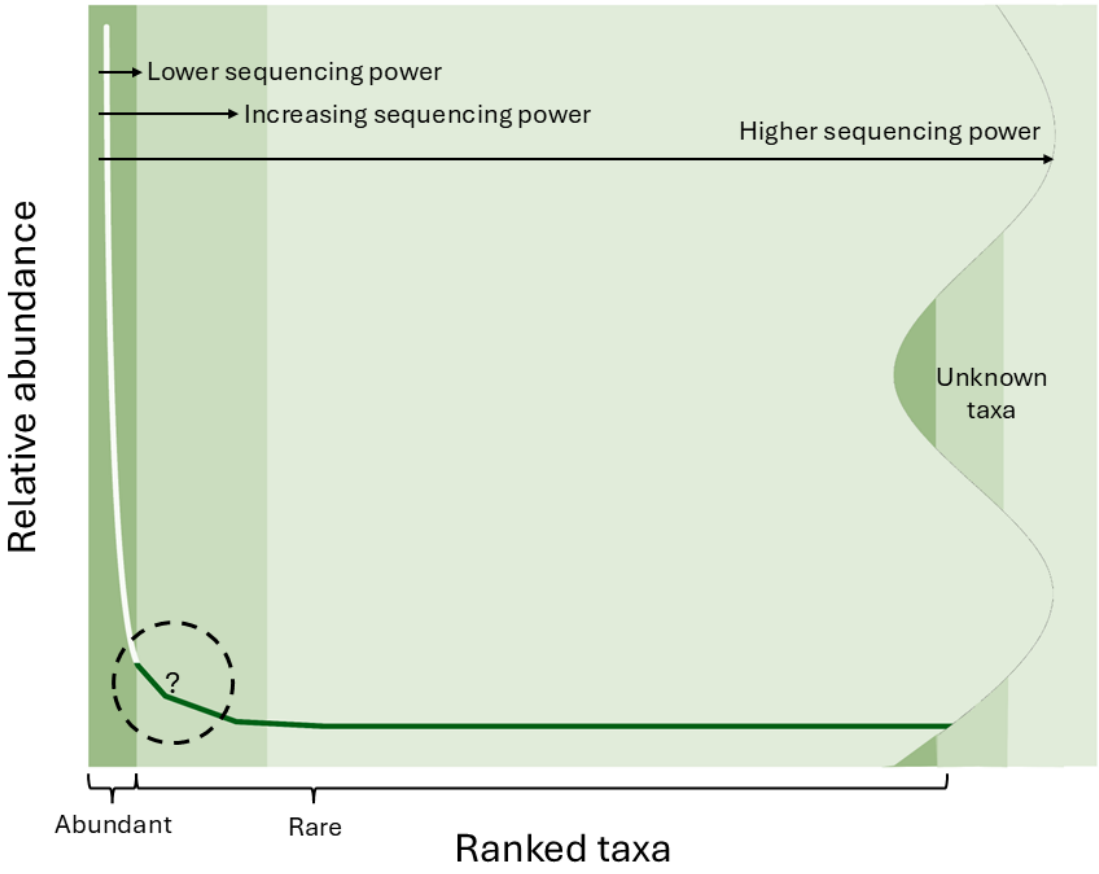


Figure 2.1. Conceptual illustration of a rank abundance curve. All taxa are ordered from the most abundant to the least abundant ones. The long-tail of the rank abundance curve illustrates the microbial rare biosphere, but the beginning of this long-tail is not obvious, as depicted by the question mark. The arrows and shades of green illustrate how different sequencing power influence the view of the rare biosphere and the waves in the end represent the taxa that are unknown.

Although next-generation sequencing methods leading to the delineation of the "microbial rare biosphere", as proposed by Sogin and colleagues in 2006, were not available to early microbial ecologists, the realization that rare populations are an integral part of biological communities is far from new (McGill *et al.*, 2007). During the last century, cumulative knowledge of biological rarity enabled key improvements in ecological theory and modelling (McGill *et al.*, 2007; Jia, Dini-Andreote and Falcão Salles, 2018). More recently, advances in molecular biology opened new ways of cataloguing microbial diversity with unprecedented speed, depth, and accuracy. Collectively, these advances now allow us to pursue greater understanding of microbial rarity in an integrative fashion as addressed below. Nevertheless, most of the studies approaching low abundance microbial populations focused mostly on the taxonomic

composition of the prokaryotic rare biosphere (Pedrós-Alió, 2012). Only more recently, the research expanded towards other microbial groups like microeukaryotes (Logares *et al.*, 2014) and viruses (Holmfeldt *et al.*, 2013), as well as host-microbiome associations (Webster *et al.*, 2009), while several methodological advances have allowed the exploration of metabolic traits within the rare biosphere (Hausmann *et al.*, 2019).

2.2. An historical perspective on the concept of biological rarity

Naturalists have reported on the existence of a few high abundance and many rare species in biological communities since the 19th century (McGill *et al.*, 2007). Perhaps the most iconic example stems from Darwin's seminal work on evolution by means of natural selection, where it is recognized that "rarity is the attribute of a vast number of species of all classes, in all countries" (Darwin, 1859). In *The Origin of Species*, the author also stresses the relevance of understanding the mechanisms of rarity and species distributions in biology:

> "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, 1859).

In Darwin's work the ubiquity of rarity was a trivial observation that laid important questions not fully answered up to today. An important distinction which needs to be made in any survey approaching rarity in biological communities is the one between local and geographical types of "rarity", since determined species may be locally very abundant but display low frequency of occurrence at regional or global geographical scales. Unless otherwise stated, in this review the terms "rare" and "rarity" are used to refer to local abundance, that is low abundance populations within a given microbial community or "microbiome". In the early 20th century, the field of biology expanded progressively to encompass, besides purely descriptive approaches to natural communities, quantitative, statistically driven analyses of species and their abundance distributions. This shift is relevant, because the concept of rarity implies a statement on abundance, one taxon is rare relative to another in the same observed universe (community). This can be described by species abundance distribution curves (Figure 2.2), which revealed uneven abundance distributions in early studies of plants and animals (Corbet, 1942; Fisher, Corbet and Williams, 1943; Preston, 1948). Interestingly, independently of the model used, the species abundance distribution curve is similar across studies and biological communities, previously reviewed (McGill *et al.*, 2007).

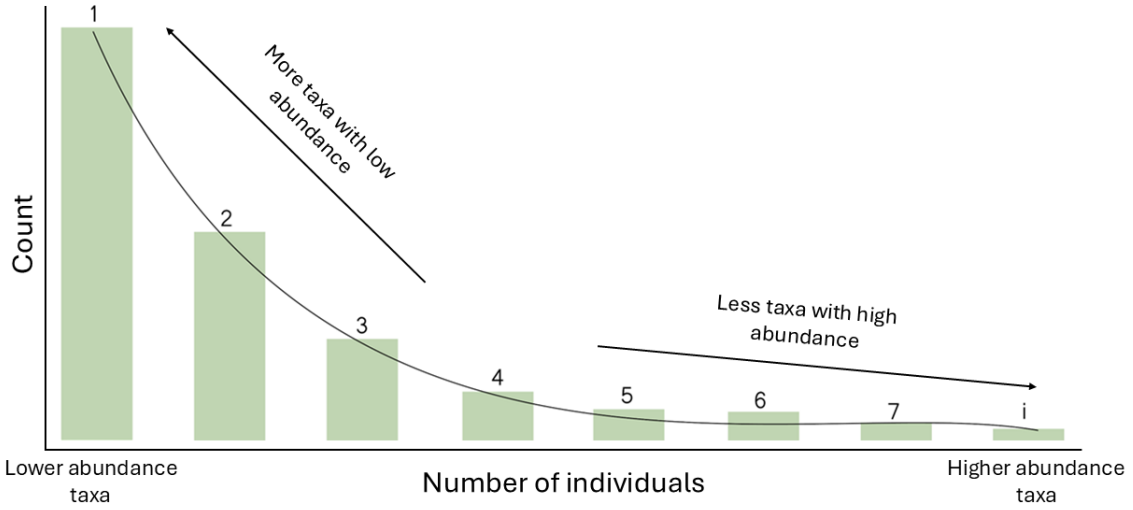


Figure 2.2. Representation of the species abundance distribution curve, which is one histogram where each bin represents the number of taxa with ni individuals. Most taxa have low abundance, while a few have higher abundance, as depicted by the arrows.

Species abundance distribution curves for microorganisms were only possible in the 21st century, due to advances in molecular biology, such as the use of 16S ribosomal RNA (rRNA) gene sequences to re-organize the tree of life (Woese and Fox, 1977). This allowed the development of microbial diversity surveys based on culture-independent methods, such as cloning and sequencing (Giovannoni *et al.*, 1990; Ward, Weller and Bateson, 1990) and community fingerprinting analysis like Denaturing Gradient Gel Eletrophoresis of rRNA genes (Muyzer, de Waal and Uitterlinden, 1993) or Automated Ribosomal Intergenic Spacer Analysis (Fisher and Triplett, 1999). Those methods were useful to describe abundant microbial taxa, but did not detect rare biosphere members (Pedrós-Alió, 2012).

Before the description of the microbial rare biosphere, ecological studies on animals and plants laid the foundations for some concepts used today in microbial ecology. For instance, the concept of abundance was divided in “global” and “local” abundance as well as “relative” and “observed” abundance (Preston, 1948), *e.g.*, the number of organisms from species A, collected in the sample analyzed, is the “observed” abundance, while the proportion of the number of organisms from species A relative to the total number of organisms from species A, B and C, collected in the same sample, is the “relative” abundance of species A.

Early definitions of rarity, in the context of plant communities, emerged from the perspective of habitat specificity, geography and local abundance (Rabinowitz, 1981). Importantly, the microbial rare biosphere concept is currently associated with relative

abundance, while geographic distribution and habitat specificity of microorganisms are indicators of frequency of occupancy, they can be abundant or rare but found in few samples (low occupancy/restricted) or be found in many samples (high occupancy/ubiquitous). The relative abundance concept of rarity was used, for example, to relate size with abundance in animals, whereby larger animals tend to be rarer (Lawton, 1989). In a study of the reproduction and dispersal of rare vs abundant populations, across datasets of different animal and plant species, several common patterns were found (Kunin and Gaston, 1993). In general, while dispersal was observed to be more limited, asexual reproduction strategies were considered an advantage for low abundance taxa. If these trends are to be mirrored by natural microbial communities, it could be predicted that rarity is not as dangerous for microorganisms as it is for animals and plants that are limited to sexual reproduction, because of the ability of clonal multiplication and independence from encounter rates exhibited by haploid microbial populations (Magurran and Henderson, 2003; Pedrós-Alió, 2006).

In summary, rare biosphere research has long been constrained by the available methods to characterize the diversity of different forms of life. Notwithstanding, several historical discoveries in general biology were crucial to improve our current view of the microbial rare biosphere, unlocked by today's advanced methodologies. In recent years, attention has been also given to the diversity and putative roles of low abundance microeukaryotes and viruses besides prokaryotes, and the next sections review such findings.

2.3. The prokaryotic and microeukaryotic rare biosphere

Most studies on the microbial rare biosphere have focused on prokaryotes and have shaped our current understanding of this important component of microbial diversity (Pedrós-Alió, 2012; Lynch and Neufeld, 2015; Jousset *et al.*, 2017).

The prokaryotic rare biosphere is thought to be composed of cells in different metabolic states, including dormant but viable cells (Jones and Lennon, 2010). This pool of low relative abundance microorganisms can eventually multiply under favorable conditions (Shade *et al.*, 2014), supporting early hypotheses which suggested the rare biosphere to function as a seed bank to leverage microbial community responses to environmental changes (Pedrós-Alió, 2007). The prokaryotic rare biosphere also comprises a highly diverse genetic pool that encompasses both novel and already described prokaryotes (Elshahed *et al.*, 2008; W. Zhang *et al.*, 2018). In addition, there are low abundance prokaryotes that, despite not becoming abundant, may display disproportionately high activity for specific functions (Pester *et al.*, 2010; Sauret *et al.*, 2014). However, the genetic mechanisms responsible for high activity with close-to-zero net growth have only recently been addressed (Hausmann *et al.*, 2019). Moreover, recent evidence suggests that rare prokaryotes may equip abundant or growing prokaryotes with novel functions by means of horizontal gene transfer (Wang *et al.*, 2017; Yang *et al.*, 2017). Finally, modern microbial biogeography studies suggest that the abundant populations have a more constant abundance and are composed of generalists, while the rare biosphere is more prone to respond to environmental fluctuations (Du *et al.*, 2020; Ji *et al.*, 2020). These examples highlight the fast-paced emergence of concepts and trends derived from the study of the prokaryotic rare biosphere across nature's ecosystems. An integrative understanding of the ecology and evolution of the rare biosphere requires extension and validation of such concepts to other key microbial groups.

Caron and Countway (2009) proposed that the ecological mechanisms underlying the dynamics of the rare microeukaryotic and prokaryotic biospheres are similar – they both represent a genomic pool that can work as seed bank and respond to changing conditions. This view was largely confirmed by early studies on the protistian rare biosphere, later reviewed by Logares, Mangot and Massana (2015). Surveys on the composition and structure of the microeukaryotic rare biosphere found specific biogeographic and temporal patterns (Nolte *et al.*, 2010; Logares *et al.*, 2014),

suggesting responsiveness to environmental factors and metabolic activity, based on the ratio of 18S rRNA genes to 18S rRNA transcripts (Logares *et al.*, 2014; Debroas, Hugoni and Domaizon, 2015). Recent efforts made to further understand protist biogeography at a global scale include the Tara Oceans expedition (Karsenti, 2015), where different models (*e.g.*, power law) fitted well with the species abundance distribution curve, suggesting that common ecological processes are in place worldwide (Ser-Giacomi *et al.*, 2018). These studies also found rare biosphere populations that occasionally appeared and disappeared at different locations, probably due to dispersal events, despite the possible co-occurrence of core rare taxa at specific sites (Ser-Giacomi *et al.*, 2018), as supported by previous studies at smaller scales (Logares *et al.*, 2014). The biogeography of rare microeukaryotes was also assessed for intertidal areas (Y. Zhang *et al.*, 2018), subtropical bays (Mo *et al.*, 2018) and epipelagic waters (Wu *et al.*, 2017) where it was suggested that rare biosphere populations were differently influenced by local (selective pressures) and regional (dispersal) effects.

Studies addressing co-occurrence between microeukaryotes and prokaryotes found that diatom blooms induced shifts in the structure and diversity of rare prokaryotic communities (Zhang *et al.*, 2020) and, similarly, that the diversity of rare microeukaryotes significantly changed during a cyanobacterial bloom (Xue *et al.*, 2018). Rare biosphere ciliate populations have also been addressed (Weisse, 2014). Based on earlier morphospecies analysis of abundant ciliate populations, they were more diverse at the photic layer. However, the identification of rare ciliates by means of high throughput 18S rRNA gene sequencing revealed that most of ciliate diversity lies at deeper water layers instead (Grattepanche *et al.*, 2016).

Most studies on rare microeukaryotes have focused on biogeography, as detailed above, and in general are congruent with biogeography patterns of the prokaryotic rare biosphere. Rare and abundant populations are under different ecological processes, but both include biogeographically restricted and cosmopolitan groups (Ji *et al.*, 2020). Furthermore, both rare prokaryotes and eukaryotes tend to display more niche specialization than their abundant counterparts (Kurm *et al.*, 2017; Xue *et al.*, 2018). Single-celled eukaryotes are the ones usually identified with high throughput sequencing methods, and their strategy of asexual reproduction is more advantageous than sexual reproduction of multicellular, even if microscopic (*e.g.*, rotifers), eukaryotes, as it avoids the need for outcrossing (Weisse, 2014).

2.4. Viral control of rare microbial populations

Few studies have addressed viral particles in the context of the rare biosphere, but viral interactions with microorganisms are relevant for understanding microbial rarity. From the simplest point of view, viral predation is dependent on the probability of encounter with the host, and this probability is lower for low abundance microbes (Pedrós-Alió, 2007). Thus, not considering other virulence factors (*e.g.*, species specificity, host resistance), low abundance is an advantage against predators (Pedrós-Alió, 2007; Lynch and Neufeld, 2015). This aligns well with the killing-the-winner hypothesis, where viral predation increases with host density, thus favoring the lytic cycle in high host density situations (Thingstad, 2000; Thingstad *et al.*, 2014). It has also been proposed that with increasing host abundance, factors such as host defense and horizontal transfer of resistance genes gain importance (Knowles *et al.*, 2016). This provides a positive feedback loop favoring the most abundant populations, termed king-of-the-mountain hypothesis (Våge, Storesund and Thingstad, 2013). Other conceptual models have been proposed, where the lysogenic cycle is not considered (Weitz and Dushoff, 2008). By comparing these models across different datasets, an extension to the killing-the-winner perspective was added, termed piggyback-the-winner, where the viral response to high host density is weighted by the switch from lytic to lysogenic cycles (Knowles *et al.*, 2016).

One study analyzed the effect of marine microbial cells storage in bottles, the so called “bottle effect”, on abundant and rare populations (Haro-Moreno, Rodriguez-Valera and López-Pérez, 2019), and found that low abundance viruses can play an important role in the microbial species abundance distribution dynamics. Essentially, near the wall of the receptacle used to store cells after water sample filtration, the viral particles bounce back and forth from the walls to the microorganisms. Therefore, the probability of encountering and infecting a host increases, resulting in a wall-chain-reaction (Haro-Moreno, Rodriguez-Valera and López-Pérez, 2019). Thus, the community initially sampled can quickly change.

The study of low relative abundance viruses (Holmfeldt *et al.*, 2013), or “rare virosphere”, is dependent on high throughput sequencing based approaches. Traditional difficulties include the small amount of viral nucleic acids and annotation of small sequences (Duhaime *et al.*, 2012; Mizuno *et al.*, 2013). Metavirome assessment has been extensively reviewed elsewhere (Herath *et al.*, 2017). In a study of marine phages (*Cellulophaga* phage), the phage diversity was associated both with genetic novelty and

a range of low abundance phages (Holmfeldt *et al.*, 2013). Due to phage specificity, it is possible that rare phages follow the same patterns of their rare hosts, if able to coexist at low abundance, without extinction. For instance, patterns of seasonality have been found for viruses in the environment, following their hosts behaviors (Sabbagh *et al.*, 2020). Furthermore, a dilution experiment, using marine samples, found that low abundance viruses invest in capsid resistance, rather than burst size (Köstner *et al.*, 2017), supporting the co-existence of viruses and hosts at low abundances. Co-occurrence analysis of rare viral particles, prokaryotes and eukaryotes could lead to important new insights into these dynamics.

The characteristics, patterns and mechanisms governing low abundance viral communities are uncharted territory, but they can be relevant. For example, the white spot syndrome virus is responsible for significant losses in shrimp aquaculture (Sánchez-Paz, 2010) and one of the key aspects in fighting this disease is the early identification of the virus while in low abundance (Pan *et al.*, 2017).

2.5. Rare biosphere and host associations

Microbial symbionts live on or within their hosts, being animals, plants, fungi, or other microorganisms, and while some can be necessary for host survival, others will be present transiently. As is the case for most natural environments, host-associated prokaryotic and microeukaryotic communities seem to follow a species abundance distribution where a few taxa are abundant and most are rare, for example, in sponge host associations (Webster *et al.*, 2009).

Hosts can receive their microbial symbionts from the surrounding environment, a phenomenon known as horizontal symbiont transmission, or they can be transferred during ontogeny (vertical symbiont transmission). For example, some species from the *Symbiodinum* genus are thought to be transmitted vertically in corals (Padilla-Gamiño *et al.*, 2012). It is generally accepted that horizontal transmission of microorganisms into multiple hosts is supported by the existence of a widespread seed bank of rare species in both marine (Webster *et al.*, 2009) and soil environments (Dawson *et al.*, 2017). Hosts contribute to the maintenance of a widespread rare biosphere, either through long distance transportation (mobile animals) or as reservoirs (sessile hosts) (Troussellier *et al.*, 2017), as detailed below.

In marine environments, the rare biosphere in host associations has been studied mostly in sponges (Reveillaud *et al.*, 2014) and corals (Gaidos, Rusch and Ilardo, 2011; Boulotte *et al.*, 2016; Ziegler *et al.*, 2018), although other host organisms have also been surveyed for symbiotic rare microbial species, such as macroalgae (Califano *et al.*, 2020) and fish (Califano *et al.*, 2017).

In sponges, early high throughput sequencing based microbiome surveys found a highly diverse rare biosphere and evidence for horizontal transmission of rare symbionts from the surrounding seawater (Webster *et al.*, 2009). Later studies confirmed this trend, further elucidating that abundant symbionts within sponge hosts usually display wide range of occurrence as rare planktonic populations across marine habitats and were previously considered absent in the surrounding seawater because they escaped detection limits of less-resolving molecular techniques (Taylor *et al.*, 2013).

Generalist symbionts (*i.e.*, found across several host species) contribute greatly to community assembly of each host and may possess a worldwide mode of dispersion as rare members of bacterioplankton communities (Thomas *et al.*, 2016). Intriguingly, low-abundance, generalist and cultivable *Alphaproteobacteria* were found to display genomic

features typically enriched in the microbial metagenomes of marine sponges, thought to be involved in host detoxification (*e.g.*, metabolic waste removal, degradation of xenobiotics), nutrient provision (*e.g.*, B6 and B12 vitamins) and chemical defense (*e.g.*, polyketide biosynthesis) processes (Karimi *et al.*, 2019). This suggests that functions of relevance in host-microbe interactions may be shared between abundant and rare populations of the symbiotic consortium. Furthermore, transient (rare) and core (abundant) microbes may form a pool of phylogenetically redundant organisms in complex symbiotic consortia, which could improve communitarian resilience (Hardoim and Costa, 2014). A spatial and temporal analysis of the *Xestospongia muta* symbiotic consortium also associated its rare populations with transient rarity (Villegas-Plazas *et al.*, 2019). Thus, there is evidence of marine sponges as sessile organisms collecting and storing nearby microbes that remain in its microbiome as core or transient rare taxa (Hardoim and Costa, 2014; Villegas-Plazas *et al.*, 2019).

Early microbiome studies found a potentially inactive rare biosphere within corals, based on rRNA gene and rRNA transcripts comparisons, suggesting that abundant populations are active because they are well adapted to the host environment (Gaidos, Rusch and Ilardo, 2011). Other studies have focused specifically on the association with *Symbiodinium* spp., the iconic photosynthetic symbionts of corals, finding that coral microbiomes include few abundant populations, potentially early colonizers, followed by many low abundance members of that genus (Boulotte *et al.*, 2016; Ziegler *et al.*, 2018). Rare *Symbiodinium* spp. colonizers may have a role in coral stability under fluctuating temperatures, being thus relevant in the current context of climate change and increasing coral bleaching events (Boulotte *et al.*, 2016). The microbiome of green macroalgae (*Ulva*) specimens grown in aquaculture has been compared with that of wild specimens (Califano *et al.*, 2020), and it was found that some rare microbial taxa associated with wild *Ulva* specimens showed increased abundances in the aquaculture environment and may be important to support bottleneck host functions (*i.e.*, algal morphogenesis). A similar pattern was found for animal-associated microbiomes in fish larviculture systems, whereby rare bacterial populations in rearing tanks may be enriched in the fish live feed (*e.g.*, rotifers), which may serve as a vector of both potentially pathogenic and probiotic microorganisms to the feeding fish (Califano *et al.*, 2017).

Studies have also addressed the relevance of the rare biosphere for plant microbiomes (Hol *et al.*, 2010, 2015; Mora-Ruiz *et al.*, 2016; Dawson *et al.*, 2017). In a dilution to extinction study, rare soil microbes were found to induce plants to invest less energy in defense mechanisms and more in nutritional content (Hol *et al.*, 2010). Furthermore, loss

in rare biosphere diversity can negatively impact crop production (Hol *et al.*, 2015). The importance of the rare biosphere for plants was also highlighted by experimental studies linking the establishment of the rhizosphere microbiome with the rare soil biosphere (Dawson *et al.*, 2017). This rare biosphere was possibly active, because of the positive correlation between plant growth and the presence of rare microbial populations in the soil. Another study further compared microbiome structures in distinct plant microhabitats, finding a gradient of decreasing diversity from the plant rhizosphere to the endophytic environment, shaped by horizontal transmission of symbionts from the soil rare biosphere (Mora-Ruiz *et al.*, 2016).

The general ecological processes of the rare biosphere (Jousset *et al.*, 2017) can be contextualized to the host environment, a microecosystem, where both parts, the host and its symbionts, influence each other (in positive, negative and neutral ways). For example, while the rare biosphere is useful in the dispersal of host-specific symbionts across vast geographical ranges, macroorganisms can store and distribute rare microbes, which will eventually find an optimal environment to grow (Trousselier *et al.*, 2017). Certainly, host-symbiont relationships entail a multitude of factors that are likely to dictate microbiome structures across distinct host organisms, such as horizontal and vertical transmission of symbionts, symbiont/pathogen evasion of the host´s immune system, microbe-microbe competitive interactions and host-symbiont cross talk. Therefore, the fundamental differences between the ecological and evolutionary mechanisms shaping the assembly of the free-living versus host-associated rare biosphere should be addressed in future studies.

2.6. Methodological assessment of the microbial rare biosphere

Most studies on the microbial rare biosphere have focused primarily on diversity assessments (examples in Table 2.1), although currently there is an increasing interest in the functional component of low abundance microbes in open and host-associated environments (examples in Table 2.2).

Regardless of the study system under investigation, marker gene high throughput sequencing based methods have been predominantly used for comprehensive analyses of the microbial rare biosphere. For prokaryotes, 16S rRNA gene sequences of specific regions, for example, the V4V5 hypervariable region (Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016), are routinely used in rare biosphere diversity assessments. For eukaryotes, the homologous 18S rRNA gene V9 region is often used (Stoeck *et al.*, 2009, 2010; Nolte *et al.*, 2010). A recent comparison of V4 and V9 18S rRNA gene hypervariable regions advised the usage of both regions for a complete taxonomic view of the rare microeukaryotic diversity and phylogeny (Choi and Park, 2020).

There are several alternatives for the bioinformatics processing of marker gene sequences. Traditionally, the most common method is the clustering of sequences into OTUs, based on a percent identity (*e.g.*, 97%). OTUs can be made *de novo* or reference based. Reference OTUs are limited to the known taxa within the database that was utilized and may change under new database versions. The *de novo* OTUs are created independently of an external reference, meaning that they can identify novel diversity. More detailed information on the consequences of each approach can be found in Callahan, McMurdie and Holmes (2017). A third approach is the use of ASVs (Eren AM *et al.*, 2013; Callahan *et al.*, 2016). This approach does not cluster sequences based on percent identity. Instead, a score of the error rate is calculated and, based on that value, sequences are grouped as the same variant or separated as a distinct one. The score, basically, allows to distinguish true biological differences from random errors (in the process of sequencing and PCR amplification). One consequence of the algorithm is the rejection of singleton sequences, which are assumed to be a variant of another sequence, resulting from a random error, rather than a distinct taxon. This approach provides the sequence itself as output, consequently, it is sample/dataset independent, contrary to OTU picking strategies. Thus, one ASV can be compared across different

studies, while *de novo* OTUs cannot and one ASV can belong to an unknown biological entity, while reference OTUs cannot. Because of these advantages, some authors argue that ASV-based analyses are, currently, the best approach for microbial ecology (Callahan, McMurdie and Holmes, 2017). This has been supported by a study directly comparing the performance of the most used bioinformatics programs to produce ASVs and OTUs (Prodan *et al.*, 2020), but the study focused on taxonomic resolution and did not address its impact on large scale ecological interpretation of data.

One study clustered OTUs (99% identity) and then used the clustered sequences (differing up to 1%) within each OTU to produce ASVs, and found distinct taxa within single OTUs, supporting the importance of high resolving taxonomic classifications (Needham, Sachdeva and Fuhrman, 2017). This study only used abundant OTUs, thus it raises the question of what are the implications of taxonomical resolution in the rare biosphere, does the rank abundance curve change in any significant way? A recent study has shown that community assembly processes of soil rare bacteria are influenced by the level of taxonomical resolution, finding stronger biogeographic patterns for ASVs, rather than OTUs (Bay *et al.*, 2020). Although, at this stage it is still unclear how the taxonomical resolution and broad scale ecological interpretation of the rare biosphere will be influenced by using either approach.

Besides the bioinformatics aspects, DNA based microbial surveys are prone to different biases (McLaren, Willis and Callahan, 2019). Regarding DNA extraction, tested protocols present more differences in rare rather than abundant populations for planktonic communities, in terms of diversity, composition, DNA quality and reproducibility, leading to the suggestion that direct comparison of datasets using different DNA extraction kits are safer for abundant bacteria (Liu, Xue and Yang, 2019). Contamination during sampling and/or DNA extraction can also be problematic. One possibility to monitor the origin of potential contaminations is to characterize the biological communities present in the reagents and materials used in DNA extraction, so-called “kitomes” (Glassing *et al.*, 2016; Fidler *et al.*, 2020). Further, useful bioinformatics tools have been developed to mitigate potential sequencing contamination (Ramírez *et al.*, 2019). Regarding PCR amplification in marker gene studies, it is known that the approach presents bias towards specific taxonomic groups (Sipos *et al.*, 2007), because PCR primers are designed based on well-known taxa.

Table 2.1. Recent publications applying advanced methodologies to access microbial rare biosphere diversity.

Relevance for the assessment of rare biosphere diversity	**Environment**	**Methodology to estimate relative abundance**	**Methodology to identify phylotypes**	**Reference**
Cultivation dependent and independent methods are complementary for the rare biosphere assessment.	Seawater.	16S rRNA gene sequencing.	16S rRNA gene; Traditional cultivation approaches.	(Crespo *et al.*, 2016)
High throughput single-cell cultivation on microfluidic streak plates if effective for rare prokaryotes.	Soil.	NA	High throughput single-cell cultivation on microfluidic streak plates.	(Jiang *et al.*, 2016)
Rare and (so far) uncultivated taxa identified by metagenome assembled genomes.	Forest soil.	Metagenomes.	Metagenome assembled genomes.	(Kroeger *et al.*, 2018)
Cultivation dependent and independent methods are complementary for the rare biosphere assessment.	Cow rumen.	16S rRNA gene sequencing.	16S rRNA gene; Traditional cultivation approaches.	(Zehavi, Probst and Mizrahi, 2018)
Description of metagenome assembled genomes from rare biosphere members, by overlap with 16S rRNA gene.	Crustal subseafloor.	16S rRNA gene sequencing.	16S rRNA gene; Metagenome assembled genomes.	(Ramírez *et al.*, 2019)
Description of previously uncultivated *Spongia officinalis* symbionts, identified as rare by the metagenomics community profiles.	Marine sponge (*Spongia officinalis*).	Metagenomes.	Cultivation under low Carbon conditions.	(Karimi *et al.*, 2019)
Cultivation dependent and independent methods are complementary for the rare biosphere assessment.	Dry desert soil.	16S rRNA gene sequencing.	16S rRNA gene. Traditional cultivation approaches.	(Rego *et al.*, 2019)
High throughput single-cell cultivation on microfluidic streak plates was superior to traditional cultivation approaches in surveying the rare biosphere.	Deep sea sediments.	16S rRNA gene sequencing.	16S rRNA gene; High throughput single-cell cultivation on microfluidic streak plates.	(Hu *et al.*, 2020)
Identification of uncultivated rare taxa by single cell sorting of a selected functional group (cellulose degrading cells).	Wood.	16S rRNA gene.	16S rRNA gene.	(Doud *et al.*, 2020)

NA – Not Applicable

Table 2.2. Recent publications applying advanced methodologies to access microbial rare biosphere functions.

Relevance for functional assessments of the rare biosphere	**Environment**	**Methodology to survey functions**	**Methodology to estimate potential activity**	**Methodology to link function with the rare biosphere**	**Reference**
Sulfate turnover associated with rare biosphere members. Rare taxa remained rare, although the proportion of 16S rRNA: to rRNA gene increased.	Microcosm from peatland soil.	NA	16S complementary DNA (cDNA) amplicon sequencing (rRNA estimation).	16S rRNA gene.	(Hausmann *et al.*, 2016)
Methanogenic archaea rare biosphere.	Soil.	*mcrA* gene sequencing.	NA	*mcrA* gene sequencing.	(Yang, Wang and Qian, 2016)
Description of uncultivated member of the rare biosphere.	Surface water.	Single amplified genomes and multi locus sequencing analysis of phylogenetic markers and functional markers.	NA	Fragment recruitment analysis of metagenomes.	(Royo-Llonch *et al.*, 2017)
Classification of prokaryotes as inactive using 16S rRNA to rRNA gene ratios can be misleading.	Forest floor.	NA	16S cDNA amplicon sequencing (rRNA estimation).	16S rRNA gene.	(Steven *et al.*, 2017)
Sulfite and sulfate reducers belong to the rare biosphere Novelty identified in the rare biosphere.	Multiple ecosystems (*e.g.*, freshwater and artificial).	*dsrB* gene sequencing.	NA	*dsrB* gene sequencing.	(Vigneron *et al.*, 2018)
Study of low abundance sulfate reducers with disproportionately high	Peatland soil mesocosm.	Metagenome assembled genome.	qPCR and metatranscriptomics,	Metagenomes and previous literature.	(Hausmann *et al.*, 2019)

activity, despite constant low abundance for 50 days.					
Functional differences between planktonic and biofilm communities identified by metagenome assembled genomes.	Crustal sub seafloor.	Metagenome assembled genomes.	NA	16S rRNA gene; Metagenome assembled genomes.	(Ramírez *et al.*, 2019)
In-depth view of cellulose degrading functions among rare biosphere taxa.	Wood.	Single cell sorting of cellulose degrading cells using functional screenings.	NA	Single amplified genomes and metagenomics.	(Doud *et al.*, 2020)
Prokaryotic rare communities have the potential to increase functional redundancy under a climate change scenario.	Soil.	Microarrays for nitrogen cycle marker genes (*amoA*, *napA*, *nifH*, *nirK*, *nirS*, *norB* and *nrfA*).	NA	16S rRNA gene.	(Liang *et al.*, 2020)

NA – Not Applicable.

It is possible to calculate the exact bias in specific methodological steps by using comparisons between natural samples and mock communities (McLaren, Willis and Callahan, 2019). This would allow for a deeper understanding of how the biological interpretation of the rare biosphere is influenced by different protocols. The bias in any methodological steps has the potential to distort the view of relative abundance and taxa identification of the rare biosphere.

The problem with amplicon sequencing of marker genes for taxonomical assessments is well illustrated by its inefficiency in describing members of the so-called candidate phyla radiation, because they are (so-far) uncultivated microorganisms and phylogenetically distant from the cultivated microorganisms (Brown *et al.*, 2015). Their unique biology includes self-splicing introns within ribosomal genes and protein coding genes within rRNA genes, which makes marker gene based approaches harder (Brown *et al.*, 2015). Their genome is very small, and it is thought that many are symbionts, based on the lack of essential metabolic pathways (He *et al.*, 2015; Nelson and Stegen, 2015). The candidate phyla radiation has a significant impact on the breadth of the tree of life (Hug *et al.*, 2016) and there is a growing interest in the candidate phyla radiation metabolism (*e.g.*, Jaffe *et al.*, 2020). We argue that the rare biosphere component that belongs to the candidate phyla radiation has been overlooked, because of the marker gene limitations. Open, untargeted metagenome sequencing of total community DNA avoids the PCR bias, but the lack of an amplification step might miss rare sequences (Pedrós-Alió, 2012).

Culture-dependent methods can identify members of the rare biosphere, including those not identified by molecular methods (Shade *et al.*, 2012; Hardoim and Costa, 2014; Crespo *et al.*, 2016; Zehavi, Probst and Mizrahi, 2018; Karimi *et al.*, 2019; Rego *et al.*, 2019) (examples in Table 2.1). For example, in a study using seawater samples, marker gene sequencing with very high sequencing power was compared with traditional cultivation methods, finding that most of the cultivated diversity was missing in the marker gene survey (Crespo *et al.*, 2016). The unique set of conditions used for cultivation was suggested to allow the growth of very low abundance cells, contributing to the high diversity observed in the cultivated consortium (Crespo *et al.*, 2016).

High throughput culturing approaches are also relevant (examples in Table 2.1). For example, microfluidic streak plate, based on cell sorting followed by cultivation, enabled consistent and comprehensive diversity assessments, including the detection of rare

prokaryotes from soil (Jiang *et al.*, 2016). This approach was useful in the isolation of low abundance prokaryotes from deep sea sediments (Hu *et al.*, 2020) and from soil samples (Kurm, van der Putten and Hol, 2019), where it was also found that the rarity status of a given population was not associated with the time necessary for growth. Doud *et al*. (2020) took this approach one step further, using a functional perspective *a priori*, based on preliminary cell sorting of cellulose degrading bacteria, followed by single cell genomics. Using 16S rRNA gene sequencing data available for the whole community, these authors were able to describe a novel cellulose degrading bacterium, *Candidatus* Cellulosimonas argentiregionis, which was rare in the source environment.

Regarding the functional contribution of the rare biosphere to ecosystem dynamics, an interesting combination of multiple methodologies was used to study rare prokaryotes involved in the sulfur cycle. Using ^{13}C stable isotope probing, *Desulfosporosinus* spp. was associated with sulfate reduction in peatland samples and identified as rare prokaryotes in the system (0.006% relative abundance) by 16S rRNA gene sequencing (Pester *et al.*, 2010). Later, by comparing the ribosome and gene copy numbers estimated by qPCR, *Desulfosporosinus* spp. was suggested to be able to persist with an active metabolism, despite not growing abundant (Hausmann *et al.*, 2016). Finally, using Metagenome Assembled Genomes (MAG/MAGs) and metatranscriptomes, it was possible to reveal how the sulfate metabolism was driven by a rare species of *Desulfosporosinus* genus, which remained in low abundance despite being active – The acidic conditions of the soil induced *Desulfosporosinus* spp. to allocate resources to increase the pH of the environment before investing energy for growth (Hausmann *et al.*, 2019).

MAGs from rare populations are difficult to obtain, but some strategies include selective sample enrichment to reduce community complexity before sequencing (Delmont *et al.*, 2015) or differential coverage binning coupled with high sequencing power (Albertsen *et al.*, 2013). Different studies have combined MAGs with phylogenetic markers to study the functions of the rare biosphere (examples in Table 2.2). For example, by overlapping the 16S rRNA gene amplicons with the 16S rRNA gene sequence from the collected MAGs. One difficulty in this strategy resides on the phylogenetic breadth of the rRNA gene of some groups, *e.g.*, *Saccharibacteria*, formerly known as *TM7* (Albertsen *et al.*, 2013), which may be a poor indicator of their full genome diversification. Moreover, it is difficult to reconstruct rRNA genes in MAGs due to their different nucleotide frequency patterns relative to the rest of the genome (Yuan *et al.*, 2015). Using the approach of overlapping 16S rRNA gene of MAGs and 16S rRNA amplicons and metagenomes to

subseafloor crustal samples, planktonic taxa found to colonize biofilms were thoroughly characterized, including rare types, *e.g.*, *Colwellia* (Ramírez *et al.*, 2019). With a similar approach, a survey on Amazon forest soil samples, despite not focusing on the rare biosphere, associated several MAGs with taxa from the rare biosphere, *e.g.*, *TM6* (Kroeger *et al.*, 2018).

Single-cell sorting followed by whole genome amplification is yet another alternative to retrieve genomes from rare prokaryotes, without the need of sequencing whole community metagenomes (Jiang *et al.*, 2016) (examples in Table 2.2). Royo-Llonch *et al*. (2017) described a so far uncultivated *Kordia* species (*Bacteroidetes*) using phylogenetic and functional marker gene analysis of Single Amplified Genomes (SAG/SAGs) obtained from marine samples. A recent study found an improvement in MAG recovery from uncultivated and endosymbiont rare members of soil samples, when applied pre-sorting of cells (Alteio *et al.*, 2020). Another example used samples from wastewater treatment plant to characterize the rare *Chloroflexi* members with a similar approach (targeted SAGs), finding novel diversity missing from previous metagenomics surveys (Dam *et al.*, 2020).

A simpler approach to survey specific functional groups is the use of functional marker genes, instead of (or in combination with) taxonomical marker genes (examples in Table 2.2). For instance, the methanogenic archaea and sulfate reducer functional groups were shown to have a pattern of a few abundant gene variants and many rare ones (Yang *et al.*, 2017; Vigneron *et al.*, 2018). This method has also been useful in increasing our understanding of the rare biosphere impact on the soil nitrogen cycle under climate change (Liang *et al.*, 2020), expanding the knowledge of nitrifiers diversity in wastewater treatment plants using *nxrB* amplicons (Gruber-Dorninger *et al.*, 2015) and in soils, using *amoA* amplicons (Pester *et al.*, 2012).

Mere calculations of gene relative abundances are not necessarily informative of microbial activity. One approach to estimate whether members of the microbial rare biosphere are active is based on the ratio between rRNA gene and rRNA transcript counts (Jones and Lennon 2010), which revealed that rare prokaryotes (Campbell *et al.*, 2011; Hugoni *et al.*, 2013; Richa *et al.*, 2017; Peter *et al.*, 2018) and rare microeukaryotes (Logares *et al.*, 2014; Debroas, Hugoni and Domaizon, 2015) can be active (examples in Table 2.2). However, this methodology has some flaws (Steven *et al.*, 2017), because the half-life time of RNA is very variable across different species and conditions (Steiner *et al.*, 2019). In a simulation study of the usage of this ratio to distinguish active from dormant cells, it was found that factors such as RNA amplification and “sampling depth”

are sources of misclassifications, mostly corresponding to false negative results, that is, phylotypes identified as dormant, but actually active (Steven *et al.*, 2017).

With the fast-paced improvement of molecular and bioinformatics methodologies, knowledge of the microbial rare biosphere is increasing rapidly, expanding from taxonomical and compositional analyses towards a more complex profiling of the metabolism of rare microbial populations (Tables 2.1 and 2.2).

2.7. Fundamental ecological principles and dynamics of the microbial rare biosphere

Many conceptual models have been developed to address microbial community assembly in nature and they usually reflect either deterministic or stochastic processes, or a combination of both. The first model was proposed based on the niche theory (Chesson, 2000; Fargione, Brown and Tilman, 2003), where communities are a direct result of the fitness of different species relative to a set of selective pressures. In contrast, stochastic mechanisms contemplated by the neutral theory of community assembly (Chave, 2004; Hubbell, 2005) imply that species share equivalent probabilities to colonize an area and that, consequently, the species abundance distribution is controlled by random processes (*e.g.*, ecological drift), rather than the difference in fitness. Today it is accepted that there is a continuum of stochastic and deterministic processes, not mutually exclusive, governing community composition and function (Chase, 2010; Ofiteru *et al.*, 2010; Stegen *et al.*, 2013, 2015, 2016). This continuum was summarized in four principles that, in theory, suffice to explain the underpinnings of community assembly patterns in natural ecosystems: selection, dispersal, speciation and ecological drift (Vellend, 2010; Hanson *et al.*, 2012). It is not the objective of this review to scrutinize these principles, extensively reviewed by Zhou and Ning (2017). Instead, we aim at summarizing some important models and patterns found for the rare biosphere to provide a deductive reasoning able to group the large number of perspectives, possibly facilitating the agreement between different studies.

The microbial rare biosphere can be explained by considering both deterministic and stochastic processes as complementary forces that shape community assembly (Ai *et al.*, 2013). Based on the patterns of abundance through space/time suggested by Lynch and Neufeld (2015), different types of rarity have been proposed for rare biosphere species (Jia, Dini-Andreote and Falcão Salles, 2018), ruled by the selective pressures exerted by the repertoire of environmental conditions and biotic interactions which locally affect low abundance populations. In any given time-window, environmental conditions may be constant (*e.g.*, constant levels of salinity) or vary over time (*e.g.*, light availability varying with seasonality). The first case leads to homogenous selection while the second to variable selection (Zhou and Ning, 2017). Homogenous selection explains the existence of taxa that appear to be always rare, *e.g.*, Galand *et al.*, (2009), and have been termed ”permanently rare taxa” (Jia, Dini-Andreote and Falcão Salles, 2018). Variable selection explains why some microorganisms oscillate from rare to abundant,

and these have been termed “conditionally rare taxa” (Jones and Lennon, 2010; Shade *et al*., 2014). Conditionally rare taxa are the environmental equivalent to laboratory enrichment of specific microorganisms (Pascoal, Magalhães and Costa, 2020), and the concept reflects the common observation that some apparently absent or very rare microorganisms in the environment bloom after perturbations, (*e.g.*, Fuentes *et al*., 2016), or the effect of seasonality (*e.g.*, Alonso-Sáez *et al*., 2014). We warn that, in theory, any microorganism can be conditionally rare, because one could not test all possible environmental conditions to determine whether microorganisms exist which will never grow abundant. This note is important, because it highlights the fact that distinguishing different types of rarity is no more than a description of how the relative abundance of microorganisms behaves across a specific set of observed environmental conditions, and not a statement on the biological mechanisms inherent to those microorganisms.

The process of organismal dispersal usually involves both deterministic and stochastic components (Lowe and McPeek, 2014). Due to their small size, the movement of microorganisms across large spatial scales is essentially passive and independent of species fitness, thus microbial dispersal can be assumed as mostly stochastic (Zhou and Ning, 2017). Dispersal can be limited by physical barriers, for example, or may be virtually unlimited in the absence of such barriers, thus resulting in high dispersal rates. Dispersal limitation contributes to the emergence of “transiently rare taxa” (Jia, Dini-Andreote and Falcão Salles, 2018), that is, rare taxa that appear and disappear from a set of observations. This can happen, for example, when a taxon has been randomly moved to another habitat where it does not have the right conditions to grow and, consequently, it can adapt as a new, resident species within the local community or extinguish (diversification and ecological drift effect). Importantly, dispersal limitation is dependent on the number of cells, so it can be more stringent to rare than abundant populations (Taylor and Buckling, 2010). High dispersal rates result in unlimited dispersal, meaning that if a taxon is rare, it will have periodical inputs and outputs of members because of dispersal, as well as ecological drift, these taxa were termed “permanently rare taxa, with variation” (Jia, Dini-Andreote and Falcão Salles, 2018).

The view that the rare biosphere works as a seed bank, providing local communities with resilience to changing conditions (Pedrós-Alió, 2006, 2007), much contributed to the resurgence, in recent years, of the dictum “everything is everywhere, but the environment selects” (Pedrós-Alió, 2007). “Everything is everywhere” is tightly associated with the perspective of unlimited dispersal of microorganisms. Once a microorganism arrives in

a new environment, it may not adapt and extinguish locally (transient) or grow abundant (conditionally rare). Thus, the existence of conditional rarity supports the "everything is everywhere, but the environment selects" perspective (Trousselier *et al.*, 2017). Microbial population growth dynamics under selective pressures are regulated by bottom-up and top-down processes, which will elicit different responses from *r* and K strategists. The first favors rapid reproduction, more common in small organisms, the second favors slower growth rates. As conditionally rare taxa grow abundant, top-down forces will impose a limitation to growth, for example, the killing-the-winner effect. Both bottom-up and top-down effects on microbial growth are dependent on the fitness of the species, thus there are selective pressures, within deterministic patterns. This theoretical framework, schematized in Figure 2.3, can be applied to several types of environments and interactions.

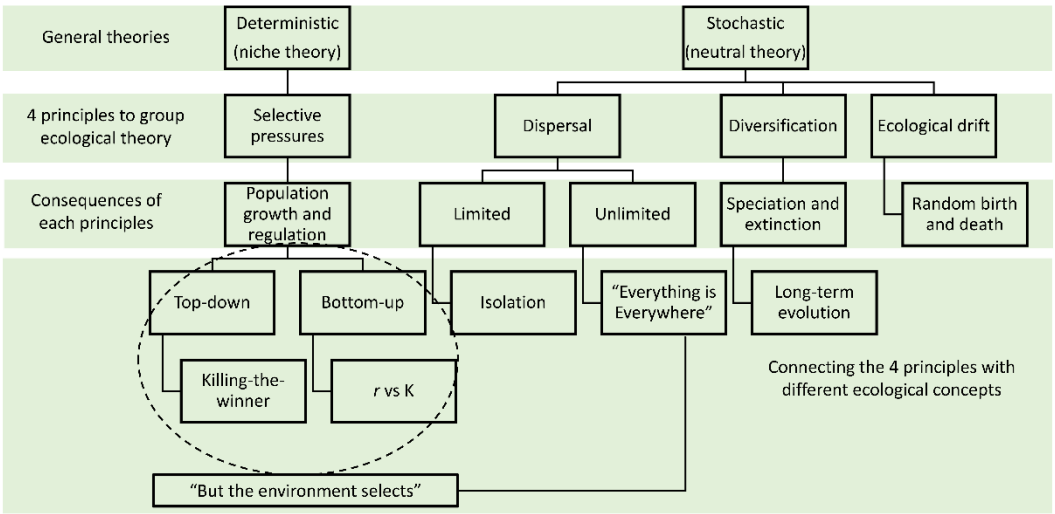


Figure 2.3. Schematic representation of how different ecological theories and concepts are intertwined. Ecological mechanisms exist in a gradient of deterministic and stochastic processes, divided by selection, dispersal, diversification and ecological drift. Selective pressures directly result in population growth dynamics, and these dynamics can be explained by top-down and bottom-up associated strategies (*e.g.*, *r* vs K). Dispersal can be limited or unlimited, the first resulting from some form of isolation and the second resulting in stochastic distribution in the environment. From this point of view, unlimited dispersal drives the "Everything is everywhere" perspective, but the specific conditions at each site pose a selective pressure, driving the "But the environment selects" perspective. Diversification can lead to speciation or extinction. Although these processes may emerge locally because of stochastic events, in the long term they result in patterns of evolution that can be explained from a deterministic point of view. Ecological drift is the random birth and death of cells.

For marine host-associated microbiomes, mobile macroorganisms provide stochastic dispersal (unlimited), while both sessile and mobile hosts provide unique environments that select for specific taxa (Trousselier *et al.*, 2017). This also applies to particles, as microorganisms can randomly (during passive dispersal) attach to particles and form biofilms that provide local selective advantage. This phenomenon was identified, for

instance, in marine microplastics (Kirstein *et al.*, 2019). In marine environments, water currents are limiting factors to microbial dispersal, as previously described for water masses in the Arctic Ocean (Galand *et al.*, 2009, 2010; de Sousa *et al.*, 2019). For estuarine habitats, where the coalescence of river and ocean microbial communities takes place, dispersal homogenizes communities and the local response to the new selective pressures (resulting from water and particle mixing) leads to a decrease/increase in the abundance of different taxa, and one consequence of such structural shifts is the emergence of taxa that are rare in the new, mixed system (Rocca *et al.*, 2020). Microbial dispersal is limited between lakes, and the set of specific selective pressures (homogenous or variable) determines population growth. For example, in lakes from China with homogenous conditions, abundant populations were associated to regional factors (under dispersal limitation) and the rare biosphere to local factors (selective pressures) (Liu *et al.*, 2015). This contrasts with observations made for saltmarsh soil microorganisms, where local factors influenced the structure of both rare and abundant populations (Du *et al.*, 2020). The latter study associated stochastic processes with the assembly of rare types (Du *et al.*, 2020). Interestingly, several studies have identified similar biogeographic patterns for abundant and rare populations that result from different assembly processes (*e.g.*, Mo *et al*., 2018). One untested hypothesis is that there is co-variance between both the environmentally driven selective pressures and geographical dispersal factors (Ji *et al.*, 2020). One last example comes from a survey on soil diatoms biogeography, suggesting spatial radiation, followed by geographical isolation (dispersal limitation) and diversification as major drives of the spatial distribution of these organisms across global geographical scales (Pinseel *et al.*, 2020).

The methodological and mathematical approaches commonly used to disentangle the effects of stochastic from deterministic processes on the assembly of microbial communities have been addressed previously (*e.g.*, Swenson, 2014; Ning *et al*., 2019). One approach is based on the comparison of phylogenetic signals with null models (Webb, 2000; Webb *et al.*, 2002), extended by Stegen *et al*. (2012) and assuming a conserved phylogenetic signal, *i.e.*, a phylogenetic tree driven by non-random taxonomic relatedness (Losos, 2008). First, the phylogenetic signal and phylogenetic null model are compared to test whether the communities are influenced by deterministic or stochastic variables using the nearest taxon index. If they are stochastic, then dispersal and unaccounted processes can be further distinguished (Chase and Myers, 2011). This approach has been used to disentangle stochastic from deterministic patterns in the rare

biosphere (Richter-Heitmann *et al.*, 2020) and was suggested to resolve the different types of rarity (Jia, Dini-Andreote and Falcão Salles, 2018).

2.8. Conclusion and open questions

The microbial rare biosphere is one important piece in the puzzle of microbial ecosystems, with known relevance for the biogeochemical cycles that sustain life on earth and hamper the consequences of human activity. The research on rarity began with animals and plants, revealing patterns that hinted for a microbial rare biosphere, later discovered by high throughput sequencing methods. There are methodological caveats to be resolved, such as sample protocols and bioinformatic methods biases, with consequent implications to the biological interpretation of rarity. However, the improvement and the use of different methodologies will allow for a deeper and more precise view of the rare biosphere.

Low abundance microbial communities have now been addressed across several ecosystems and contexts and making sense of the cumulative knowledge in the field, though challenging, is a timely demand. This review provided deductive reasoning, balancing deterministic and stochastic ecological principles as “best praxis” to understand the structure and function of microbial rare biosphere from an ecosystemic perspective. Despite the recent growth in this field of research, many questions remain open. For example, the co-occurrence patterns between rare and abundant prokaryotes, microeukaryotes, viral particles and hosts across different ecosystems, co-variance effects of deterministic and stochastic processes on the structure of rare microbial communities, and the implications of different reproduction strategies in rare microeukaryotes. The answers to these questions will provide new insights into community assembly theory.

2.9. Declarations

Author contributions

Conceptualization (all authors); Supervision (RC and CM); Visualization (FP); Funding acquisition (all authors); Investigation (FP); Project administration (RC and CM); Writing – original draft (FP); Writing – review and editing (all authors).

Conflict of interest statement

The authors declare no conflict of interest.

Funding

FCT funded this study through a grant to CM (PTDC/CTA-AMB/30997/2017), and a PhD grant to FP (2020.04453). This research was also supported by national funds by FCT through the projects UIDB/04423/2020 and UIDP/04423/2020.

Part II. Advances in marine microbial ecology methods: from seawater sampling to bioinformatics.

Chapter 3. Inter-comparison of marine microbiome sampling protocols.

The materials from this chapter were previously published in:

Pascoal, F. *et al.* (2023) “Inter-comparison of marine microbiome sampling protocols,” *ISME Communications*, 3(1), p. 84. Available at: https://doi.org/10.1038/s43705-023-00278-w.

See author contributions in section **3.6 Declarations**.

3. Inter-comparison of marine microbiome sampling protocols

Francisco Pascoal[1,2], Maria Paola Tomasino[1], Roberta Piredda[3], Grazia Marina Quero[4], Luís Torgo[5], Julie Poulain[6], Pierre E. Galand[7], Jed A. Fuhrman[8], Alex Mitchell[9], Tinkara Tinta[10], Timotej Turk Dermastia[10], Antonio Fernandez-Guerra[11], Alessandro Vezzi[12], Ramiro Logares[13], Francesca Malfatti[14], Hisashi Endo[15], Anna Maria Dąbrowska[16], Fabio De Pascale[12], Pablo Sánchez[13], Nicolas Henry[17,18], Bruno Fosso[19], Bryan Wilson[20], Stephan Toshchakov[21], Gregory Kevin Ferrant, Ivo Grigorov[22], Fabio Rocha Jimenez Vieira[23], Rodrigo Costa[24,25], Stephane Pesant[26] and Catarina Magalhães[1,2]

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal; [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal; [3]Integrative Marine Ecology Department, Stazione Zoologica Anton Dohrn, Naples, Italy; [4]Institute for Biological Resources and Marine Biotechnologies, National Research Council (IRBIM-CNR), Largo Fiera della Pesca 2, 60125, Ancona, Italy; [5]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada; [6]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France; [7]Sorbonne Université, CNRS, Laboratoire d'Écogéochimie des Environnements Benthiques (LECOB), Observatoire Océanologique de Banyuls, Banyuls-sur-Mer, France; [8]Department of Biological Sciences, University of Southern California, Los Angeles; EMBL's European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom; [9]National Institute of Biology, Marine Biology Station Piran, Piran, Slovenia; [10]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark; [11]Department of Biology, University of Padova, Italy; [12]Institute of Marine Sciences (ICM), CSIC. Passeig Marítim de la Barceloneta, 37-49, ES08003, Barcelona, Spain; [13]Department of Life Sciences, University of Trieste, Trieste, Italy; [14]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Japan; [15]Institute of Oceanology Polish Academy of Sciences, Marine Ecology Department, Poland; [16]Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M ECOMAP, UMR 7144, Roscoff, France; [17]CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Sorbonne Université, Roscoff, France; [18]Department of Biosciences, Biotechnologies and Environment, University of Bari, 70126, Bari, Italy; [19]Department of Biology, John Krebs Field Station, University of Oxford, Wytham OX2 8QJ, United Kingdom; [20]Kurchatov Center for Genome Research, Moscow, Russia; [21]Technical University of Denmark, National Institute of Aquatic Resources, Denmark; [22]Département de Biologie, École normale supérieure, Paris, France; [23]Department of Bioengineering, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisbon, Portugal; [24]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisbon, Portugal; [25]European Molecular Biology Laboratory.

Abstract

Research on marine microbial communities is growing, but studies are hard to compare because of variation in seawater sampling protocols. To help researchers in the inter-comparison of studies that use different seawater sampling methodologies, as well as to help them design future sampling campaigns, we developed the EuroMarine Open Science Exploration initiative (EMOSE). Within the EMOSE framework, we sampled thousands of liters of seawater from a single station in the NW Mediterranean Sea (Service d'Observation du Laboratoire Arago [SOLA], Banyuls-sur-Mer), during one single day. The resulting dataset includes multiple seawater processing approaches, encompassing different material-type kinds of filters (cartridge membrane and flat membrane), three different size fractionations (> 0.22 μm, 0.22–3 μm, 3–20 μm and > 20 μm), and a number of different seawater volumes ranging from 1 L up to 1000 L. We show that the volume of seawater that is filtered does not have a significant effect on prokaryotic and protist diversity, independently of the sequencing strategy. However, there was a clear difference in alpha and beta diversity between size fractions and between these and “whole water” (with no pre-fractionation). Overall, we recommend care when merging data from datasets that use filters of different pore sizes, but we consider that the type of filter and volume should not act as confounding variables for the tested sequencing strategies. To the best of our knowledge, this is the first time a publicly available dataset effectively allows for the clarification of the impact of marine microbiome methodological options across a wide range of protocols, including large-scale variations in sampled volume.

Keywords: Marine microbiome, standardized sampling, metagenomics, metabarcoding, microbial diversity, seawater sampling.

3.1. Introduction

The characterization of microbial life on Earth has become a topic of transdisciplinary interest. Indeed, it is now recognized that acquiring knowledge in microbial ecology across multiple biomes is crucial to develop a deeper understanding of life from cells to ecosystems. As a result, massive international collaborative research projects have focused on microbiomes associated with humans (Gevers *et al.*, 2012; Proctor *et al.*, 2019), corals (Huggett and Apprill, 2019), seagrass (https://seagrassmicrobiome.org) or sponges (Thomas *et al.*, 2016; Moitinho-Silva *et al.*, 2017). In addition, over the past 20 years, global coordinated marine planktonic microbiome sampling initiatives have been launched, such as the Global Ocean Sampling (2003-2010) (Venter *et al.*, 2004; Rusch *et al.*, 2007), the International Census of Marine Microbes (ICoMM) (Amaral-Zettler *et al.*, 2010), Malaspina 2010 Circumnavigation Expedition (Duarte, 2015) and Tara Ocean expeditions (2009-2012) (Sunagawa *et al.*, 2020), together with census programs such as the Earth Microbiome program (Gilbert *et al.*, 2010; Gilbert, Jansson and Knight, 2014), and the Micro B3-led Ocean Sampling Day (OSD) (Kopf *et al.*, 2015). Details on advances and perspectives on global ocean microbial ecology, their relevance and future challenges have been extensively reviewed elsewhere (Tara Ocean Foundation, Tara Oceans and European Molecular Biology Laboratory, 2022).

The current massive effort to study the world's microbiomes has given rise to multiple standardization initiatives including the use of common protocols for sampling the microbiome of different environments and host tissues, and of common sequencing procedures. Relevant initiatives with methodological standardization efforts include, for example, OSD (Kopf *et al.*, 2015), Earth Microbiome (Gilbert, Jansson and Knight, 2018), European Marine Omics Biodiversity Observatory Network (EMO BON) (Santi *et al.*, 2021) and Metagenomics for Human Intestinal Tract (MetaHIT) (Qin *et al.*, 2010).

The large-scale analysis of free-living and host-associated microbiomes constitutes a huge step forward in understanding microbes-animal (Bahrndorff *et al.*, 2016; Moitinho-Silva *et al.*, 2017; Huggett and Apprill, 2019; Bosch and McFall-Ngai, 2021) and microbes-plant (Zilber-Rosenberg and Rosenberg, 2008) interactions, as well as the structure, function and diversity of microbial communities in diverse Earth habitats (Alivisatos *et al.*, 2015). However, gaps need to be filled to better standardize and harmonize the best practice and strategies to sample, describe and study microbial diversity. In particular, in ocean microbiome studies it is known that the estimation of microbial richness depends on several factors, including the marker genes and primers

used for metabarcoding (Parada, Needham and Fuhrman, 2016), different DNA extraction protocols (Santos *et al.*, 2012; Shaffer *et al.*, 2022), the sequencing depth and genomic approach (amplicon sequencing vs metagenome sequencing) and clustering criteria (Kirchman, Cottrell and Lovejoy, 2010). Although the sampling strategy is recognized to influence estimations of microbial plankton diversity (Kirchman, Cottrell and Lovejoy, 2010), there is a lack of studies designed to systematically test the effect of methodological variables on the sampling procedures to study ocean microbiome diversity and taxonomic composition. These studies are crucial to design precise protocols to sample the entire size-range of marine microbial communities (Sunagawa *et al.*, 2015).

To study marine microbes, it is necessary to collect seawater and then to concentrate the cells through filtration. The filtered volume is usually in the range of 0.5 L, 1 L, 3-10 L and 100 L (*e.g.*, (Rusch *et al.*, 2007; Hardoim and Costa, 2014; de Sousa *et al.*, 2019; Davidov *et al.*, 2020)) or until the filter gets clogged, depending on sediment particles present, organic matter detritus, cell biomass and/or growing microalgae (Li *et al.*, 2018). Thus, according to the trophic status of the system, its hydrographic conditions and the proximity to terrestrial runoff sources, different volumes of water might be needed, or a pre-filtration step added. It is uncommon to use volume in the range of microliters, but it has been used, for example, to test bacterium-bacterium interactions at millimeter scales (Long and Azam, 2001). It is often possible to find variations in the filtered volume within the same study, for example, because of on-site methodological constraints, *e.g.*, (de Sousa *et al.*, 2019), or for samples intended for different purposes, like DNA and RNA collection, *e.g.*, (Hunt *et al.*, 2013). In addition, there are two main types of filters widely used by the scientific community: 1) cartridge membrane filters, with a pore size of 0.22 µm; and 2) flat membrane filters, which can also be used for size fractionation. The flat membrane filters also differ in the material they comprise (polyethersulfone, polycarbonate, cellulose, etc.), which affects their properties (Boström *et al.*, 2004). A previous study compared amplicon sequencing results of 16S, 18S and 12S rRNA genes for five flat membrane filters with different compositions, and found no significant differences (Djurhuus *et al.*, 2017). Another study, however, highlighted that different filter materials and DNA extraction protocols can introduce false negative detection of microeukaryotic OTUs and underestimate diversity (Deiner *et al.*, 2018). Size fractionation is used to separate microbial cells by size (Matsui *et al.*, 2004; Maruyama *et al.*, 2008), thus selecting prokaryotes from larger microeukaryotes and discriminating free-living cells from particle attached ones. Classically, samples are divided into

picoplankton (0.22-3 µm), nanoplankton (3-20 µm) and microplankton (20-200 µm) size fractions, based on the historic division of planktonic size fractions (Sieburth, Smetacek and Lenz, 1978). However, there are variations across different studies, for example, selecting for picoplankton in the 0.8-3 µm range (Piredda *et al.*, 2018). Previous studies have addressed the distribution of microeukaryotes across size fractions, observing that size fractionation can introduce artifacts from cell collapse and subsequent retention on smaller sized fractions (Massana *et al.*, 2015). Additionally, variables such as the shape and life cycle stage of protists can result in the identification of the same species across different size fractions, as observed for diatoms (Piredda *et al.*, 2018).

Within the framework of the EMOSE 2017 initiative we sampled water from a coastal site of the NW Mediterranean Sea at SOLA station and constructed a unique dataset of deeply sequenced metagenomes and 16S/18S rRNA gene amplicons (MetaB16SV4V5 and MetaB18SV9). To our knowledge, this dataset represents the largest sequencing effort ever conducted at a single site on a single day. The experiment was designed to compare different filtration volumes, filter types, size fractionations and sequencing strategies earlier used in some of the most relevant global ocean initiatives, *i.e.*, Tara Oceans (Sunagawa *et al.*, 2020), Malaspina (Duarte, 2015) and the OSD (Kopf *et al.*, 2015).

The EMOSE sampling was designed to evaluate how microbial diversity estimates change with (i) changing volume of filtered seawater, (ii) different filter types (10 L of water on 0.22 µm cartridge versus flat membrane filters), (iii) whole water filtration versus size fractionation (10 L of water on flat membrane filters); (iv) different size fractions (100 L through 20 µm, 3 µm and 0.22 µm pore size filters); and (v) a single 2.5 L filter versus 4 pooled filters of 2.5 L (0.22 µm pore size, whole water, cartridge membrane). All the comparisons were considered independently for MetaB16SV4V5, MetaB18SV9 and metagenomics, for prokaryotes and protists. Note that several studies include viruses and fungi in their definition of “microbiome”, *e.g.*, Tara Ocean Foundation, Tara Oceans and European Molecular Biology Laboratory, (2022). For the purposes of this article, we are only considering prokaryotes and unicellular eukaryotes, unless stated otherwise.

The size, uniqueness and accessibility of the EMOSE dataset has great potential to help clarify the impact of methodological differences between studies and to contribute to the standardization of applied procedures. It is also open source and freely available and will allow for further investigations beyond the scope of this study.

3.2. Methods

3.2.1. Seawater sampling

Sampling took place at a single location, the SOLA station (42°29'300 N – 03°08’700 E) in the bay of Banyuls-sur-Mer (NW Mediterranean Sea), aboard the research vessel RV Nereis II from the Oceanological Observatory of Banyuls-sur-Mer (OOB). A total of 75 carboys of 20 L were collected on a single day (2017-05-30) from subsurface water (3 m depth) using a high-volume well pump for about 45 minutes. Carboy containers were divided into different seawater sampling strategies to minimize the sampling bias of ship drifting and diurnal variation. The carboy containers used to store the seawater were washed with diluted bleach (10% v/v) the day before and thoroughly rinsed twice with sample water before being filled.

Water filtration was performed following several seawater sampling protocols (Figure 3.1, Supplementary Data 3.1). For the analysis of the total volume of water collected: 1 L was filtered through a single cartridge membrane filter of 0.22 µm; 2.5 L through a single cartridge membrane filter of 0.22 µm; 10 L through a single flat membrane filter of 0.22 µm. For the size fractions analysis, microbes were collected by serial filtration through three filters of decreasing pore sizes, according to the following procedures: 10 L through a mesh filter of 20 µm followed by 3 µm and 0.22 µm flat membrane filters; 100 L through a mesh filter of 20 µm, followed by flat membrane filters of 3 µm and 0.22 µm. All filters were flash-frozen in liquid nitrogen and stored in a freezer at -80 ºC.

The protocols for cartridge membrane filtration were performed with the use of the Sterivex cartridge membrane filter unit (Product Code SVGPB1010, Millipore) with a polyethersulfone membrane, while the protocols for membrane filtration of whole water community (> 0.22 µm) and of the 0.22 µm to 3 µm size fractions used 142 mm diameter polyethersulfone Express Plus membrane filters (Product Code GPWP14250, Millipore). For the 3 µm to 20 µm fractionations, 142 mm diameter polycarbonate membrane filters were used (Product Code TSTP14250, Millipore). As for the large size fractions (> 20 µm), the 47 mm diameter nylon mesh filter was used instead (referred to as flat membrane from here on).

Following the sampling scheme in Figure 3.1 and Supplementary Table 11.1, the present study analyzed 79 seawater samples (n = 157, including successful replicates) according to commonly applied methodologies, expanded to include (and compare) the use of various filter types, different volumes of filtered water, and the division of plankton based

on its size (Supplementary Table 11.1). Note that some replicates were lost during seawater filtration, DNA extraction and/or sequencing, we refer the reader to Supplementary Table 11.1 for information on the replicates successfully obtained. Furthermore, some samples were discarded during rarefaction due to low number of reads (more details below). The metadata relative to each sample, including the sub-samples used for pooling of larger volume samples, are described in detail in Supplementary Data 3.1.

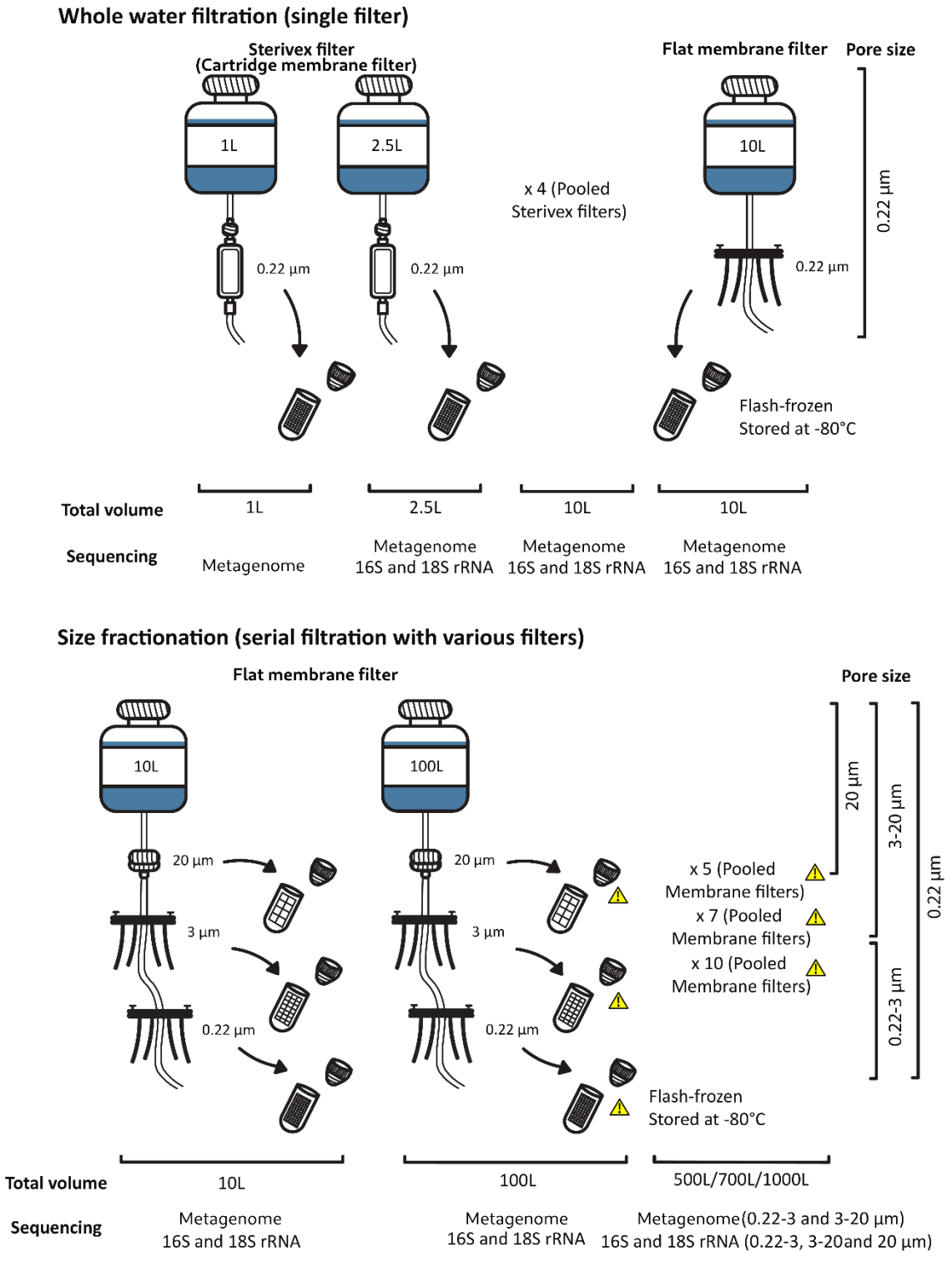


Figure 3.1. Schematic representation of sampling campaign. An attempt was made to have at least three replicates of each step, however, some steps lost replicates and/or volume, those situations are highlighted with an attention sign in this figure. For more details on replicates, see Supplementary Table 11.1.

3.2.2. DNA extraction, amplification, and sequencing

A full description of the following protocols for molecular data production is available in Alberti *et al*. (2017).

Briefly, DNA extraction began by cryo-grinding (SPEX SamplePrep 6870 Freezer/Mil, Fisher Scientific) the filters with lysis buffer, followed by a filter column XL (Macherey-Nagel), with lysis buffer. Purification was done with Nucleospin RNA II, with 1 volume of filtrate and equal volume of ethanol (70% v/v), followed by elution of DNA with nucleopsin buffer set (Macherey Nagel). DNA was quantified by a dsDNA-specific fluorimetric quantitation method using Qubit 2.0 Fluorometer (Thermo Fisher Scientific) with Qubit dsDNA BR (Broad-Range) and HS (High-Sensitivity) Assays and stored at −20 °C.

For DNA shotgun sequencing (metagenome), the library preparation was performed using a protocol for low DNA input. 10 ng of total DNA were sonicated, and sequencing libraries prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England BioLabs). Fragments were end-repaired, 3'-adenylated and NEXTflex DNA barcoded adaptors were added as per the manufacturer's instructions. After two consecutive 1x Ampure XP (Fisher Scientific) clean ups, the ligated products were PCR-amplified with the NEBNext Ultra II Q5 Master Mix (included in the kit), followed by 0.8x AMPure XP purification. Prepared libraries were first quantified by Qubit dsDNA HS Assay measurement. A size-profile analysis was then conducted in an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and by qPCR with the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems, Wilmington, MA, USA) on an MXPro instrument (Agilent Technologies, Santa Clara, CA, USA). Libraries were subjected to Illumina sequencing on a HiSeq 4000 instrument (Illumina), with 150 bp paired end reads layout. Three samples were sequenced with the Rapid HiSeq 4000 instrument, again using the 150 bp paired-end reads layout.

For amplicon sequencing of the V9 hypervariable region of the 18S rRNA gene (MetaB18SV9), DNA was amplified with the primers 1389F 5'- TTGTACACACCGCCC -3' and 1510R 5'- CCTTCYGCAGGTTCACCTAC -3' (Amaral-Zettler *et al.*, 2009). Three PCR reactions per sample were set up using PCR mixtures (25 µL final volume) containing 5 to 10 ng of total DNA template with 0.35 µM final concentration of each primer, 3% of DMSO and 1X Phusion Master Mix. PCR amplifications were performed as follows: 98 °C for 30 s; 25 cycles of 10 s at 98 °C, 30 s at 57 °C, 30 s at 72 °C; and 72 °C for 10 min. PCR products were then pooled and purified by 1.8x AMPure XP beads (Beckman Coulter Genomics) cleanup. The PCR products varied in length from 170-180

bp. A negative control (Nuclease-free water) was included. All libraries were prepared using the NEBNext DNA Modules Products and NextFlex DNA barcodes with 100 ng of purified PCR product as input and sequenced using HiSeq 2500 Rapid (Illumina) machine (150 bp paired-end reads). Three samples were sequenced with the MiSeq instrument.

For amplicon sequencing of the V4V5 hypervariable regions of the 16S rRNA gene (MetaB16SV4V5), DNA was amplified with the primers 515F 5' - GTGYCAGCMGCCGCGGTAA - 3' and 926R 5' - CCGYCAATTYMTTTRAGTTT - 3' (Caporaso et al., 2011; 2012; Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016). For each sample, six reactions were used using the same PCR mixtures as above, with thermal cycling of 30s at 98ºC, followed by 37 cycles of 10s at 98ºC, 30s at 53ºC and 30s at 72ºC, ending with a final 10 min at 72 ºC. Please be aware that these PCR conditions were changed from those in Parada, Needham and Fuhrman (2016), because of modifications due to the polymerase used. PCR products were then pooled and purified by 1x AMPure beads. The PCR product size varied from 300 bp to 700 bp. All libraries were prepared using the NEBNext DNA Modules Products and NextFlex DNA barcodes with 250 ng of purified PCR product as input. In parallel, one negative control (water) and 16 mock communities were used: 8 mock communities of prokaryotes and 8 of eukaryotes (provided by the Jed Fuhrman laboratory, described in Parada, Needham and Fuhrman (2016) and Yeh *et al.*, (2021). After AMPure XP purification (1 volume) and quantification by Qubit fluorometric measurement (HS assay), equimolar pools of amplified libraries were run on a 2% (w/v) agarose gel to select 500–650 bp gel slices (amplicon size increased by Illumina adapters). This sizing step separated the prokaryotic 16S amplicons from the eukaryotic amplification products, which were not sequenced in this study. The sized library was finally purified using the Nucleospin Extract II DNA purification kit. Sequencing was carried out on a HiSeq 2500 Rapid machine, with 250 bp paired-end reads. Eight samples were sequenced, together with respective mock communities, using a MiSeq machine instead, with a 2x300 bp paired-end mode. Please note that sequencing results for samples relative to read length size division and mock communities were made publicly available but were not reported in this article.

3.2.3. Bioinformatics processing of raw sequences

The FASTQ files of the produced sequences were submitted to the European Nucleotide Archive (ENA: PRJEB87662), where they are publicly available. Raw reads were

processed by the MGnify platform (Mitchell *et al.*, 2019). More specifically, Version 5.0 was used for the amplicon data (MataB16SV4V5 and MetaB18SV9), while Version 4.1 was used for the metagenomic data. Briefly, forward and reverse reads were merged with SeqPrep v1.2, quality filtered with Trimmomatic v0.36, reads with less than 100 bp and with more than 10% bp ambiguity were removed. Infernal v1.1.2 (Nawrocki and Eddy, 2013) was used together with Rfam 13.0 (Kalvari *et al.*, 2018) for identification of small subunit (SSU) of the rRNA genes. Amplicon reads were directly attributed taxonomic lineages using pre-computed operational taxonomic units with MAPSeq v1.2.3 (Matias Rodrigues *et al.*, 2017) and Silva database v132 (Quast *et al.*, 2012). For shotgun metagenomics data, mOTU2 (Milanese *et al.*, 2019) were used with Silva database v132 (Quast *et al.*, 2012) for taxonomic lineages. The number of reads at each step for each sample and the respective accession numbers and links are available in Supplementary Data 3.2. Despite the recent revision in the taxonomy of microorganisms at the phylum level, and hence substantial nomenclature modifications (Oren and Garrity, 2021), we used the taxonomic information as provided by the MGnify platform after the Silva database v132 (Quast *et al.*, 2012).

3.2.4. Abundance table pre-processing

The tables with abundance per taxonomic lineage and sample were directly transferred from the MGnify platform to R environment (R Core Team, 2020). We started by dividing the sequencing runs into prokaryotes and protists. Specifically, the MetaB16SV4V5 was filtered to include taxonomic lineages attributed to prokaryotes. We removed any taxonomic lineages attributed to organelles (mitochondria and chloroplasts). Using the same reasoning, for the MetaB18SV9 approach we focused on protists instead and excluded any taxonomic lineages assigned to prokaryotic, metazoan, fungi or viridiplantae taxonomy. Metagenomic data were subdivided into a prokaryotic dataset and a protist dataset, because sequencing all DNA without primer bias allows to identify both biological groups. Notwithstanding, we considered them to be independent biological groups and reasoned that it would be more informative to separate them. This separation was performed after removal of any taxonomic lineages associated with organelles, metazoan, fungi and viridiplantae. For either approach, the taxonomic lineages with NA taxonomy at phylum level were discarded. The abundance tables were then downloaded for manual curation of taxonomy to add a “fake rank” column with relevant taxonomy of both prokaryotes and protists. For protists, we focused on phyla and classes of most interest, while for prokaryotes the “fake rank” included all phyla, but *Proteobacteria* was subdivided by class level.

After taxonomy curation, we removed singleton taxonomic lineages from amplicon sequencing approaches (MetaB16SV4V5 and MetaB18SV9). The sequencing depth was not homogeneous between the variables that we intend to directly compare, which could result in biased comparisons of diversity, likewise we decided to apply rarefaction after removing singletons. The threshold for rarefaction was considered individually for each sequencing approach (MetaB16SV4V5, MetaB18SV9 and metagenomes) and biological group (prokaryotes and protists), because they are independent and represent different orders of magnitude of sequencing depth. Additionally, the specific rarefaction threshold applied should counterbalance the cost of losing too many high-quality reads against losing too many valid samples. Specifically, samples from MetaB16SV4V5 were rarefied to 250 000 reads and three samples were discarded; samples from MetaB18SV9 were rarefied to 100 000 reads and six samples were discarded; samples from metagenomes, considering only the prokaryotic taxonomic lineages, were rarefied to 10 000 reads and 15 samples were discarded; and samples from metagenomes, considering protist taxonomic lineages, were rarefied to 1000 reads and four samples were discarded. The samples discarded are available in Supplementary Table 11.2.

3.2.5. Statistical analysis

All statistical analyses were performed in R (R Core Team, 2020). Alpha and beta diversity metrics were calculated using the *vegan* v2.5.3 package (Oksanen *et al.*, 2018), all figures, except Figure 3.1, were produced with the ggplot2 v3.4.0 package (Wickham *et al.*, 2016) or base R. Statistical tests and their assumptions were tested using the rstatix v0.7.0 package (Kassambara, 2022) and followed the guidelines for best practices proposed in (Kassambara, 2019). The alpha diversity metric used was the total number of taxonomic lineages in a given sample, *i.e.*, species richness. We decided to use a single alpha diversity metric to simplify readability of the results and selected the species richness metric, because it is the most straightforward alpha diversity metric. This alpha diversity metric allows us to assess the direct output of the methodologies compared, making it the most general purpose one. We acknowledge that using a single alpha diversity metric would be very reductive in an environmental research setting. However, we did not make inferences on the ecology of the system and several alpha diversity metrics could be redundant with each other. For a sanity check, we verified that the Shannon index would probably get similar results, while the Simpson index could provide different results, based on correlation analysis (Supplementary Figure 10.1).

We decided to use nonparametric statistical tests to compare species richness between variables, because of the limited number of samples for specific comparisons. For the comparison of two independent groups, we used the Mann-Whitney U test (Bauer, 1972). For more than two independent groups we used the Kruskal-Wallis test (Kruskal and Wallis, 1952). Statistically significant results from Kruskal-Wallis were followed by the Dunn *post-hoc* test (Dunn, 1964). The tests were performed for comparisons with at least three replicate samples per independent group. For each test we specified if the *p* value was significant or not (alpha = 0.05), after adjusting with the Bonferroni method for multiple comparisons (Jafari and Ansari-Pour, 2019).

To describe beta diversity, we compared the dissimilarity between methodological variables. For that purpose, we used Bray-Curtis dissimilarity matrices and visualized the distance between samples with ordination plots, specifically nMDS, with the metaMDS() function of the *vegan* package (Oksanen *et al.*, 2018). Significance values were calculated by PERMANOVA with adonis2() function and homogeneity of variance was verified with betadisper() function, both from the *vegan* package (Oksanen *et al.*, 2018). Finally, the significance of distance to centroid was accessed with Tukey test, with base R functions.

3.3. Results

3.3.1. Environmental communities

For the MetaB16SV4V5 sequencing results (n = 60), we initially obtained between 714 103 and 2 841 890 raw reads per sample (median = 1 462 584 reads, IQR = 267 472 reads). The final number of high-quality reads attributed to taxonomic lineages ranged between 169 945 and 1 517 860 reads per sample (median = 660 878 reads, IQR = 438 932 reads). Thus, between 13.38% and 72.47% of reads were kept after the quality filtering and processing into taxonomic lineages (median = 50.69%, IQR = 30.47%). MetaB18SV9 sequencing results (n = 47) obtained between 670 823 and 3 876 463 raw reads per sample (median = 1 322 612 reads, IQR = 319 425 reads). From those reads, the final, high-quality reads processed into taxonomic lineages ranged between 66 912 and 1 889 838 reads per sample (median = 1 733 379 reads, IQR = 348 105 reads). Thus, between 6.21% and 60.33% of reads were kept after the quality filtering and processing into taxonomic lineages (median = 13.05%, IQR = 26.13%). For metagenomes (n = 50), we obtained between 36 573 050 and 123 310 150 raw reads per sample (median = 58 122 461 reads, IQR = 12 863 822 reads). A minor fraction of the metagenome reads was used for taxonomic identification. Specifically, for prokaryotes, the 16S rRNA reads ranged from 628 to 98 749 (median = 23 376 reads, IQR = 43 212 reads), thus corresponding to a ratio between 0.0011% and 0.0976% (median = 0.0394%, IQR = 0.0664%) of final vs initial raw reads. For protists, the range of final metagenome 18S rRNA reads used was between 396 and 9160 reads (median = 2 568 reads, IQR = 2 807 reads), thus corresponding to a ratio between 0.0006% and 0.0121% (median = 0.0051%, IQR = 0.0046%). The values of the sequencing results are summarized in Supplementary Table 11.3, with additional centrality metrics.

We estimated the predicted number of taxonomic lineages for each level of sequencing depth with rarefaction curves (Supplementary Figure 10.2 and 10.3). MetaB18SV9 was the only sequencing approach that clearly reached the plateau of the rarefaction curve, but just for the 3-20 µm size fraction samples and because of a higher number of reads (Supplementary Figure 10.2). While MetaB16SV4V5 did not reach a clear plateau of the rarefaction curve, it was fairly close (Supplementary Figure 10.2 and 10.3). Metagenomes were closer to the plateau of the rarefaction curve for protists than for prokaryotes (Supplementary Figure 10.2 and 10.3).

3.3.2. The effect of seawater filtered volume in marine microbial diversity, accounting for filter pore size

A considerable range of seawater volumes, from as low as 1 L up to 1000 L, was filtered using several pore sizes (whole water with 0.22 µm, or size fractions with 0.22-3 µm, 3-20 µm and > 20 µm). Additionally, whole water cartridge membrane volumes from 1 L to 10 L were also compared for the metagenomes. Below, we consider the prokaryotes and protists results independently.

3.3.2.1. Prokaryotes

In Figure 3.2, we illustrate the number of prokaryotic taxonomic lineages obtained by filtering the different seawater volumes, separated by filter type (cartridge membrane and flat membrane), pore sizes (whole water and size fractions) and sequencing strategy (MetaB16SV4V5 and metagenomes). The broad view highlights the absence of any clear effect of the filtered volume (1 L to 1000 L) on the number of prokaryotic taxonomic lineages obtained, independently of the filter type and sequencing approach (MetaB16SV4V5 and metagenome) for whole water and the fraction 0.22-3 µm. For flat membrane filters the number of prokaryotic taxonomic lineages increased with increasing pore size, but not with increasing volume (Figure 3.2). Furthermore, the rarefaction curves of MetaB16SV4V5 revealed that the number of prokaryotic taxonomic lineages were consistently divided by size fractions (Supplementary Figure 10.2), but not by volume (Supplementary Figure 10.3). The same differences were observed for metagenomes, but only at a lower number of reads (Supplementary Figure 10.3). The number of prokaryotic taxonomic lineages obtained for each sample are available at Supplementary Table 11.3.

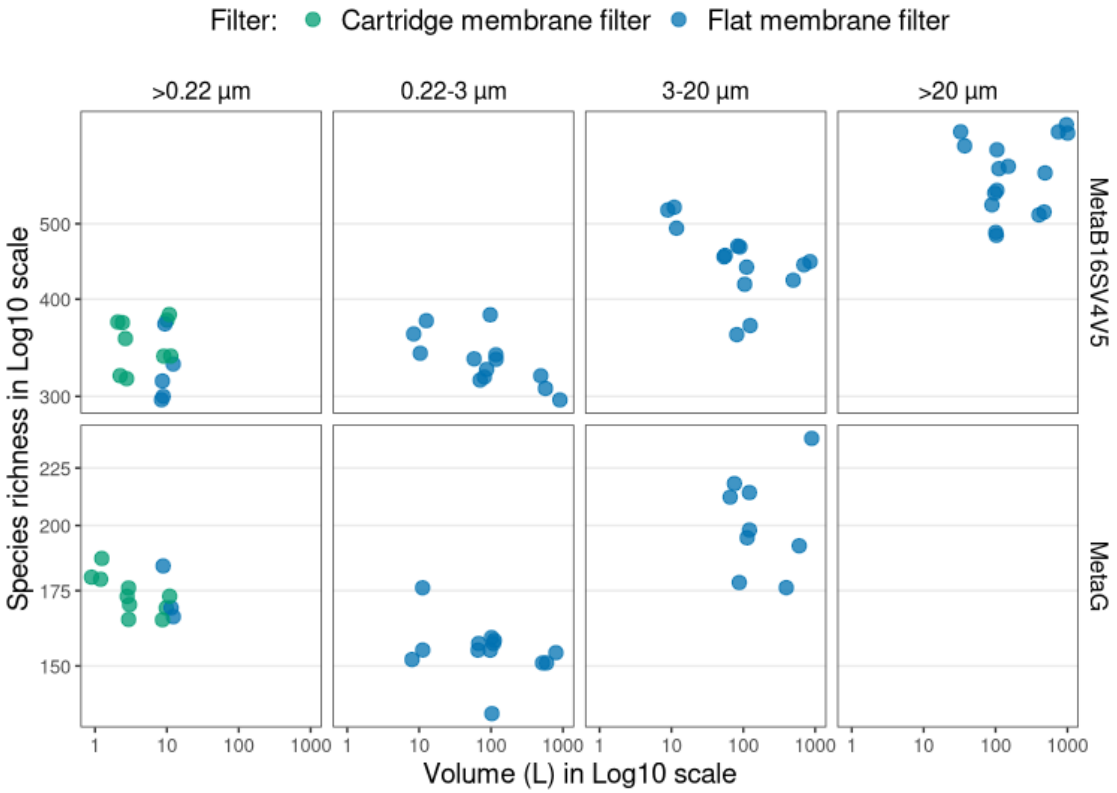


Figure 3.2. Overview of the prokaryotic species richness obtained. The grid divides the possible sequencing strategies in rows (MetaB16SV4V5 or metagenomes - MetaG) and the utilization of whole water (> 0.22 µm) or size fractions (0.22–3 µm, 3–20 µm and > 20 µm) in columns. Color distinguishes between flat and cartridge membrane filters. Within each grid unit, the prokaryotic species richness is plotted against volume, which ranges from 2.5 L to 1000 L.

To clarify the effect of using either a single filter (whole water) or several consecutive filters (size fractions) in the number of prokaryotic taxonomic lineages obtained, we directly compared samples of the same filter (membrane) and volume (10 L). For MetaB16SV4V5, whole water (> 0.22 µm) and 0.22-3 µm size fraction samples presented a similar number of prokaryotic taxonomic lineages (Figure 3.3A) and both presented fewer prokaryotic taxonomic lineages than the 3-20 µm size fraction (Figure 3.3A). Accordingly, the statistical test indicated significant differences in the species richness obtained after > 0.22 µm, 0.22-3 µm and 3-20 µm ($p < 0.05$, Kruskal-Wallis), more specifically, between > 0.22 µm and 3-20 µm size fractions ($p < 0.05$, *post-hoc* Dunn test). On the metagenomes side, for the same comparison, there were no appreciable differences in the number of prokaryotic taxonomic lineages (Figure 3.3A) and they were not significant ($p > 0.05$, Kruskal-Wallis). Details on the above-mentioned statistical tests are available in the Supplementary Table 11.7.

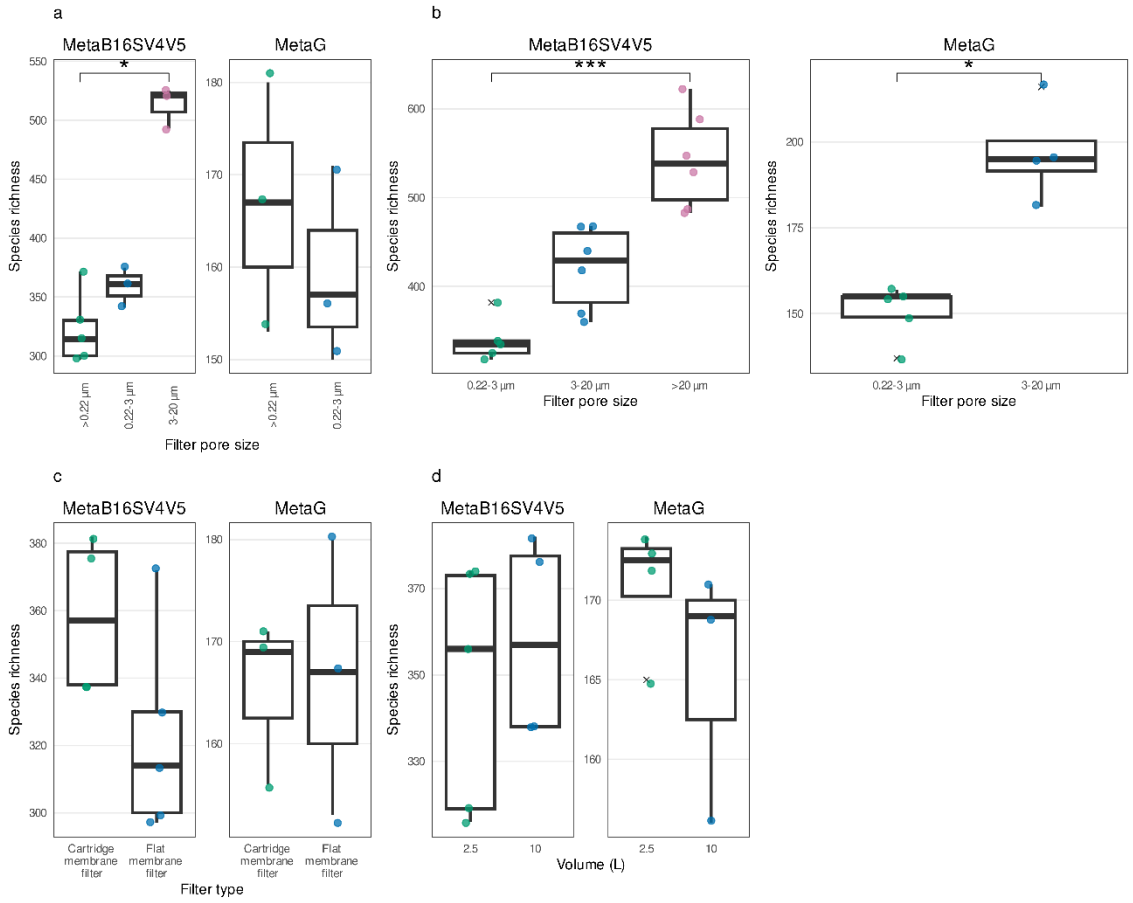


Figure 3.3. Detailed comparison of the prokaryotic species richness across methodological variables. **a.** Comparison for whole water (> 0.22 µm), 0.22–3 µm and 3–20 µm size fractions for the same volume (10 L) and filter (flat membrane), for MetaB16SV4V5 (left) and metagenomes (right). Note that metagenomes did not include samples in 3–20 µm size fraction in (a). **b.** Comparison for size fractions (0.22–3 µm, 3–20 µm and > 20 µm size fractions) for the same volume (100 L) and filter (flat membrane), for MetaB16SV4V5 (left) and MetaG (right). Note that metagenomes did not include samples in > 20 µm size fraction in (b). **c.** Comparison for flat membrane vs cartridge membrane, for the same volume (10 L) and whole water (> 0.22 µm), for MetaB16SV4V5 (left) and metagenomes (right). **d.** Comparison between 2.5 L (single filter) and 10 L (four 2.5 L filters pooled together), using the same filter (cartridge membrane) and whole water (> 0.22 µm), for MetaB16SV4V5 (left) and MetaG (right). All panels illustrate the species richness obtained for each sample (point). To help the reader compare the variables, we added boxplots on top of the points. Significance was determined using either Mann–Whitney test for two independent groups, or Kruskall–Wallis for more than two independent groups, followed by a *post-hoc* Dunn test, if needed. Significance was illustrated with the symbols: $p > 0.05$ (empty); $p < 0.05$ (*); $p < 0.01$ (**); and $p < 0.001$ (***).

Following the same reasoning, we compared the size fractions of 0.22-3 µm, 3-20 µm and > 20 µm, using the flat membrane filter, which revealed an increase in the prokaryotic species richness with increasing pore size, for both MetaB16SV4V5 and metagenomes (Figure 3.3B). In fact, the median number of prokaryotic taxonomic lineages obtained by MetaB16SV4V5 increased significantly from 335 (0.22-3 µm) to 429 (3-20 µm) and 538 (> 20 µm) ($p < 0.05$, Kruskal-Wallis, Figure 3B), more specifically between 0.22-3 µm and > 20 µm size fractions ($p < 0.05$, *post-hoc* Dunn test). Similarly, metagenomes increased the median number of prokaryotic taxonomic lineages from 155 (0.22-3 µm) to 195 (3-20 µm) (Figure 3.3B), which was also significant ($p < 0.05$, Mann-Whitney U

test). Details on the above-mentioned statistical tests are available at Supplementary Table 11.5. Please note that for metagenomes in Figure 3.3B there are no samples for the > 20 µm size fraction, because some samples were lost due to insufficient DNA for sequencing, while some samples that were successfully sequenced were later discarded due to low number of reads (below 10 000 reads, for a list of discarded samples in the rarefaction step see Supplementary Table 11.4). The overview of prokaryotic species richness was overall consistent and supported by the rarefaction curves, because the different size fractions had similar levels of alpha diversity, while the same did not apply for volume (Supplementary Figures 10.2 and 10.3).

To verify the specific effect of the filter (cartridge membrane or flat membrane), filters were compared for the same volume (10 L) and pore size (whole water, 0.22 µm) (Figure 3.3C). The number of prokaryotic taxonomic lineages identified by MetaB16SV4V5 was higher for the cartridge membrane filter (Figure 3.3C), but this difference was not very appreciable, because the range of values for the flat membrane filter included almost the entire range of values from the cartridge membrane filter. More specifically, the number of prokaryotic taxonomic lineages obtained with the flat membrane filter ranged from 297 to 372 (median = 314, IQR = 30, n = 5), while for the cartridge membrane filter this number ranged from 332 to 382 (median = 357, IQR = 40, n = 4). Accordingly, the difference between the number of prokaryotic taxonomic lineages between cartridge and flat membrane filters was not significant ($p > 0.05$, Mann-Whitney U test). Metagenomes provided an equivalent number of prokaryotic taxonomic lineages between either filter (Figure 3C) and the differences were not significant ($p > 0.05$, Mann-Whitney U test). Although we compared cartridge and flat membrane filters under the same volume (10 L), the cartridge membrane filters reached 10 L by pooling together four cartridge membrane filters of 2.5 L. However, the single 2.5 L cartridge membrane filter and 10 L pooled from four cartridge membrane filters of 2.5 L obtained an equivalent number of prokaryotic taxonomic lineages, without significant differences ($p > 0.05$, Mann-Whitney U test) for either sequencing approach (Figure 3.3D). Details on the above-mentioned statistical tests are available in the Supplementary Table 11.5.

The alpha diversity results for the full range of volumes and size fractions were consistent with beta diversity. For either MetaB16SV4V5 and metagenomes, whole water and 0.22-3 µm size fraction samples were clustered together in the nMDS analysis (Figure 3.4A and 3.4B), followed by two other distinct clusters of the samples from the 3-20 µm and > 20 µm size fractions. Additionally, the volume did not follow any clear direction in the ordination figures (Figure 3.4A and 3.4B). PERMANOVA tests were made to support the

ordination figures, with similar results for MetaB16SV4V5 and metagenomes. Specifically, both volume and size fractions significantly changed the community composition ($p < 0.05$, PERMANOVA), but this result should be interpreted with caution, because if the same test considers the division of samples by size fraction, then community composition did not change significantly across volume ($p > 0.05$, PERMANOVA). Details of the PERMANOVA statistical tests for prokaryotes are available at Supplementary Table 11.6. The variation within size fractions, measured by distance to centroid, further supported the clustering of prokaryotic community composition by size fractions (Figure 3.4C and 3.4D, Supplementary Table 11.7).

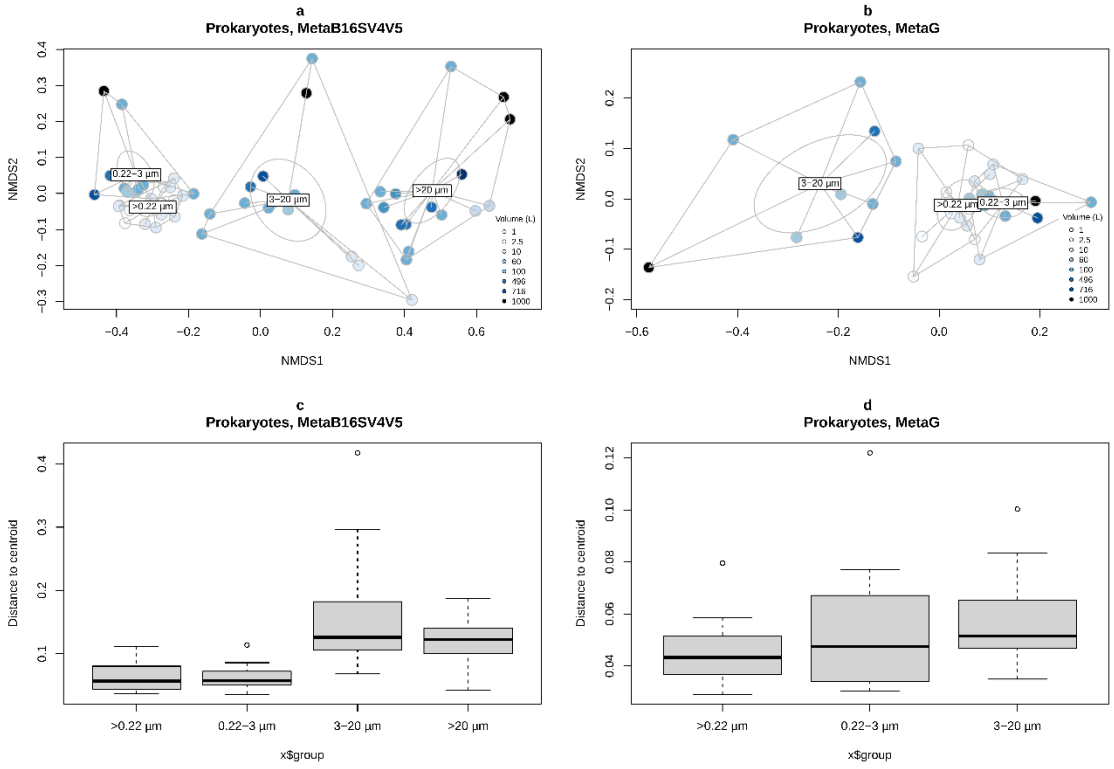


Figure 3.4. Prokaryotic community composition overview. nMDS ordination of dissimilarity values (Bray–Curtis) for the prokaryotic community obtained in each sample. Samples were colored by volume and grouped by whole water (> 0.22 µm), 0.22–3 µm, 3–20 µm and > 20 µm size fractions. Division by (a) MetaB16SV4V5 and (b) metagenomes. Additionally, boxplots represent the distance to centroids of samples within each size fraction, divided by (c) MetaB16SV4V5 and (d) metagenomes. Note that metagenomes didn't include the > 20 µm size fraction. For details on missing replicates, see Supplementary Table 11.2.

The former alpha and beta diversity patterns could be the reflection of a restricted group of dominant taxa, instead of the entire microbial community. To verify possible differences due to taxonomy, the number of taxonomic lineages for each prokaryotic taxonomic group (see Methods) was compared against the volume and size fractions (Figure 3.5). This comparison revealed that size fractions, and not volume, affected the species richness within high level taxonomic groups, for either MetaB16SV4V5 (Figure

3.5), or metagenomes (Supplementary Figure 10.4). A more detailed analysis revealed that prokaryotic species richness across volume changed differently depending on size fraction, as was the case for some major groups (*Gammaproteobacteria*, *Alphaproteobacteria*, *Bacteroidetes*, *Deltaproteobacteria*, *Acidobacteria* and *Plantcomycetes*) (Figure 3.5). Although the metagenomes presented less taxonomic groups at phylum and class level, most were consistently separated by size fraction across volumes, like *Gammaproteobacteria* and *Alphaproteobacteria*, but some were not, like *Betaproteobacteria* and *Thaumarchaeota* (Supplementary Figure 10.4). Note that we analyzed species richness within the selected taxonomic groups and not their relative abundance. To illustrate the difference, we plotted the number of taxonomic lineages attributed to *Candidatus* Marinimicrobia and their relative abundance for each size fraction at 100 L of volume (Figure 3.6). The example from *Candidatus* Marinimicrobia shows that even though it did not change the number of taxonomic lineages (Figure 3.6A), their relative abundance decreased with increasing pore size of the size fractions (Figure 3.6B). It is possible that finer differences exist at lower taxonomic levels for other phyla, but the full analysis of such possibilities goes beyond the scope of this study.

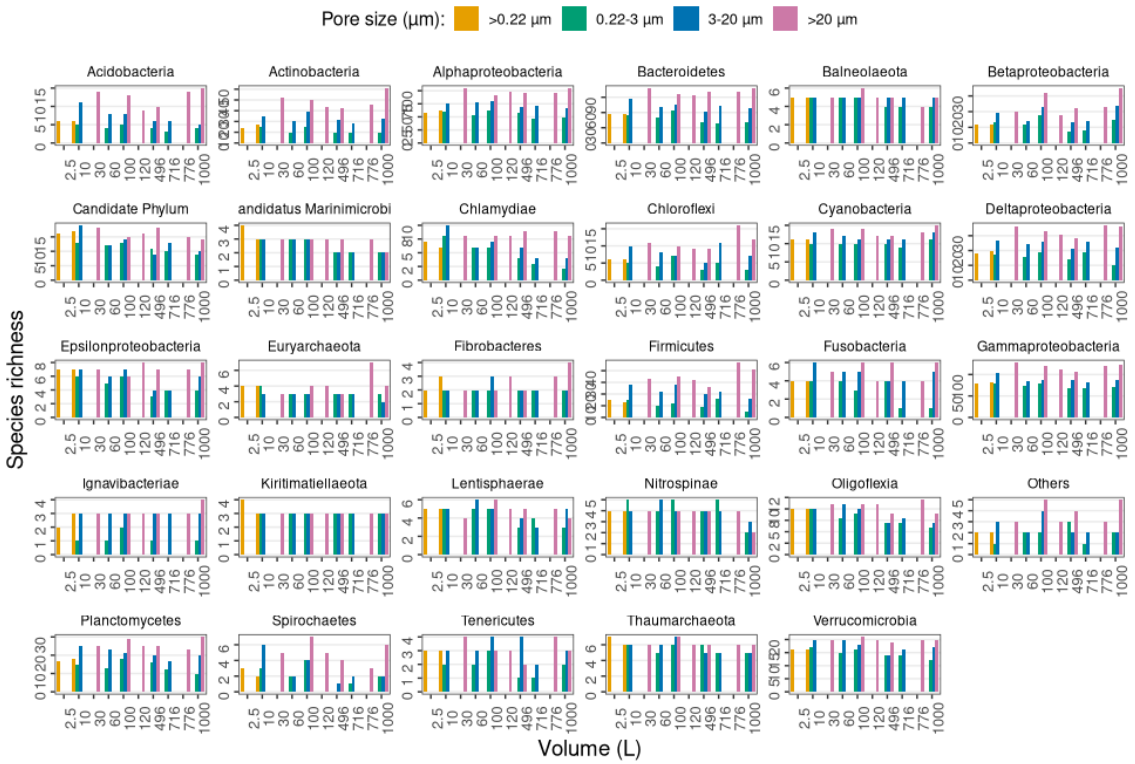


Figure 3.5. Prokaryotic species richness per taxonomic group, from MetaB16SV4V5. Each panel represents the species richness of a specific prokaryotic phyla or class for each volume (1–1000 L). Bar plots indicate species richness and are colored by pore size. The taxonomic group called “Others” includes all phyla that did not reach, at least, 100 taxonomic

lineages across all samples, to avoid an excessive amount of uninformative, redundant panels. The taxonomic group called “Candidate phylum” includes all phyla with Candidate designation, except for *Candidatus* Marinimicrobia.

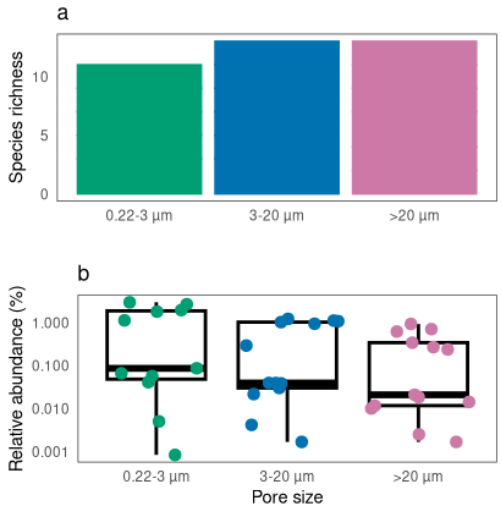


Figure 3.6. Detailed analysis of *Candidatus* Marinimicrobia, from MetaB16SV4V5. **a.** Number of *Candidatus* Marinimicrobia taxonomic lineages; **b.** Relative abundance of the taxonomic lineages from (a) . The values from (a) and (b) were compared for each size fraction (0.22–3 µm, 3–20 µm and > 20 µm) using the same volume (100 L) and filter (flat membrane).

3.3.2.2. Protists

Generally, the protist species richness was more affected by the pore size and by the utilization of consecutive filters, than by the volume (Figure 3.7). More specifically, only whole water filtration (> 0.22 µm) for MetaB18SV9 showed any appreciable change in the protist species richness from 2.5 L (median = 343, IQR = 6.75, n = 4) to 10 L (median = 348, IQR = 34.8, n = 12) (Figure 3.7). However, for either MetaB18SV9 or metagenomes, there was no appreciable difference in the protist species richness from 10 L to 1000 L, within any of the size fractions (Figure 3.7). Comparing pore sizes, whole water (> 0.22 µm), 3-20 µm and > 20 µm size fractions identified more protist taxonomic lineages than 0.22-3 µm size fraction samples (Figure 3.7). The number of protist taxonomic lineages obtained after each sample are available at Supplementary Table 11.8. The higher impact of size fraction, rather than volume, on protist species richness was further supported by rarefaction curves (Supplementary Figure 10.2 and 10.3), even though the size fractions were not as distinct from one another as they were with the prokaryotic data.

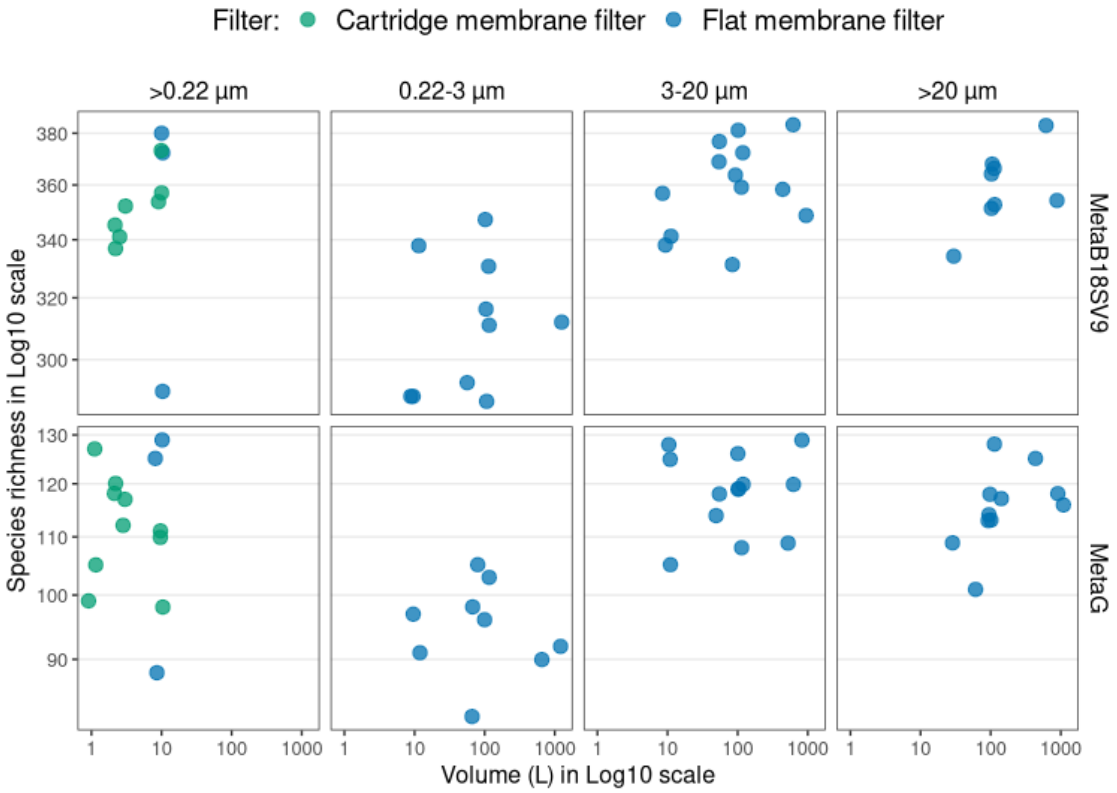


Figure 3.7. Overview of protist species richness obtained. The grid divides the possible sequencing strategies in rows (MetaB18SV9 or MetaG) and the utilization of whole water (> 0.22 µm) or size fractions (0.22–3 µm, 3–20 µm and > 20 µm) in columns. Color distinguishes between flat membrane and cartridge membrane filters. Within each grid unit, the protist species richness is plotted against volume. For details on missing replicates, see Supplementary Table 11.2.

To verify the impact of using either whole water or size fractions, we compared these samples for the same volume (10 L) and filter (flat membrane). Both MetaB18SV9 and metagenomes had fewer protist taxonomic lineages in the 0.22-3 µm size fraction than in the whole water (> 0.22 µm) or 3-20 µm size fraction (Figure 3.8A). However, the range of the number of protist taxonomic lineages obtained for whole water included the range of values for both the 0.22-3 µm and 3-20 µm size fractions (Figure 3.8A). More specifically, the number of protist taxonomic lineages obtained by MetaB18SV9 varied between 290 and 380 for the whole water, 289 and 338 for 0.22-3 µm size fraction, and 338 to 357 in 3-20 µm size fractions (Figure 3.8A), which were not significantly different (p > 0.05, Kruskal-Wallis). The number of protist taxonomic lineages obtained by metagenomes varied between 88 and 129 for whole water, 91 and 97 for 0.22-3 µm, and 105 and 128 for 3-20 µm size fractions (Figure 3.8A); these differences were also statistically non-significant (p > 0.05, Kruskal-Wallis). We note, however, that the number of samples for the metagenome provides little support for the described differences in

this specific comparison. Details on the above-mentioned statistical tests are available at Supplementary Table 11.9.

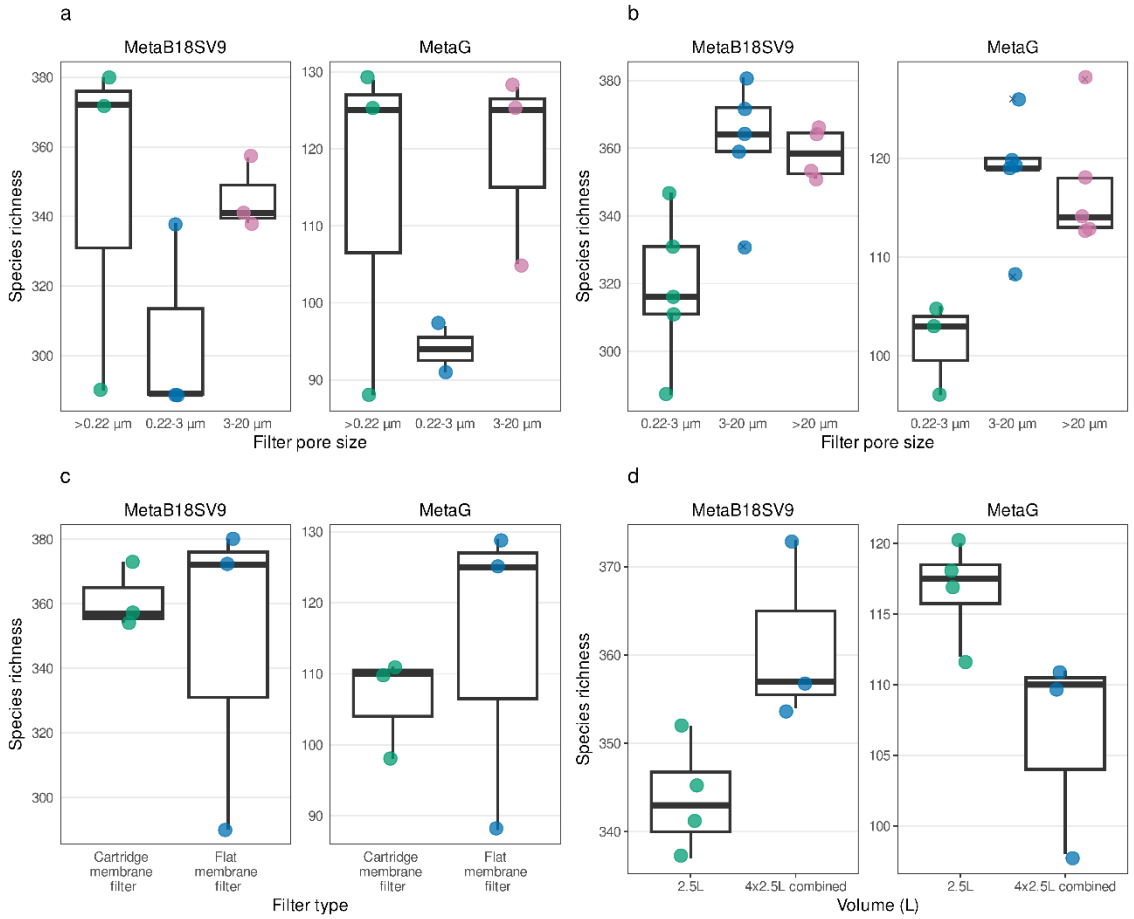


Figure 3.8. Detailed comparison of the protist species richness across methodological variables. **a**. Comparison for whole water (> 0.22 µm), 0.22–3 µm and 3–20 µm size fractions for the same volume (10 L) and filter (membrane), for MetaB18SV9 (left) and MetaG (right). **b**. Comparison for size fractions (0.22–3 µm, 3–20 µm and > 20 µm size fractions) for the same volume (100 L) and filter (membrane), for MetaB18SV9 (left) and MetaG (right). **c**. Comparison for flat membrane vs cartridge membrane, for the same volume (10 L) and whole water (> 0.22 µm), for MetaB18SV9 (left) and MetaG (right). **d**. Comparison between 2.5 L (single filter) and 10 L (four 2.5 L filters pooled together), using the same filter (cartridge membrane) and whole water (> 0.22 µm), for MetaB18SV9 (left) and metagenomes (right). All panels illustrate the species richness obtained for each sample (point). To help the reader compare the variables, we added boxplots on top of the points. Significance was determined using either Mann–Whitney U test for two independent groups, or Kruskal–Wallis for more than two independent groups, followed by a *post-hoc* Dunn test, if needed. Significance was illustrated with the symbols: $p > 0.05$ (empty); $p < 0.05$ (*); $p < 0.01$ (**); and $p < 0.001$ (***).

Directly comparing the size fractions of 0.22-3 µm, 3-20 µm and > 20 µm size fractions for the same filter (membrane) and volume (100 L), the 0.22-3 µm size fraction had fewer protist taxonomic lineages than the 3-20 µm and > 20 µm size fractions (Figure 3.8B), for both MetaB18SV9 and metagenomes. These differences were significant for the MetaB18SV9 ($p < 0.05$, Kruskal-Wallis), but not for the metagenomes ($p > 0.05$, Kruskal-Wallis). However, the significance of the test was not very strong and the *post-hoc* test for MetaB18SV9 was not significant for any combination of size fractions, after

adjustment (p > 0.05, *post-hoc* Dunn test). Details on the above-mentioned statistical tests are available at Supplementary Table 11.9.

To verify the specific effect of the filter type, cartridge and flat membrane filters were compared for the same volume (10 L) and pore size (whole water, > 0.22 µm). The differences in the number of protist taxonomic lineages between cartridge and flat membrane filters were small (Figure 3.8C) and not significant (p > 0.05, Mann-Whitney U test). However, the range of values was wider for the flat membrane filter than the cartridge membrane filter with the MetaB18SV9 approach (Figure 3.8C). The number of protist taxonomic lineages within the replicates of flat membrane filters varied between 290 and 380 (difference of 90 taxonomic lineages), while in the cartridge membrane filters varied between 354 and 373 (difference of 19 taxonomic lineages) (Figure 3.8C). For metagenomes, the values were equivalent between both types of filters (Figure 3.8C). Please note that the cartridge membrane and flat membrane filters were compared at 10 L volume, but the cartridge membrane samples obtained 10 L by pooling together four cartridge membrane filters of 2.5 L together. For MetaB18SV9, the number of protist taxonomic lineages obtained after pooling four 2.5 L cartridge membrane filters was higher than using a single filter of 2.5 L (Figure 3.8D), but not significant (p > 0.05, Mann-Whitney U test). However, this was not the same for the metagenomes, where the number of protist taxonomic lineages was equivalent and slightly higher for a single filter of 2.5 L (Figure 3.8D), but also not significant (p > 0.05, Mann-Whitney U test). Details on the above-mentioned statistical tests are available at Supplementary Table 11.9.

Beta diversity was consistent with species richness, because samples were grouped according to the pore size of the filter (Figure 3.9A and 3.9B). Samples from smaller pore sizes (whole water, 0.22-3 µm and 3-20 µm) were near each other, while the samples for > 20 µm size fractions were distant from the remaining, in either MetaB18SV9 and metagenomes (Figure 3.9A and 3.9B). This was further supported by the significant results of PERMANOVA for the volume and size fractions independently (p < 0.05, PERMANOVA), but once they were considered together the effect on community composition was no longer significant (p > 0.05, PERMANOVA). Note that the variable for size fractions did not meet the homogeneity of variance pre-requisite of PERMANOVA (p > 0.05, betadisper). Details on the PERMANOVA statistical tests for protists are available in Supplementary Table 11.10. Additionally, a more detailed look into the betadisper results, *i.e.*, a measure of distance to the centroid of samples within each size fraction, revealed that samples were very consistent within size fractions (Figure 3.9C and 3.9D, Supplementary Table 11.11).

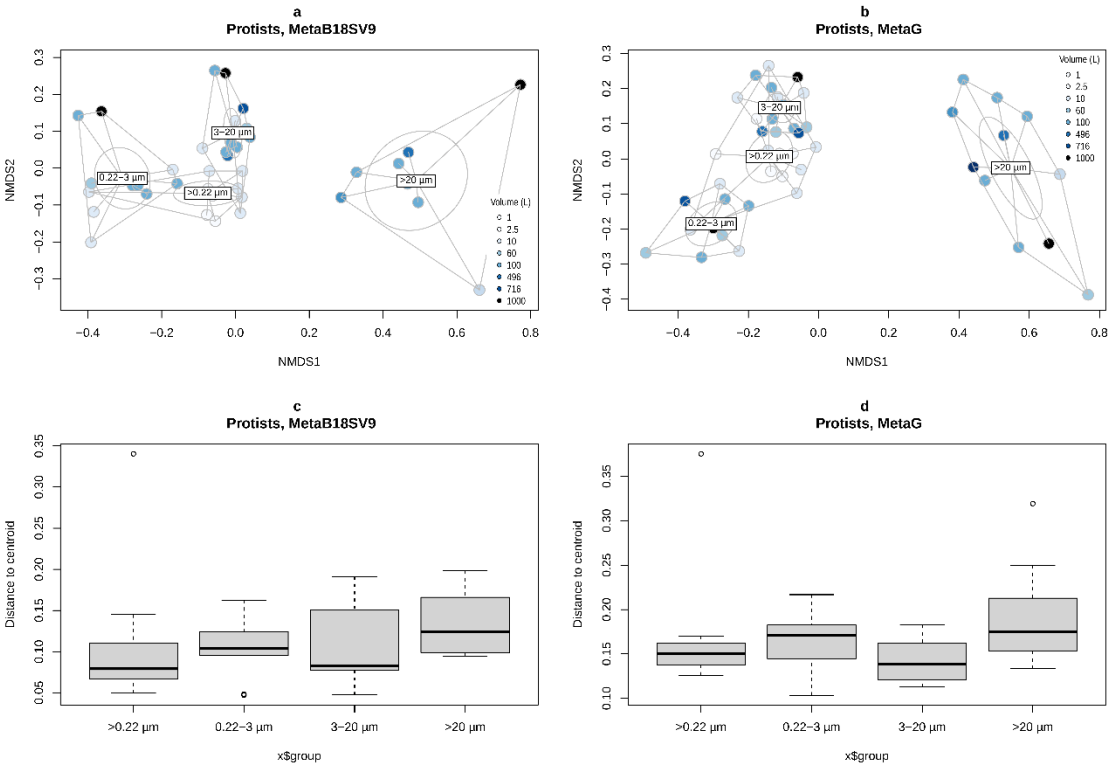


Figure 3.9. Protist community composition overview. nMDS ordination of dissimilarity values (Bray–Curtis) for the protist community obtained in each sample. Samples were colored by volume and grouped by whole water (> 0.22 µm), 0.22–3 µm, 3–20 µm and > 20 µm size fractions divided by (a) MetaB18SV9 and (b) MetaG. Additionally, boxplots represent the distance to centroids of samples within each size fraction, divided by (c) MetaB16SV4V5 and (d) metagenomes.

The taxonomic analysis of MetaB18SV9 did not reveal clear relations between the number of protist taxonomic lineages and volume, although some groups like *Dinophyceae* did show a small increase in their species richness with increasing volume (Figure 3.10). For several protist taxonomic groups, the > 20 µm size fraction consistently identified more taxonomic lineages, independently of the volume, for example, *Dinophyceae*, *Bacillariophyceae* and *Foraminifera* (Figure 3.10). In contrast, other groups were more prevalent in the 3-20 µm size fraction, like *Cercozoa*, *Hacrobia* and *Haptophyta* (Figure 3.10). Several groups did not seem to favor any specific size fraction, like *Excavata* or *Syndinales* (Figure 3.10). In the metagenomes, from 10 L to 1000 L, some groups had more protist taxonomic lineages in the > 20 µm size fraction, like *Bacillariophyceae*, or fewer, like *Hacrobia* (Supplementary Figure 10.5). Additionally, the metagenomes did not reveal any specific taxonomic group that increased the number of protist taxonomic lineages with increasing volume (Supplementary Figure 10.5).

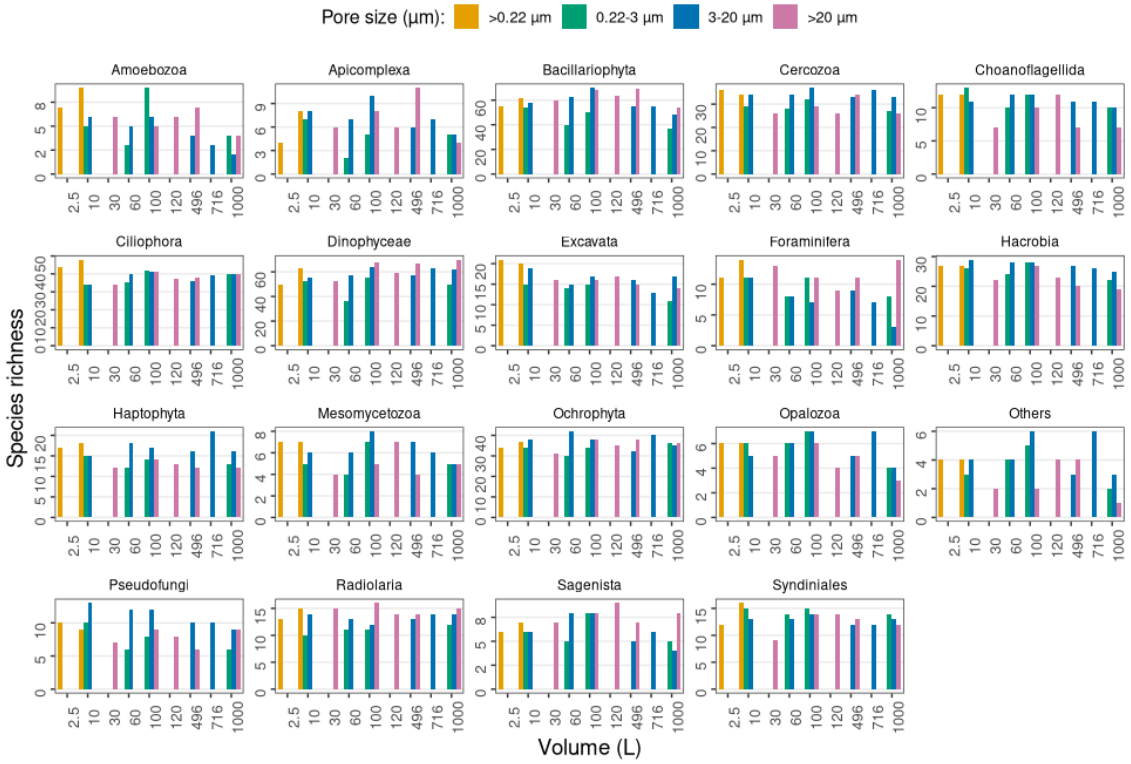


Figure 3.10. Protist species richness per taxonomic group, from MetaB18SV9. Each panel represents the species richness of a specific protist group for each volume (1–1000 L). Bar plots indicate species richness and are colored by pore size. Selected taxonomic groups follow a “fake rank” manually curated to highlight groups of interest, with the less representative groups merged into “Others” designation.

3.4. Discussion

3.4.1. The effect of methodological variation in common seawater sampling protocols

This study directly compared the seawater sampling methodologies used in major sampling campaigns of the global ocean for marine microorganisms – Tara Oceans (Sunagawa *et al.*, 2020), Malaspina (Duarte, 2015), Ocean Sampling Day (Kopf *et al.*, 2015) and European Marine Omics Biodiversity Observation Network (Santi *et al.*, 2021). Our inter-comparison includes cartridge and flat membrane filters, by whole water filtration (single filter), or by size fractionation (serial water filtration of three filters of different pore sizes). The volumes filtered through cartridge membrane filters ranged from 1 L to 2.5 L, with an additional 10 L sample resulting from pooling together 4 samples of 2.5 L. This pooling step is common practice in several laboratories, since cartridge membrane filters are not practical for large volumes of seawater (Walsh, Zaikova and Hallam, 2009). Such studies require the usage of flat membrane filters instead, which in our case served for filtration from 10 L to 100 L, and for pooling 100 L samples into even larger volumes, up to 1000 L.

Our results clearly showed that pore size was the only methodological variable that significantly affected the description of microbial communities in terms of alpha and beta diversity. This was more evident for protists, as expected (Massana *et al.*, 2015; Piredda *et al.*, 2018), and underpins the reason why some expeditions, like Tara Oceans, used size fractionation (Karsenti *et al.*, 2011). Furthermore, studies have shown that pre-filtration steps and size fractions do, in fact, change the perception of microbial diversity (Liang and Keeley, 2013; Piredda *et al.*, 2018). From a taxonomic point of view, most taxonomic groups could be identified in all size fractions. This was most surprising for the prokaryotes, where the > 20 µm size fraction consistently had more taxonomic lineages, indicating that several taxonomic lineages were specifically found in that size fraction. One possible explanation for the identification of taxonomic lineages specific to the > 20 µm size fraction is that those prokaryotes were attached to particles, or to the filter material itself. Considering that the turbidity of the water was very low, the only particles plausible for the prokaryotes to attach to would be the protists or other cell debris, including aggregates. Thus, we suggest that the prokaryotic taxonomic lineages specific to the large size fraction could be prokaryotes associated with microeukaryotes, colonial bacteria and/or specialized in colonizing larger particles. Given the presence of

prokaryotes on > 20 µm size fractions and protists on 0.22-3 µm size fractions, we cannot rule out the possibility that extracellular DNA, besides actual cells, is retained in the filters, for example, by sorption (Liang and Keeley, 2013). However, the general picture is that free-living prokaryotes are identically identified in whole water (> 0.22 µm) and 0.22-3 µm size fraction, while particle-attached prokaryotes can be retained within larger pore size fractions (3-20 µm and > 20 µm). This is consistent with previous studies that account for the effect of pre-filtration on prokaryotic diversity with 16S rRNA gene sequencing (Padilla *et al.*, 2015). Protists also follow the same general picture described in previous studies (Massana *et al.*, 2015), with contamination between smaller size fractions, for example, because of cell fragments. In this study, neither biological group was most unique in composition at > 20 µm size fraction. Notwithstanding, we highlight that it was unexpected to find more prokaryotic and protist taxonomic lineages in the > 20 µm size fraction than in whole water, which cannot be fully explained by our experimental design and should be addressed in future work.

The patterns related to pore size were independent of the filtered volume. In fact, our work demonstrates that the volume of filtered water does not affect species richness and beta diversity of the analyzed sample. In other words, collecting more water, *i.e.*, more cells, did not translate into more prokaryotic or protist taxonomic lineages, nor significant variations in community dissimilarity or different high level taxonomic composition. Additionally, it is noteworthy that we compared both 1 L to 10 L after whole filtration by cartridge membrane, and 10 L to 1000 L after size fractionation by three consecutive flat membrane filters, which exceeds by far the volumes filtered in most marine microbial ecological studies (usually up to 100 L, *e.g.*, Rusch *et al.*, (2007)), and is a higher range than previous studies on the effect of filtered volume and pre-filtration (Padilla *et al.*, 2015). Regarding the utilization of cartridge membrane with 2.5 L or 10 L as a result of pooling together 4 samples of 2.5 L, the diversity metrics used were similar and the differences in community composition were close to zero. However, for specific taxonomic groups, the general rule does not necessarily apply. For example, the number of taxonomic lineages attributed to *Dinophyceae* increased with volume, even though it was a small increase.

Generally, we can argue that, except for pore size, there is little to no difference between protocols, and that the small differences can be due to stochastic events. Notwithstanding, we note that our analysis was mainly based on quantitative estimations of diversity and in specific situations, based on the study design and question, seemingly unimportant differences from a quantitative point of view might be relevant. For example,

if the objective was focused on the *Candidatus* Marinimicrobia, which is abundant in several marine environments and can play roles in marine biogeochemical cycles (Martinez-Gutierrez and Aylward, 2019), it could be relevant to know that even though we can obtain a similar number of taxonomic lineages for each size fraction, their relative abundance decreases with size fraction. The mechanism to justify the higher relative abundance of Candidate phylum Marinimicrobia on 0.22-3 µm size fraction is unknown to us. Despite that, we were expecting to find more abundant members of this candidate phylum in metagenomes, based on previous work comparing 16S rRNA gene amplicon sequencing and metagenomes from Arctic seawater samples (Pascoal *et al.*, 2022a) (Chapter 5). Instead, *Candidatus* Marinimicrobia was rare in the metagenome data (relative abundance below 0.1%). Besides this example, we did not explore taxonomy further, leaving this challenge to future research.

2.4.2. Consistency between sequencing strategies

In this study, we compared the above-mentioned protocol variations in seawater sampling with distinct sequencing strategies. Specifically, amplicon sequencing of the V4V5 hypervariable regions of the 16S rRNA gene, and of the V9 hypervariable region of the 18S rRNA gene, based on well-established primers (Amaral-Zettler *et al.*, 2009; Caporaso et al., 2011; 2012; Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016). Total DNA shotgun sequencing was also included, which is a gene-untargeted approach lacking an amplification step and thus it results in a lower number of single gene reads that can be used to determine microbial diversity, but is not affected by primer bias, *e.g.*, Brown *et al*. (2015). The metagenome derived taxonomic lineages were further divided into prokaryotes and protists. Although the analyses were independently performed for each of the above-mentioned groups, we were able to see that results were consistent between different sequencing strategies. The practical difference was on the number of taxonomic lineages, which was lower for prokaryotes and protists under metagenomes, while the relative difference between methodological variables tested were similar. Few exceptions include the comparison between using a single filter of 2.5 L or using 10 L (4 pooled filters of 2.5 L), where amplicon-based approaches identified more taxonomic lineages in the pooled approach, while the metagenomes identified more taxonomic lineages in the single filter with 2.5 L. We note, however, that the differences were small in either situation. Collectively, it is clear that the methodological variables pose similar effects on observed species richness and beta diversity independently of sequencing strategy. Regarding taxonomy, metagenomes obtained lower resolution at high-level analysis than what was obtained for the amplicon-based approaches, for either

prokaryotes or protists. This lower resolution was a consequence of lower number of SSU gene reads overall, and not any of the sampling methodological variables tested. This was expected from previous studies where similar samples were compared for amplicon and metagenome-based approaches, *e.g.*, Pascoal *et al.*, (2022a).

3.4.3. Strengths and limitations of the EMOSE dataset

This dataset is publicly available (see Data Availability in section 3.6) and contains samples from more methodological variables than the ones presented in this work. For example, we did not use the samples from amplicon sequencing results of the 16S rRNA gene divided by read length at the library preparation stage. Another variable which we did not consider was the sequencing machine (HiSeq or MiSeq), which in principle did not affect results. Despite that, a recent study compared replicates from HiSeq and MiSeq platforms, after 16S rRNA gene amplicon sequencing, and found differences in community composition (Epstein *et al.*, 2021). We note, however, that the aforementioned study focused on the coral microbiome, while ours is focused on seawater microbial communities.

Due to the high number of different variables tested in this study, the EMOSE dataset has hundreds of samples, which is a significant advantage over up-to-date studies, but also has some limitations. Firstly, it might be complex to use the data available due to the high number of variables. Secondly, the high number of samples is a result of many variables tested, not of many replicates (n = 3, wherever possible). However, it should be noted that keeping the same number of replicates across all variables was not possible, due to on-site methodological and logistic impossibilities, including the effort and time needed to filter 100 L of seawater using three different filters. Some limitations are all the more understandable when we take into account the fact that the presented dataset assesses the effect of filtration of several thousands of liters of seawater carried out, as far as we know, for the first time on such a scale during a one-day/one-place campaign.

We decided to compromise with a traditional, easy to understand, rarefaction procedure, but we are aware of its limitations and possible implications. More specifically, rarefaction results in the loss of valid reads and of valid samples and does not account for the compositional nature of high-throughput sequencing data (McMurdie and Holmes, 2014; Gloor *et al.*, 2017). Additionally, another study that directly compared seawater sampling strategies suggested not to standardize the sequencing results (Djurhuus *et al.*, 2017). Some alternatives solve the normalization of compositional data without the need of

removing valid reads, for example, centered-log ratio transformation (Gloor *et al.*, 2017). However, alternative normalization procedures can make interpretation more difficult, for example, by giving negative values of diversity, and can be harder to understand than rarefaction, which is common practice between peers and relatively easy to interpret. We also note that the differences in sequencing depth between samples were considerably large in some cases and there is not a single “best” normalization tool to solve that problem.

Regarding the statistical analysis of species richness, we note that we used non-parametric tests that compare distributions and, as such, they should be accompanied by median and interquartile information, which we illustrated by means of boxplots. Although few replicates were used in some comparisons, which compromises confidence in data distributions (Krzywinski and Altman, 2014), our analyses clearly demonstrate the impact of size fractionation on species richness for both metagenome and amplicon sequencing data. Furthermore, we acknowledge that using a single alpha diversity metric (species richness) might miss some tendencies of the data, we believe this would not be the case for the Shannon index, but it could be for the Simpson index based on correlation analysis (Supplementary Figure 10.1).

In addition to the undeniable scientific importance, the presented research has also significant practical meaning. Considering our results, especially in the case of research which, due to limited resources and time, cannot afford to filter large volumes of seawater, we recommend partial filtration. As we demonstrated, dividing the efforts of filtering, *inter alia*, 1000 L into 100 samples of 10 L (which could be divided into dozens of replicates for each variable), still accurately represents the site in question and does not negatively affect the statistical power of the tests. Compared with, for example, two samples of 500 L each, there would be no improvement in the biodiversity recognized, and the statistical analysis would be compromised. Importantly, the applicative nature of this recommendation mainly concerns studies focusing on quantitative aspects of seawater microbial diversity. For studies with other objectives, the reasoning might not apply or apply differently.

3.5. Conclusion

Our findings highlight that different seawater sample volumes (from 1 L to 1000 L) and the filter types did not affect the identified prokaryotic and protists species richness and beta diversity. In contrast, through serial filtration with membranes of different pore sizes, the size fractionation was the only factor crucial for the results obtained. Furthermore, the use of whole water filtration (> 0.22 µm) was generally equivalent to the 0.22-3 µm size fraction. This metabarcoding and metagenomic comparison of sampling protocols can help researchers to design their own sampling campaigns and to compare studies using different protocols. Even though we did a tremendous effort to address many different variables in protocols used by different campaigns, there is more to be tested and compared for the purpose of standardization of protocols in the future, for example, DNA extraction protocols.

3.6. Declarations

Author contributions

Conceptualization (FP, RC, LT and CM); R script for all statistical analysis and data visualization (FP); support in statistical analysis and data visualization (LT and RP); protist taxonomy validation (RP); design of the EMOSE sampling campaign (CM, AV, and FM); seawater sampling (CM, MPT, SP); DNA extraction, amplification and sequencing (CM and JP); public availability of the EMOSE data (CM and SP); preliminary data analysis (AM, BF, BW, ST, GKF, IG and FRJV); writing – review and editing (FP, AMD, AM, BW, MPT, RP, GMQ, JP, PEG, TT, TTD, AFG, AV, RL, FM, HE, FDP, PS, NH, RC, JAF, and CM); writing – original draft (FP).

Funding

This study was supported by the Euromarine European Marine Research Network by funding the initiative EMOSE 2017 “Inter-Comparison of Marine Plankton Metagenome Analysis Methods”. FCT funded this study through the grant PTDC/CTA-AMB/4946/2020, 2020.03139 CEECIND to CM and a PhD grant to FP (2020.04453). The work of LT was undertaken, in part, thanks to funding from the NSERC. This study was also partially funded by the Strategic Funding UIDP/04423/2020, UIDB/04565/2020 and LA/P/0140/2020 through national funds provided by FCT and partially carried out under the project ATLANTIDA (ref. NORTE-01–0145-FEDER-000040), supported by NORTE 2020, under the PORTUGAL 2020 Partnership Agreement and through the European Regional Development Fund (ERDF). This work was also partially supported by financing attributed to TT from the Slovenian Research Agency (Research Core Funding P1-0237).

Acknowledgments

We are grateful to the captain and the crew of the RV “Nereis II” for their help in acquiring the samples at the Banylus Oceanographical Observatory. We acknowledge Ana Gomes for helping with the design of Figure 3.1. We acknowledge David Needham for his precious counseling on bioinformatics.

Data availability

All raw sequences from the EMOSE dataset, which includes all data used in this article, are available in ENA under the accession number ERP090011. The abundance tables are available at MGnify platform under accession number MGYS00001935. Note that both versions 5 and 4.1 include all sequencing strategies (MetaB16SV4V5, MetaB18SV9 and metagenomes). The metadata for each sample database was recorded in PANGEA (Pesant, 2017), but a cleaner version is available at Supplementary Data 3.1. The R scripts for all data manipulation, statistical tests and figures is available in GitHub (https://github.com/pascoalf/Inter-comparison-of-marine-microbiome-sampling-protocols), no manipulation of data was done outside the scripts provided, except for manual curation of taxonomy, because it needs human experts' evaluation.

3.7. Supplementary Data

Supplementary Data that does not fit within the PhD thesis document can be found in the original publication (Pascoal *et al.*, 2023), under doi: https://doi.org/10.1038/s43705-023-00278-w. Supplementary Data 3.1 of this thesis corresponds to "Supplementary Table S2" file and Supplementary Data 3.2 of this thesis corresponds to "Supplementary Table S3" file, available in ISME Communications page:

https://www.nature.com/articles/s43705-023-00278-w#Sec18

Supplementary Data 3.1. Detailed information regarding the sampling campaign. Adapted from PANGEA repository at https://doi.pangaea.de/10.1594/PANGAEA.879516.

Supplementary Data 3.2. Report of the number of reads after each step of the MGnify platform for each sample and links for quality reports and public repository.

Chapter 4. Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities.

This chapter was based on materials previously published in:

Pascoal, F. *et al.* (2024) “Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities,” *Annals of Microbiology*, 74(1), p. 29. Available at: https://doi.org/10.1186/s13213-024-01767-6.

See author contributions in section **4.6 Declarations**.

4. Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities

Francisco Pascoal[1,2], Pedro Duarte[3], Philipp Assmy[3], Rodrigo Costa[4,5] and Catarina Magalhães[1,2]

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal; [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal; [3]Norwegian Polar Institute, Fram Centre, Tromsø 9296, Norway; [4]Department of Bioengineering, Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico, University of Lisbon, Portugal; [5]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, University of Lisbon, Portugal.

Abstract

High-throughput sequencing of the full-length 16S rRNA gene has improved the taxonomic classification of prokaryotes found in natural environments. However, sequencing of shorter regions from the same gene, like the V4V5 region, can provide more cost-effective high throughput. It is unclear which approach best describes prokaryotic communities from underexplored environments. In this study, we hypothesize that high-throughput full-length 16S rRNA gene sequencing combined with adequate taxonomic databases improves the taxonomic description of prokaryotic communities from underexplored environments in comparison with high-throughput sequencing of a short region of the 16S rRNA gene. To test our hypothesis, we compared taxonomic profiles of seawater samples from the Arctic Ocean using: full-length and V4V5 16S rRNA gene sequencing in combination with either the GTDB or the Silva taxonomy database. Our results show that all combinations of sequencing strategies and taxonomic databases present similar results at higher taxonomic levels. However, at lower taxonomic levels, namely family, genus, and most notably species level, the full-length approach led to higher proportions of ASVs assigned to formally valid taxa. Hence, the best taxonomic description was obtained by the full-length and GTDB combination, which in some cases allowed for the identification of intraspecific diversity of ASVs. We

conclude that coupling high-throughput full-length 16S rRNA gene sequencing with GTDB improves the description of microbiome profiling at lower taxonomic ranks. The improvements reported here provide more context for scientists to discuss microbial community dynamics within a solid taxonomic framework in environments like the Arctic Ocean with still underrepresented microbiome sequences in public databases.

Keywords: V4V5 16S rRNA gene amplicon sequencing, full-length 16S rRNA gene amplicon sequencing, GTDB, Silva, Arctic Ocean microbiome, high-throughput sequencing.

4.1. Introduction

The Arctic Ocean is an underrepresented ecosystem in microbial ecology surveys, which is reflected by the lack of deposited metagenomes in public databases (Priest *et al.*, 2021). Studies on Arctic Ocean prokaryotic ecology are currently using second-generation sequencing technologies, such as Illumina, because their high throughput allows for a representative view of prokaryotic communities (Wilson *et al*., 2017; de Sousa *et al*., 2019; Fadeev *et al*., 2021; Pascoal *et al.*, 2022a; Thiele *et al*., 2022). Even though the Illumina biochemistry is cost-effective, it typically delivers short sequence reads of up to 300 bp or roughly double the size for the paired-reads (Tedersoo *et al.*, 2021), which limits the analytical workflow to specific hypervariable regions of the 16S rRNA gene, in turn making it difficult to design truly universal primers. Consequently, several primer pairs have been validated and optimized for different environments and contexts, for example, the V4V5 region for marine environments (Parada, Needham and Fuhrman, 2016). However, lower taxonomic level classification is inherently difficult, because the short regions do not provide enough resolution for an accurate distinction between closely related species (Johnson *et al.*, 2019). Although primer bias cannot be avoided in any amplicon-based approach, by increasing the size of the amplicon, the taxonomic resolution power increases (Johnson *et al.*, 2019). Using PacBio circular consensus sequencing (CCS) it is possible to obtain average read lengths of up to 30 000 bp (Tedersoo *et al.*, 2021) which is two orders of magnitude larger than Illumina. By means of PacBio CCS of the full-length 16S rRNA gene and raw read processing with *DADA2* (Callahan *et al.*, 2019; Kumar *et al.*, 2019), it was possible to describe mock communities with 100% accuracy and correctly distinguish pathogenic and non-pathogenic *Escherichia coli* strains (Callahan *et al.*, 2019). High-throughput long-read sequencing approaches have been extensively reviewed elsewhere (Tedersoo *et al.*, 2021).

Besides the full-length 16S rRNA gene sequencing, database selection can improve the accuracy of species-level classification (Myer *et al.*, 2016; Schloss *et al.*, 2016; Rodríguez-Pérez, Ciuffreda and Flores, 2022) and habitat-specific reference databases can further improve species-level classification compared to universal reference databases (F Escapa *et al.*, 2020; Costa *et al.*, 2022; Overgaard *et al.*, 2022). However, there are no habitat-specific databases for underexplored environments such as the Arctic Ocean. Some of the most used taxonomic databases for microbial ecology studies are the Silva (Pruesse *et al.*, 2007), Greengenes (McDonald *et al.*, 2012), and Ribosomal Database Project (RDP) (Maidak *et al.*, 1996) databases. Since Silva is updated

frequently and includes reference sequences from environmental samples, it was used in several prokaryotic ecology surveys of the Arctic Ocean (*e.g.*, (de Sousa *et al.*, 2019; Pascoal *et al.*, 2022a; Thiele *et al.*, 2022)). Recently, GTDB was introduced (Parks *et al.*, 2022), taking advantage of the increasing number of available pure culture and MAGs in taxonomy assignments. Briefly, Silva manually assigns taxonomy based on the phylogenetic tree of the small and large subunits of the rRNA gene (Quast *et al.*, 2012; Yilmaz *et al.*, 2014). GTDB identifies species clusters by whole-genome average nucleotide identity and solves higher ranks with relative evolutionary divergence; this method allows consistent and automatic classification of genomes (Parks *et al.*, 2022).

Recent studies have compared specific short regions and full-length sequencing of the 16S rRNA gene in several environments, for example, seawater (Wang *et al.*, 2022), contaminated soil (Yan *et al.*, 2023), coral microbiome (Pootakham *et al.*, 2019, 2021), cow rumen (Brede *et al.*, 2020), soybean rhizosphere (Yu *et al.*, 2022), and fish microbiome (Klemetsen, Willassen and Karlsen, 2019). However, to the best of our knowledge, the short- and full-length sequencing of the 16S rRNA gene, as well as the GTDB and Silva databases, have not been compared for seawater samples from the Arctic Ocean. The Arctic Ocean is an underrepresented and extreme environment with potentially novel biodiversity (Priest *et al.*, 2021), and where long-term microbial monitoring programs are established (Renner, Dodd and Fransson, 2018; Fadeev *et al.*, 2021).

In this study, we hypothesize that the best combination of methods to survey the microbial communities of the Arctic Ocean at lower taxonomic levels is the full-length sequencing of the 16S rRNA gene by means of third generation sequencing, with GTDB for taxonomic assignment of ASVs. To test this hypothesis, we compared the taxonomic resolution offered by full-length 16S rRNA gene sequencing (by PacBio CCS) versus V4V5 hypervariable region sequencing (short reads) of the 16S rRNA gene (by Illumina), taxonomically classified with either Silva or GTDB databases. These combinations were applied to seawater replicate samples from Kongsfjorden, Svalbard and eastern Fram Strait collected during the Norwegian Polar Institute Monitoring Cruise in 2019 within the context of the Environmental Monitoring of Svalbard and Jan Mayen (MOSJ).

4.2. Methods

4.3. Sampling campaign

During the Norwegian Polar Institute Monitoring Cruise from 8 – 13 August 2019, seawater samples were collected at surface (within upper 10 m), the Chlorophyll *a* maximum depth (varied between 5 and 28 m), and 10 m above the seafloor (except at deep Hausgarten [HG-IV] Station, where the deepest sample was taken at 2300 m). Samples were collected along a transect from inner Kongsfjorden across the West Spitsbergen shelf and out to Fram Strait (Figure 4.1). We used the seawater sampling protocol previously described for the Arctic Ocean by de Sousa *et al*. (2019). In this study, 18 seawater samples from six stations (three samples per station, each representing a different depth within each station - Table 4.1) were selected for DNA extraction followed by 16S rRNA gene amplicon sequencing using Illumina and PacBio sequencing strategies as detailed below. Therefore, a total of 36 16S rRNA gene sequencing libraries (n = 18 for each sequencing approach) were processed and subjected to downstream analyses as described below. Of the 18 samples under study, the sampling volume varied between 1 L and 5 L (mean = 3.2 L, sd = 1.3 L) and the depths varied between 1 m and 2300 m according to the unique bathymetric features of each station (see Table 4.1 for details).

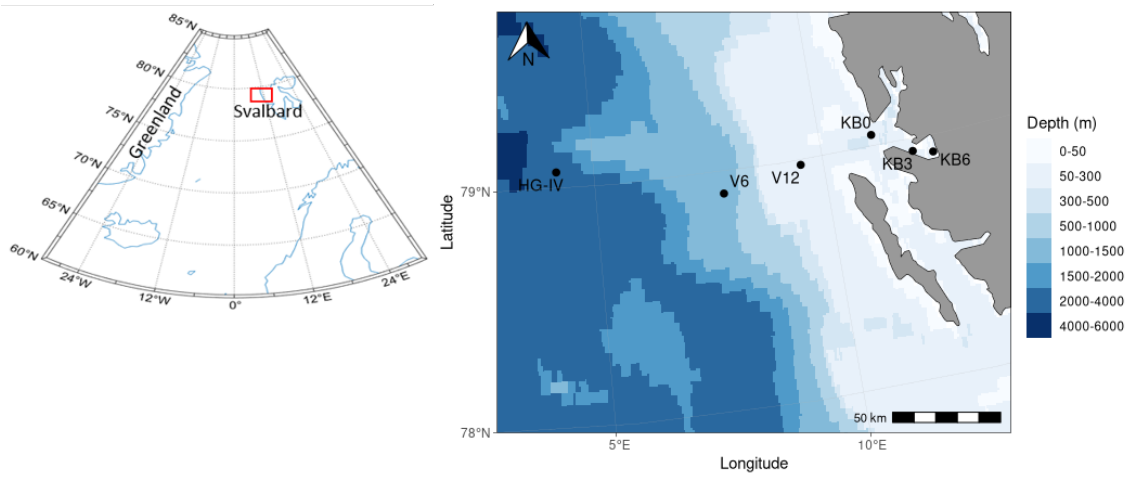


Figure 4.1. Map of sampling campaign. The left map provides a geographic context of the sampling area. The right map shows the coordinates of each sampling station through the Kongsfjorden and the Fram Strait and indicates the depth of the seafloor.

Table 4.1. Metadata of MOSJ 2019 samples that were used.

Sample code	Station	Sampling depth (m)	Station depth (m)	Latitude	Longitude	Collection Date (UTC)	Filtration Volume (mL)
KB3_S_R1	KB3	5	342	N 78 57.38	E 11 56.81	09/08/2019	1800
KB3_M_R1	KB3	15	342	N 78 57.38	E 11 56.81	09/08/2019	3000
KB3_B_R1	KB3	300	342	N 78 57.38	E 11 56.81	09/08/2019	5000
KB6_S_R1	KB6	1	52	N 78 56.01	E 12 23.11	09/08/2019	1750
KB6_M_R1	KB6	7	52	N 78 56.01	E 12 23.11	09/08/2019	1000
KB6_B_R1	KB6	50	52	N 78 56.01	E 12 23.11	09/08/2019	1000
KB0_S_R1	KB0	5	331	N79 02.81	E 11 06.62	11/08/2019	1500
KB0_M_R1	KB0	14	331	N79 02.81	E 11 06.62	11/08/2019	3500
KB0_B_R1	KB0	320	331	N79 02.81	E 11 06.62	11/08/2019	3500
V12_S_R1	V12	5	220	N 78 58.812	E 9 29.171	11/08/2019	3000
V12_M_R1	V12	28	220	N 78 58.812	E 9 29.171	11/08/2019	3200
V12_B_R1	V12	215	220	N 78 58.812	E 9 29.171	11/08/2019	3100
V6_S_R1	V6	5	1127	N 78 54.59	E 7 47.44	11/08/2019	4000
V6_M_R1	V6	25	1127	N 78 54.59	E 7 47.44	11/08/2019	5000
V6_B_R1A	V6	1100	1127	N 78 54.59	E 7 47.44	11/08/2019	4000
HGIV_S_R1	HG-IV	5	2407	N79 04.333	E 4 09.262	13/08/2019	4500
HGIV_M_R1	HG-IV	18	2407	N79 04.333	E 4 09.262	13/08/2019	4200
HGIV_B_R1	HG-IV	2300	2407	N79 04.333	E 4 09.262	13/08/2019	5000

4.2. DNA extraction, amplification, and sequencing

DNA extraction was described previously (Semedo *et al.*, 2021), using the DNeasy PowerWater Sterivex Kit (QIAGEN Laboratories, Inc.). Amplicon sequencing was performed by the Integrated Microbiome Resource (IMR), following their protocol (https://imr.bio/protocols.html). For the V4V5 region of the 16S rRNA gene sequencing by Illumina, the selected primers were 515F (5'-GTGYCAGCMGCCGCGGTAA-3') and 926R (5'-CCGYCAATTYMTTTRAGTTT-3') (Caporaso et al., 2011; 2012; Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016). For the full-length 16S rRNA gene sequencing by PacBio CCS, the primers used were 27F (5'-AGRGTTYGATYMTGGCTCAG-3') and 1492R (5'-RGYTACCTTGTTACGACTT-3') (Paliy *et al.*, 2009).

4.2.3. Bioinformatics processing of reads

Sequences generated by Illumina were processed using *DADA2* package for quality filtering and chimera removal (Callahan *et al.*, 2016), with default parameters and trim lengths of 249 nt (Forward) and 214 nt (Reverse). For the sequences generated by PacBio CCS, bam files were processed into FASTQ files by the sequencing provider. From the FASTQ files, primers were removed, and reads were filtered to fit within a range of 1000 bp and 1600 bp, filtered sequences were approximately 1500 bp length. Quality filtering and chimera removal of long reads was performed by DADA2 (Callahan *et al.*, 2016), following the specification for PacBio CCS reads (Callahan *et al.*, 2019; Kumar *et al.*, 2019).

4.2.4. Taxonomic classification

For the V4V5 16S rRNA gene sequencing, taxonomic classification of ASVs down to genus level was performed with Naive-Bayes algorithm (Wang *et al.*, 2007) and with exact matching for species level (Edgar, 2018). For the full-length 16S rRNA gene sequencing, the Naive-Bayes algorithm (Wang *et al.*, 2007) was used for all taxonomic levels in the main analysis. For both sequencing approaches, training sets were selected from Silva version 138 (Gurevich *et al.*, 2013; Yilmaz *et al.*, 2014; Glöckner *et al.*, 2017; Perez-Mon, Frey and Frossard, 2020) and GTDB r202 (Parks *et al.*, 2018, 2020, 2022; Chaumeil *et al.*, 2022).

4.2.5. Statistical analysis and data visualization

All statistical analyses and data visualization were performed in R (R Core Team, 2023). Alpha and beta diversity were calculated using *vegan* v2.5.3 (Oksanen *et al.*, 2018) and plots were created with *ggplot2* v3.4.1 (Wickham *et al*., 2016). Considering that the order of magnitude of the number of reads was different between the two sequencing platforms, and the focus on assessing the effectiveness of each sequencing approach, the main results are presented without rarefaction. However, to discard the effect of different numbers of reads on the direct comparison of sequencing approaches, we added supplementary analyses with rarefaction at 10 000 reads. In either rarefied or non-rarefied data, we removed samples with less than 10 000 reads (the list of samples removed due to low number of reads is in Supplementary Table 11.2). As a result, thirteen seawater samples were considered of high quality under both sequencing approaches and used in downstream analyses (thus, a total of 26 16S rRNA gene sequencing libraries).

To compare the difference in distribution of independent groups, we used the Mann-Whitney U test (non-parametric test) and calculated the effect size of each test, to verify if the sample size was enough to support the statistical test. Linear least squares model was used to plot tendency lines comparing depth against alpha diversity metrics, accompanied by their corresponding R-squared values. Finally, we used the PERMANOVA test to support the beta diversity plots. Alpha was set to 0.05 on all statistical tests.

Finally, we were careful to plot the points themselves whenever there were less than five observations per variable and boxplots with median, interquartile ranges, and outliers for those cases with more than 5 observations. The sampling size was small for comparing environmental variables, but it was not small for comparing sequencing approaches, because there were 13 high quality samples for each independent variable. Such considerations were considered in the Results presentation.

The map on Figure 4.1 was partially made by the *ggOceanMaps* v2.2.0 R package (Vihtakari, 2020), with bathymetry from NOAA National Centers for Environmental Information (2022). ArcGIS was used for the context map.

4.3. Results

4.3.1 Overview of sequencing results

Our results enable a direct comparison between second- and third-generation sequencing technologies in the description of seawater prokaryotic communities by means of 16S rRNA gene sequencing. The number of raw CCS reads sequenced by PacBio varied between 20 119 and 53 770 reads (median = 34 182 reads, IQR = 18 610 reads, n = 13), the final number of high-quality reads varied between 10 221 and 29 839 (median = 16 779 reads, IQR = 12 987 reads, n = 13). For Illumina sequencing, the number of raw reads varied between 14 874 and 176 856 (median = 67 998 reads, IQR = 34 824 reads, n = 13), the final number of high-quality reads varied between 12 190 and 134 744 (median = 53 783 reads, IQR = 29 845 reads, n = 13). Summary statistics are available in Supplementary Table 11.13 and rarefaction curves are available in Supplementary Figure 10.6.

4.3.2. Effectiveness of taxonomic classification

To test the effectiveness of taxonomic classification, we considered the proportion of ASVs that were classified at each taxonomic level, for each combination of short- and full-length 16S rRNA gene sequencing and taxonomic database (Silva or GTDB). The combination leading to a higher proportion of classified ASVs down to order level was full-length 16S rRNA gene combined with Silva database, but from family to species level, the best database was GTDB (Figure 4.2, Supplementary Figure 10.7). In fact, from 59.60% up to 73.71% of ASVs obtained from the full-length sequencing approach were classified at species level with GTDB. In contrast, less than 10% of the ASVs obtained from the short reads sequencing approach were classified down to species level, independently of database selection (Figure 4.2, Supplementary Figure 10.7). Differences in the proportion of classified ASVs between sequencing approaches were significant at all taxonomic levels (Mann-Whitney U test, $p < 0.05$), with large effect size ($r > 0.5$) (Supplementary Table 11.14), except for kingdom level with both databases and species level with the Silva database (test failed, Supplementary Table 11.14). Similar results were obtained for the rarefied data, except when taxonomy at the phylum level was employed in combination with the Silva database (Supplementary Table 11.14).

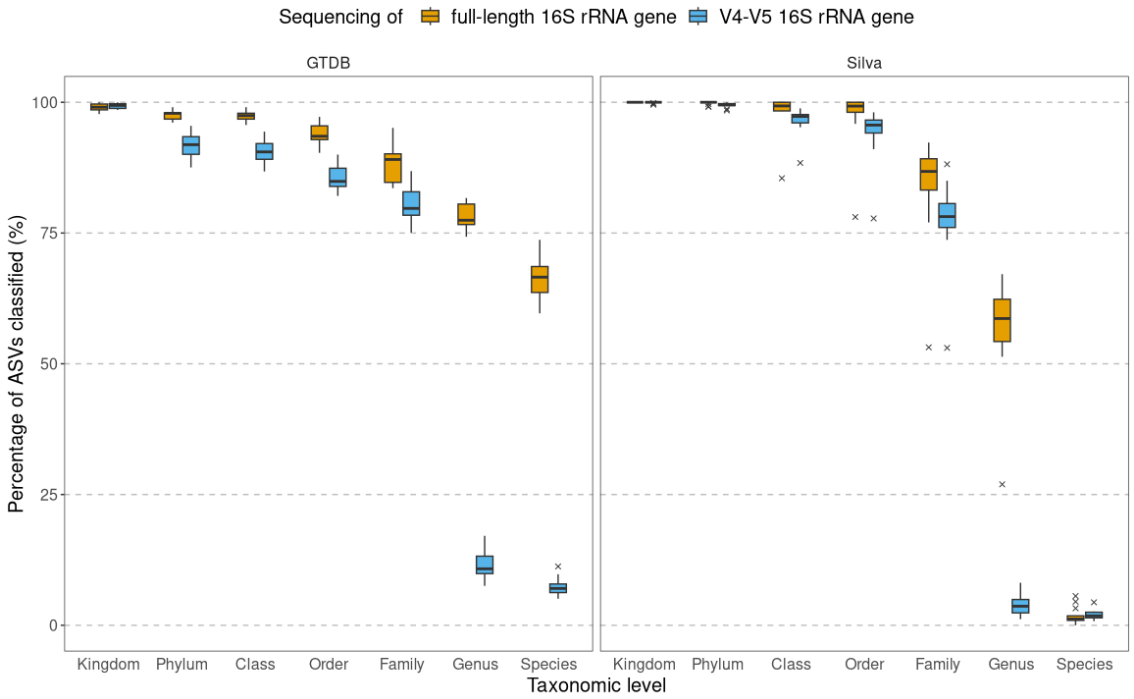


Figure 4.2. The percentage of ASVs classified at each taxonomic level. The left panel shows results for the GTDB database, and the right panel shows results for the Silva database. For each database, full-length and V4V5 16S rRNA gene sequencing were compared (orange and blue, respectively). Boxplots were used to illustrate centrality metrics of the percentage of ASVs classified, including crosses representing the outliers and were calculated with 13 observations, per independent group. Note that the "Kingdom" taxonomic level was maintained to keep consistency with the terminology used by the reference databases used.

Abundant ASVs (relative abundance > 0.1%, per sample) classified with GTDB had similar relative abundances between short- and full-length 16S rRNA gene sequencing (Figure 4.3), and 92.2% of abundant ASVs could only be classified down to species level with the combination of full-length 16S rRNA gene and GTDB database, but could not be captured otherwise, even when using full-length 16S rRNA gene and Silva database (Figure 4.3). Furthermore, some ASVs were classified at species level with either database, but for a single sequencing approach (short- or full-length). For example, *Polaribacter temperatairgensii* and *Sulfitobacter dubius* were identified with short- and full-length approaches, but only with the Silva database (Figure 4.3); while MAGs (represented by GTDB placeholders) were identified in both sequencing approaches, but only with GTDB (Figure 4.3). Generally, the databases identified different species (Figure 4.3).

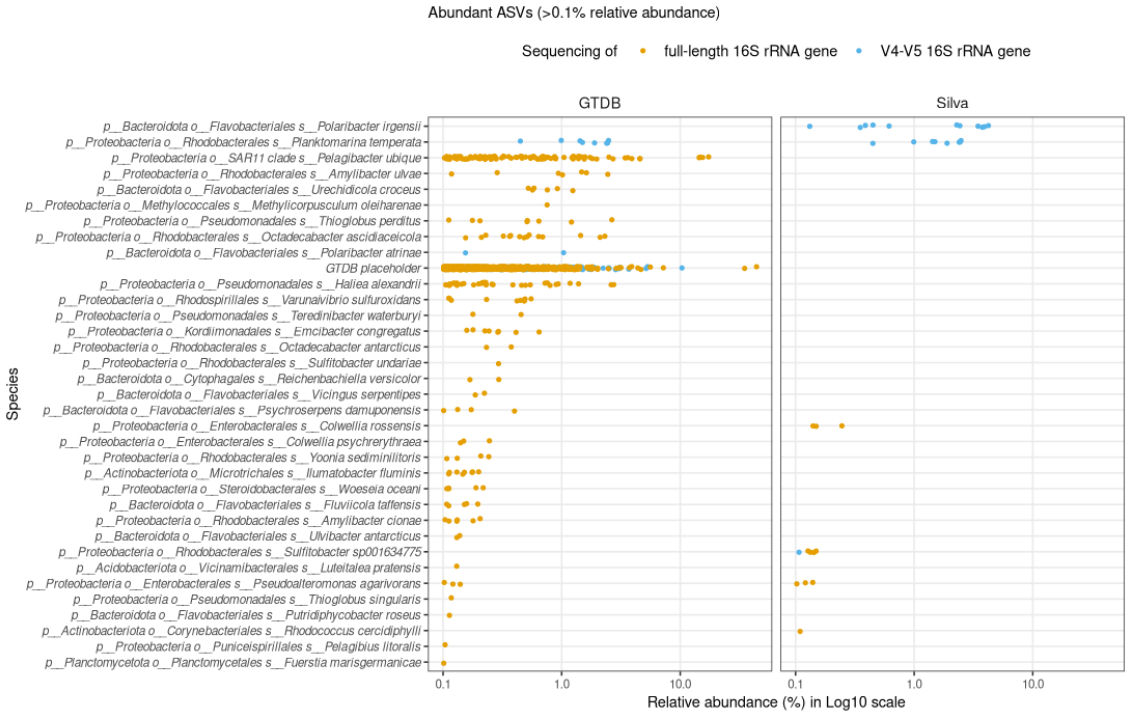


Figure 4.3. Relative abundance of ASVs classified at species level. The left panel shows results for the GTDB database, and the right panel shows results for the Silva database. For each database, full-length and V4V5 16S rRNA gene sequencing were compared (orange and blue, respectively). Only ASVs with more than 0.1% relative abundance, per sample, were illustrated. The Log10 scale was used to allow the distinction between ASVs with less than 1% relative abundance. All species derived from MAGs in the GTDB classification, but without a binomial species name, were grouped in the category “GTDB placeholder”. Phylum, order, and genus level information is provided for each species listed.

4.3.3. Alpha and beta diversity

To verify the impact of 16S rRNA gene read length and taxonomy database choice on estimating prokaryotic diversity metrics, we compared the alpha and beta diversity measures obtained at different water column depths along the transect from inner Kongsfjorden across the West Spitsbergen shelf and out to Fram Strait. These analyses were performed at phylum, order, genus, and species levels. At each taxonomic level, we considered all ASVs that obtained a valid taxonomic classification, *i.e.*, removed the ASVs with no classification. Additionally, at species level, we only used ASVs that could be assigned a species taxonomy in each sequence-database combination, *i.e.*, removed the ASVs with no classification. Unless stated otherwise, when we refer to a specific taxonomic level, we only refer to the sub-set of assigned ASVs at the specified taxonomic level.

Alpha diversity (ASV richness, Shannon index, and Simpson index) was more consistent between sequence-database combinations at phylum than at species level (Figure 4.4, Supplementary Figure 10.8). More specifically, at phylum level, ASV richness increased, Shannon index remained constant, and Simpson's index decreased along depth,

independent of sequence-database combination (Figure 4.4A, Supplementary Figure 10.8). At species level, there were no contradictory trends in alpha diversity between short- and full-length approaches, except for Simpson's index (Figure 4.4B, Supplementary Figure 10.8). Even though trends between water column depths were similar, at species level the full-length 16S rRNA gene combined with GTDB presented higher alpha diversity measures than all other combinations (Figure 4.4B, Supplementary Figure 10.8). At intermediate taxonomic levels, alpha diversity metrics were more consistent for order than genus level (Supplementary Figure 10.9 and 10.10). Generally, the trends were not significant, with few exceptions. For example, the linear regressions were significant for the phylum, order, and species ASV richness with full-length 16S rRNA gene sequencing, with either database ($p < 0.05$, $0.49 < r^2 < 0.56$, Supplementary Table 11.16). Generally, there was not enough support for a conclusive comparison of alpha diversity against depth for most combinations. However, the linear regressions were more similar with each other at phylum than species level, in accordance with Figures 4.4 and Supplementary Figure 10.8.

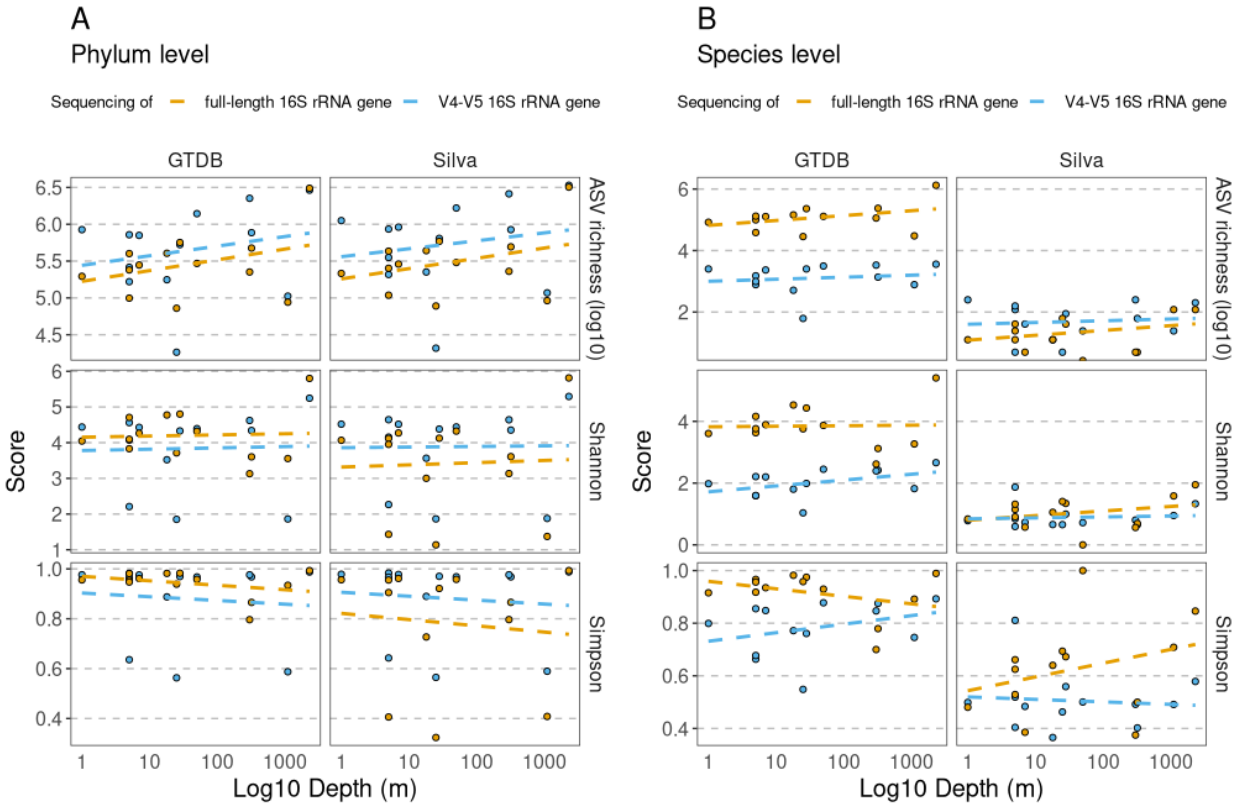


Figure 4.4. Alpha diversity scores across depth. ASVs were filtered (A) at phylum level and (B) at species level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help reading the figure. Regression equations and statistical support are available in Supplementary Table 11.15.

The combination of full-length 16S rRNA gene and GTDB obtained higher ASV richness than any other combination when accounting for the different sampling stations (Figure 4.5A, Supplementary Figure 10.11). The higher ASV richness at species level with full-length 16S rRNA gene sequencing and GTDB was most probably a consequence of the increased classification of low abundance ASVs, *i.e.*, the rare biosphere (relative abundance < 0.1% per sample), into formally valid species (Figure 4.5B, Supplementary Figure 10.11). None of the ASVs identified with full-length 16S rRNA sequencing and classified at species level by the Silva database were rare (Figure 4.5B, Supplementary Figure 10.11). Stations displayed very low ASV richness for that combination (full-length 16S rRNA gene sequencing and Silva database) (generally below 10 ASVs, Figure 4.5A, and Supplementary Figure 10.11). For the ASVs classified at phylum level, in contrast, ASV richness was high (> 100 ASVs) for all stations and rare taxa were found across all combinations tested (Figure 4.5 and Supplementary Figure 10.11). Finally, very similar ASV richness scores were obtained for phylum and order taxonomic levels, while such scores decreased for genus and species taxonomic levels (Figure 4.5 and Supplementary Figure 10.11).

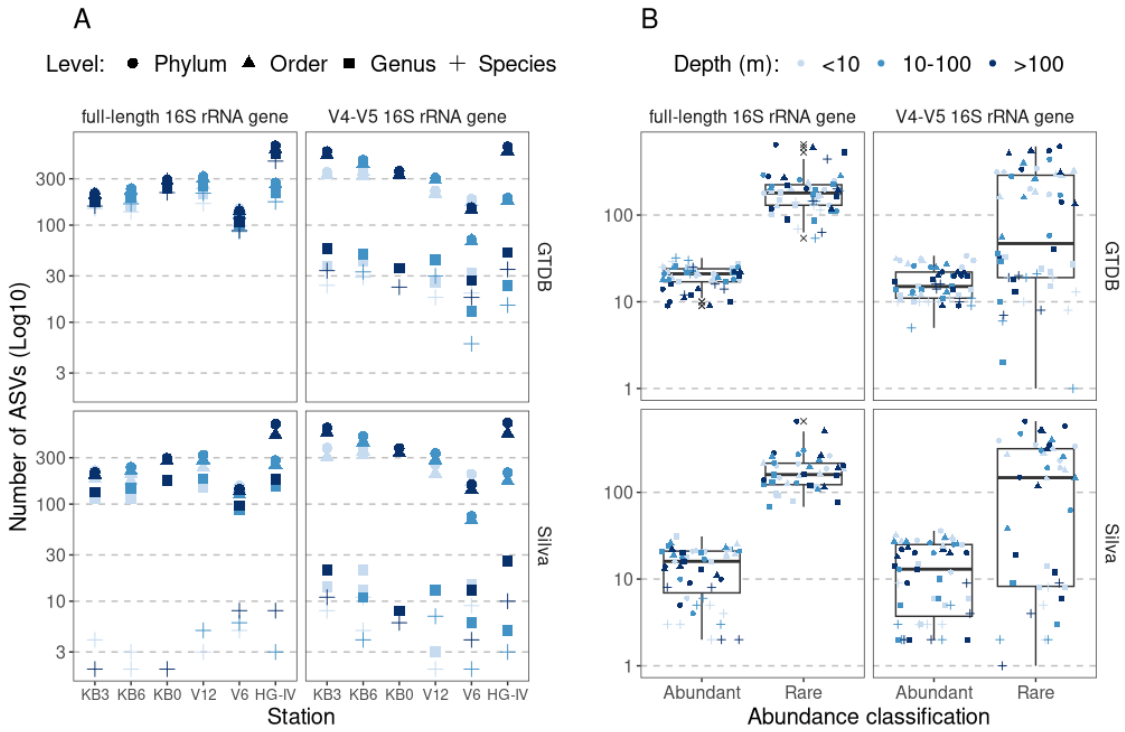


Figure 4.5. ASV richness at phylum, order, genus and species taxonomic levels across methodological combinations. **A**. ASV richness by sampling Station. **B**. ASV richness by abundance classification, with rare ASVs defined as ASVs with less than 0.1% relative abundance, per sample. For both (A) and (B), the taxonomic level is represented by different symbols. The ASV richness is illustrated in Log10 scale so that lower values could be distinguished; each panel further

divided itself into four panes, allowing direct comparison of full-length and V4V5 16S rRNA gene sequencing (columns) and Silva or GTDB databases (rows). In (A) the points were illustrated without centrality metrics, because each station included less than five distinct observations, while in (B) a centrality metric was used with boxplots, because the number of observations (n = 13 per independent group) allowed it. Nevertheless, the original points were plotted on top of the boxplots, so that it was possible to see the direct connection between (A) and (B). Note that some combinations presented less than 13 observations, which reflected samples without any species classified. The points were colored in different shades of blue, divided by depth: < 10 m; 10-100 m; and > 100 m.

Notably, more than one ASV was obtained for 27 species (with binomial name), *e.g.*, 39 different ASVs were attributed to *Pelagibacter ubique* (RS_GCF_000012345_1) using full-length 16S rRNA gene sequencing and GTDB (for a full list, see Supplementary Table 11.16). We verified that *Pelagibacter ubique* (RS_GCF_000012345_1) contained a single 16S rRNA gene copy (Supplementary Table 11.16), which indicates that the multiple ASVs obtained for this species provide information on subspecies-level diversity. Another relevant example was *Colwellia psychrerythraea*, because its GTDB entry describes a specific strain (34H) and it includes up to 9 copies of the 16S rRNA gene (Supplementary Table 11.16).

As for the beta diversity component, the results obtained by either database were very similar at phylum level, as shown by the centroids proximity (Figure 6A and B), but community composition of taxonomically classified ASVs differed at species level (Figure 4.6C and 4.6D). The community composition was the most distinct between databases for the species level ASVs identified with full-length 16S rRNA gene sequencing (Figure 4.6D). In fact, at phylum level community composition was not significantly different (p > 0.05, PERMANOVA test Supplementary Table 11.17), but was so at species level (p < 0.05, PERMANOVA test, Supplementary Table 11.17). For order and genus level, the beta diversity overlapped between databases (Supplementary Figure 10.12, p > 0.05, PERMANOVA test Supplementary Table 11.17), except for genus level, with V4V5 16S rRNA gene sequencing (Supplementary Figure 10.12, p < 0.05, PERMANOVA test, Supplementary Table 11.17).

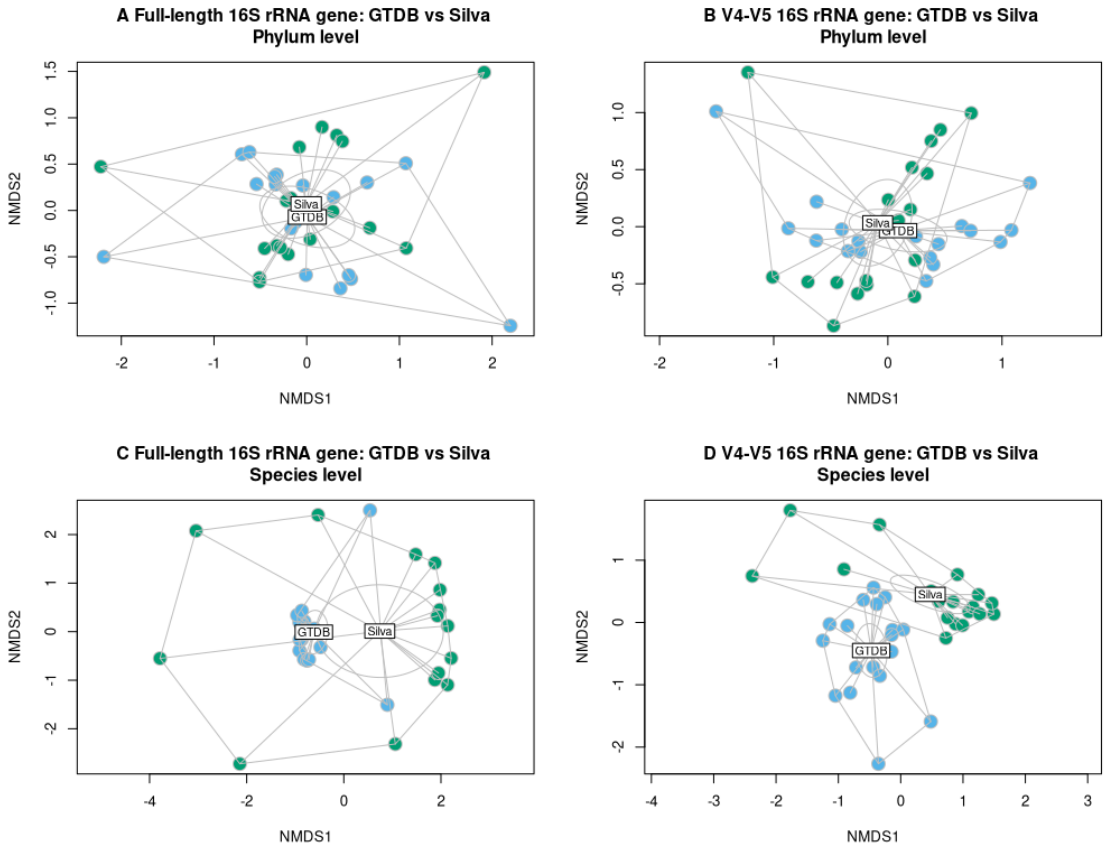


Figure 4.6. Beta diversity at phylum and species level. Community composition between Silva and GTDB databases for (A) full-length 16S rRNA gene, at phylum level; (B) V4V5 region of the 16S rRNA gene, at phylum level; (C) full-length 16S rRNA gene, at species level; (D) V4V5 region of the 16S rRNA gene, at species level. Bray-Curtis distances were illustrated with nMDS ordination. Shapes and color were used to illustrate sampling stations and depth.

4.4. Discussion

Species-level assignment of prokaryotes can improve insights gained in microbial ecology studies, but it is hard to achieve with the current, most used gene sequencing strategies (Earl *et al.*, 2018; Brede *et al.*, 2020). One possible way to improve the capacity for species level assignments is to use high-throughput sequencing of the full-length 16S rRNA gene (Johnson *et al.*, 2019), an approach validated in laboratory conditions (Callahan *et al.*, 2019) and tested for several natural environments (Klemetsen, Willassen and Karlsen, 2019; Pootakham *et al.*, 2019, 2021; Wang *et al.*, 2022; Yan *et al.*, 2023).

If we consider classification effectiveness as a function of the proportion of reads classified, then our results suggest that the best approach to improve the effectiveness of taxonomic assignments at lower taxonomic ranks is the combination of full-length 16S rRNA gene sequencing by CCS PacBio with the GTDB database. We note that the database choice was fundamental, because with Silva the number of ASVs classified at genus and species level was much lower than with GTDB (less than 10% of ASVs were classified down to species level, using Silva database). Our results highlight the relevance of database selection in delivering appropriate taxonomy profiling of natural microbial communities. Furthermore, our results are consistent with previous findings addressing the impact of sequencing approaches and/or database selection on the description of microbiomes from several environmental/host matrices (Myer *et al.*, 2016; Brede *et al.*, 2020; F Escapa *et al.*, 2020; Costa *et al.*, 2022; Overgaard *et al.*, 2022; Yu *et al.*, 2022). For instance, a few full-length 16S rRNA gene sequences were classified at species level with Silva in an experiment using an *in vitro* model of subacute rumen acidosis (Brede *et al.*, 2020). Also, the quantity of ASVs classified at species level was previously found to be less dependent on the sequencing approach (LoopSeq, PacBio and Illumina) than database selection (in this case, RDP or Silva) in a study on the effects of waterlogging on the rhizosphere microbiome (Yu *et al.*, 2022). Other studies showed that habitat-specific databases can foster the highest classification accuracy at species level (Myer *et al.*, 2016; F Escapa *et al.*, 2020; Overgaard *et al.*, 2022; Silva *et al.*, 2022). It is important to note that most databases were built using sequences shorter than current long-read amplicons and do not account for within species variation (Tedersoo *et al.*, 2021), but our full-length 16S rRNA gene results did not seem to be particularly affected by this.

Besides the database and sequencing technology selection, the software used for the processing of raw reads can have a significant impact on data interpretation. In this study, we used *DADA2* (Callahan *et al.*, 2016), which has been shown to deliver high quality results for both V4V5 16S rRNA gene (Callahan *et al.*, 2016) and full-length 16S rRNA gene (Callahan *et al.*, 2019) datasets. For an extensive review on bioinformatics software for raw read processing see Hakimzadeh *et al*. (2024).

Besides sequencing the full-length 16S rRNA gene, the operon 16S-ITS-23S could be used instead (Seol *et al.*, 2022) and other studies have provided insights into the usefulness of alternative genetic markers for the taxonomic assignment of prokaryotes. For example, the RNA polymerase subunit B gene (*rpoB*) led to higher accuracy than the V3V4 region of the 16S rRNA gene in mock communities and to a different microbiome description of entomopathogenic nematodes (Ogier *et al.*, 2019). Advances in shotgun metagenomics, metagenome assembled genomes, and culturomics, together with the expansion of reference databases, are key for a long-term improvement of species-level classification of prokaryotes from natural environments.

Some studies have suggested that combining second and third generation sequencing can improve the description of prokaryotic communities (Brede *et al.*, 2020). However, in this study, Illumina's high-throughput did not add explanatory power to the data obtained by PacBio CCS, because: (1) at phylum level, all results were very similar between short- and full-length 16S rRNA gene sequencing; (2) for lower taxonomic levels (particularly, species level), short 16S rRNA gene sequencing is less accurate than the full-length alternative (Callahan *et al.*, 2019); and (3) the ability to assign species-level taxonomy to ASVs belonging to the rare biosphere was best with full-length 16S rRNA gene sequencing and GTDB database. Furthermore, using two sequencing approaches in parallel might be prohibitively expensive. For an extensive review on sequencing costs per strategy, see Tedersoo *et al*. (2021).

In terms of ecological analysis, all combinations tested were equivalent at phylum and order level for alpha and beta diversity. In fact, at this level the values of alpha diversity were similar independently of using either full-length or V4V5 region of the 16S rRNA gene; and of using either GTDB or Silva database. At the species level, there were several differences between the combinations tested. Specifically, more ASVs were assigned a species-level taxonomy using full-length 16S rRNA gene sequencing when compared to sequencing of the V4V5 region of the 16S rRNA gene, and this effect was more pronounced whenever the GTDB database was used. In fact, the analysis of community composition further supports the importance of database selection, because

species-level community composition was grouped by database, instead of environmental variables, when using the V4V5 region of the16S rRNA gene.

It is important to note that the combination of full-length 16S rRNA gene and GTDB database was able to assign species-level taxonomy of distinct, yet closely related ASVs to the same species. We highlighted *Pelagibacter ubique*, because it was the species with the most ASVs in our dataset and we verified that the reference genome included a single 16S rRNA gene copy. Besides *Pelagibacter ubique*, we highlighted *Colwellia psychrerythraea*, because its GTDB reference sequence belongs to a formally classified strain (34H) with up to nine 16S rRNA gene copies. Both examples provided, *Pelagibacter ubique* and *Colwellia psychrerythraea*, are consistent with previous findings, that established the possibility of accurate identification of strain level diversity (Callahan *et al.*, 2019). We are aware that our analysis is not sufficient to be certain of subspecies or even strain identification, but it is indicative. Future studies using full-length 16S rRNA gene sequencing and ASVs should take into consideration the possibility of diversity overestimation, because of unaccounted subspecies-level diversity (for example, if one and each ASV is counted as a potential representative of one species, and the possibility of heterogeneous ASVs to belong to the same “species” is discarded).

A previous study compared the V3V4 and the V4V5 regions of the 16S rRNA gene to help the design of long-term monitoring campaigns of the Arctic Ocean (Fadeev *et al.*, 2021), finding that the V4V5 region was better than the V3V4 region for that purpose. In here, we argue that full-length 16S rRNA gene sequencing in combination with the GTDB database will further improve such efforts, because of the improved ability to assign species-level taxonomy and, eventually, subspecies level diversity, as well as to obtain information on the reference genomes and/or MAG representatives. This genomic information might be used to gather putative functional information. However, other methods might be more adequate for functional profiles (*e.g.*, metagenomics and metatranscriptomics), if there is enough budget. Future studies will need to address to what extent it is possible, or not, to combine datasets obtained from specific regions of the 16S rRNA gene and datasets obtained from the full-length 16S rRNA gene, so that ASVs can be compared across a wide range of data generated from long-term monitoring programs. For example, it is worth exploring if it is possible to extract short-reads from long-reads and obtain equivalent information. It is also worth exploring if specific taxonomic databases are biased towards specific sequencing approaches. These issues are relevant, because microbial ecologists need to balance methodological

consistency vs methodological updates over decade-long efforts to obtain standardized diversity information from natural environments.

4.5. Conclusions

In conclusion, full-length 16S rRNA gene sequencing improves the description of prokaryotic microbiomes from natural environments, when compared with short region sequencing (V4V5 in here). According to previous findings (Myer *et al.*, 2016; F Escapa *et al.*, 2020; Costa *et al.*, 2022; Overgaard *et al.*, 2022), databases optimized for specific habitats provide the best classification at species level, but they are unavailable for underexplored environments such as the Arctic Ocean. Thus, for this and similar natural environments, researchers depend on universal taxonomic databases. We identified no advantage besides higher throughput in the sequencing of short regions of the 16S rRNA gene, even for the description of the microbial rare biosphere, depending on the database. Overall, the higher resolution of full-length 16S rRNA gene combined with the use of GTDB resulted in the species-level taxonomy assignment of several ASVs with genome reference information, which may improve inter-comparison of ASVs across studies, thereby providing more context for microbial ecologists in the interpretation of natural microbial communities.

4.6. Declarations

Author contributions

Conceptualization (FP, RC and CM); data curation (FP and CM); formal analysis (FP); MOSJ 2019 sampling campaign (CM, PD, PA); funding acquisition (CM, RC and PA); methodology (FP); software (FP); visualization – all R plots (FP); visualization – map (FP and PD); writing – original draft (FP); writing – review and editing (all authors).

Competing interests

The authors declare no conflict of interest.

Funding

FCT funded this study through two grants to CM (PTDC/CTA-AMB/30997/2017; 2022.02983.PTDC) and through a PhD grant to FP (2020.04453). This research has been also supported by the Norwegian Polar Institute and the Research Council of Norway (project no. 244646) and by national funds by FCT through the projects UIDB/04423/2020, UIDP/04423/2020, UIDB/04565/2020, UIDP/04565/2020 and LA/P/0140/2020.

Availability of data and materials

The FASTQ files of the MOSJ 2019 project are publicly available in the European Nucleotide Archive (Project: PRJEB60815). All R code used is available in GitHub (https://github.com/pascoalf/Full_vs_V4V4_16S_Annals_of_Microbiology).

Chapter 5. Exploration of the types of rarity in the Arctic Ocean from the perspective of multiple methodologies.

This chapter used materials previously published in:

Pascoal, F. *et al.* (2022a) “Exploration of the types of rarity in the Arctic Ocean from the perspective of multiple methodologies,” *Microbial Ecology*, 84(1), pp. 59–72. Available at: https://doi.org/10.1007/s00248-021-01821-9.

See author contributions in section **5.6 Declarations**.

5. Exploration of the types of rarity in the Arctic Ocean from the perspective of multiple methodologies

Francisco Pascoal[1,2], Rodrigo Costa[3,4], Philipp Assmy[5], Pedro Duarte[5] and Catarina Magalhães[1,2]

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal; [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal; [3]Department of Bioengineering, Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico, University of Lisbon, Portugal; [4]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, University of Lisbon, Portugal; [5]Norwegian Polar Institute, Fram Centre, N-9296 Tromsø, Norway.

Abstract

The Arctic Ocean is facing rapid environmental changes with cascading effects on the entire Arctic marine ecosystem. However, we have a limited understanding of the consequences such changes have on bacteria and archaea (prokaryotes) at the base of the marine food web. In this study, we show how the prokaryotic rare biosphere behaves over a range of highly heterogeneous environmental conditions using 16S rRNA gene reads from amplicon and metagenome sequencing data from seawater samples collected during the Norwegian Young Sea Ice Expedition (N-ICE) between late winter and early summer. The prokaryotic rare biosphere was analyzed using different approaches: ASVs and OTUs from the 16S rRNA gene amplicons and operational taxonomic units from the 16S rRNA genes of the metagenomes. We found that prokaryotic rare biosphere communities are specific to certain water masses, and that most of the rare taxa identified were always rare and disappeared in at least one sample under changing conditions, suggesting their high sensitivity to environmental heterogeneity. In addition, our methodological comparison revealed a good performance of 16S rRNA gene amplicon sequencing in describing rare biosphere patterns, while the metagenome-derived data were better to capture a significant diversity of so-far uncultivated rare taxa. Our analysis on the dynamics of the rare prokaryotic biosphere, by combining different methodological approaches, improves the description of the types of rarity predicted from community assembly theory in the Arctic Ocean.

Keywords: Types of rarity, ASVs vs. OTUs, Arctic Ocean microbiome, marine microbial ecology, polar microbial ecology, molecular ecology.

5.1. Introduction

The Arctic Ocean represents an extreme environment, influenced by the strong seasonality of polar ecosystems, marked by the transition from 24 hours of darkness to midnight sun and melting of sea ice in the summer (Meyer *et al.*, 2017). Internally the Arctic Ocean is characterized by complex circulation patterns of distinct water masses that shape the structure of Arctic prokaryotic communities (Müller *et al.*, 2018), because each water mass creates one singular environment with strong selective pressures towards specific ecotypes (Müller *et al.*, 2018). It has been shown that deeper layers of the Arctic Ocean display higher microbial biodiversity than surface waters (Wilson *et al.*, 2017; de Sousa *et al.*, 2019) which was attributed to the different environmental characteristics of the respective water masses (Galand *et al.*, 2009; Kirchman, Cottrell and Lovejoy, 2010; Ghiglione *et al.*, 2012; Müller *et al.*, 2018). Thus, by studying spatial and temporal variations of Arctic microbial communities, it appears that those in deeper waters (below 200m) are more diverse and shaped by more homogeneous selective pressures than the ones in surface waters, which are subjected to large seasonality in environmental drivers such as light availability and temperature (Ghiglione *et al.*, 2012).

The prokaryotic rare biosphere, that is, the low abundance prokaryotes, has only recently been addressed in this environment. Since the first description of the marine prokaryotic rare biosphere (Sogin *et al.*, 2006), it was hypothesized to be cosmopolitan and to serve as a seed bank for organismal recruitment and enrichment across ecosystems, changing in abundance as a response to changing environmental conditions (Pedrós-Alió, 2006, 2007). Initially, it was hypothesized that the Arctic Ocean would be characterized by cosmopolitan taxa that can oscillate between rare and abundant (conditionally rare taxa - CRT), but previous studies have shown that the Arctic Ocean rare biosphere is prevalently always rare (permanently rare taxa - PRT) across biogeographic (Galand *et al.*, 2009) and seasonal patterns (Kirchman, Cottrell and Lovejoy, 2010). Lately, conditional rarity has been found in several environments (Shade *et al.*, 2014), including the Arctic Ocean, but with less prevalence than permanently rare taxa (Kirchman, Cottrell and Lovejoy, 2010). For example, some archaea of the phylum *Thaumarchaeota*, which are important chemolithoautotrophs in the Arctic Ocean (Wilson *et al.*, 2017), have been identified as conditionally rare in the winter to spring transition (de Sousa *et al.*, 2019). It has also been shown that seasonal variation in the structure of Arctic prokaryotic

communities can be associated with that of microeukaryotic communities. For example, the emergence of the *Bacteroidetes* order *Flavobacteriales* was linked with the bloom of the microalga *Phaeocystis pouchetii*, during the transition from late winter to early summer (de Sousa *et al.*, 2019).

Besides permanent and conditional rarity, transiently rare taxa (TRT) can also be considered as a further mode of rarity, whereby rare species disappear from the sampled environment (Jia, Dini-Andreote and Falcão Salles, 2018). Each type of rarity should be considered as contextual, meaning that it only applies to describe how a rare microbe behaves across time and space or to changes in tested environmental variables (Pascoal, Costa and Magalhães, 2021) (Chapter 2). It does not imply, for instance, that one specific permanently rare taxon cannot grow under different environmental conditions (Pascoal, Costa and Magalhães, 2021) (Chapter 2). These different types of rarity are explained by different stochastic and deterministic ecological mechanisms (Jia, Dini-Andreote and Falcão Salles, 2018) in the framework of community assembly theory (Vellend, 2010; Zhou and Ning, 2017). According to Jia *et al.*, (2018), stochastic mechanisms of drift and dispersal limitation are responsible for transient rarity, whereas deterministic mechanisms, equivalent to selective pressures, are responsible for permanent and conditional rarity. Permanent rarity is disclosed when the selective pressure determining the growth of the specific taxon is maintained constant, while conditional rarity may be more often observed when the selective pressure varies. Thus, stochastic mechanisms assume that random events of cells dispersal will be responsible for the emergence/disappearance of taxa in the sampled sites, whereas deterministic mechanisms assume that prokaryotes respond to abiotic/biotic pressures (Zhou and Ning, 2017; Jia, Dini-Andreote and Falcão Salles, 2018).

The low abundance populations that exist in the environment represent a genomic reservoir, which has a role in the resistance and resilience of the ecosystem (Jousset *et al.*, 2017). Thus, the rare biosphere might have an important function in the Arctic Ocean. Recently, some studies have addressed this issue, for example, the rare biosphere response to climate change in the central Arctic Ocean (Fernández-Gómez *et al.*, 2019). Despite several studies on the ecology of the rare biosphere in the Arctic region (Galand *et al.*, 2009; Kirchman, Cottrell and Lovejoy, 2010; Baltar *et al.*, 2016; de Sousa *et al.*, 2019; Fernández-Gómez *et al.*, 2019; Gokul *et al.*, 2019), a perspective on the types of rarity derived from community assembly theory in this environment is lacking.

In this study, we explore distinct methodologies to identify and describe the types of prokaryotic rarity in the Arctic Ocean. To this end, we used the N-ICE 2015 expedition dataset (Granskog *et al.*, 2018; de Sousa *et al.*, 2019), which derives from sampling events that took place in the Arctic Ocean, north of Svalbard, between January and June 2015. In this study, we used V4V5 16S rRNA gene amplicon sequencing reads processed into ASVs and OTUs, as well as OTUs derived from shotgun-sequenced metagenomes (mOTU), a multi-phasic approach to data processing, to describe the rare prokaryotic communities of different water masses in the Arctic Ocean. The reason for the utilization of ASVs and mOTUs in addition to the traditional OTUs, is that the latter were not used before to explore the rare biosphere in the Arctic Ocean. Thus, the question of whether these approaches provide novel insights into our perspective of the rare biosphere can be answered in this study.

5.2. Methods

5.2.1. Environmental context

During N-ICE 2015, research vessel Lance was drifting with sea ice north of Svalbard in the southern Nansen Basin and the Yermak Plateau with a focus on atmosphere-snow-ice-ocean-ecosystem interactions in a thinning sea ice regime (Granskog *et al.*, 2018). The N-ICE sampling points correspond to three different regions, sampled in March, April, and June, along the drift of research vessel Lance (Granskog *et al.*, 2018). At each sampling region, 20 L of seawater (± 3 L) were collected under the ice at three different depths (5 m, 20 m or 50 m and 250 m), and microbial communities concentrated (for details, see de Sousa *et al.*, (2019)). These sampling points correspond to different water masses (Meyer *et al.*, 2017): Polar surface waters (PSW), at 5 m and 50 m; warm polar surface water (PSWw), at 20 m in June; modified Atlantic water (MAW) at 250 m from March to April; Atlantic water (AW) at 250 m, in June. The sampling points' characteristics are summarized in Table 5.1. Importantly, water samples collected in March and April were covered by thick ice and snow and thus were characterized by low light availability. Conversely, samples collected in June were underneath thinner ice and melting snow, possessing higher light availability and being closer to the inflow of AW.

Table 5.1. Spatial and temporal characteristics of the nine sampling points, according to de Sousa *et al.*, (2019) and using the same sample IDs. This table also includes the run accession numbers for the raw sequence samples, that can be accessed through the ENA database (see Availability of Data in section 5.6).

Sample ID	Month of sampling	Depth (m)	Ocean region	Water mass	Latitude	Longitude	ENA 16S rRNA sequences run accession	ENA metagenomic sequences run accession
NB_5	March	5	Nansen Basin	PSW	83° 10.002´N	22° 01.998'E	ERR2044662	ERR2017139
NB_50	March	50	Nansen Basin	PSW	83° 10.002´N	22° 01.998'E	ERR2044663	ERR2017140
NB_250	March	250	Nansen Basin	MAW	83° 10.002´N	22° 01.998'E	ERR2044664	ERR2017141
TR_5	April	5	Transition Region	PSW	82° 23.195´N	15° 9.198´E	ERR2044665	ERR2017142
TR_50	April	50	Transition Region	PSW	82° 23.195´N	15° 9.198´E	ERR2044666	ERR2017143
TR_250	April	250	Transition Region	MAW	82° 23.195´N	15° 9.198´E	ERR2044667	ERR2017144
YP_5	June	5	Yermak Plateau	PSW	80° 30.775´N	07° 52.428´E	ERR2044668	ERR2017145
YP_20	June	20	Yermak Plateau	PSWw	80° 30.775´N	07° 52.428´E	ERR2044669	ERR2017146
YP_250	June	250	Yermak Plateau	AW	80° 30.775´N	07° 52.428´E	ERR2044670	ERR2017147

5.2.2. Seawater sampling and processing for microbiome analysis

Seawater for physical and chemical variables was collected with the rosettes of the ship-based and the on-ice conductivity, temperature and depth devices described in Meyer *et al*. (2017), through a hole in the ice. Microbial collection and processing prior to sequencing of both V4V5 16S rRNA gene amplicons and metagenomes followed the methodology described by the Ocean Sampling Day Consortium (Kopf *et al.*, 2015) further described in detail for the N-ICE 2015 expedition by de Sousa *et al.* (2019). Briefly, samples for genetic sequencing were collected using a GoFlow bottle (20 L) and then filtered through a cartridge filter (Sterivex) with a 0.22 µm pore size, hydrophilic, PVDF, Durapore membrane (SVGV010RS, Merck Millipore, Portugal) with the help of a peristaltic pump. The filters were sealed and stored at −80 °C until further processing. DNA extraction was described in detail in de Sousa *et al.* (2019) and was performed using the PowerWater DNA Isolation Kit protocol (MO BIO Laboratories, Inc.). Amplicon sequencing of 16S rRNA genes used the primer set 515F-Y/926R (Caporaso et al., 2011; 2012; Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016), which encompasses the V4V5 hypervariable regions. For both 16S rRNA genes and total metagenomes, DNA sequencing was performed on an Illumina MiSeq platform using 2x 300bp, V3 chemistry (Illumina). The library preparation for V4V5 16S rRNA gene sequencing and total community DNA sequencing, as well as DNA quality assessment, are described in de Sousa *et al.* (2019).

4.2.3. Bioinformatics processing of 16S rRNA gene and total community DNA sequences

The 16S rRNA gene amplicons generated in the framework of the N-ICE 2015 project were processed in this study for the assignment of ASVs with *DADA2* R package v1.15.0 (Callahan *et al.*, 2016), using default parameters, after primers removal. Considering the quality profiles for forward and reverse reads (Supplementary Figures 10.13 and 10.14) the trimming at position of 196nt for both forward and reverse reads was selected. By default, *DADA2* only accepts merged sequences with at least a 12nt match, thus the maximum merged read size was 380bp (196 bp x 2). The same raw sequences used to produce ASVs were processed by the MGnify platform version 5.0 (Mitchell *et al.*, 2019) to generate a list of hierarchical OTUs (at least, 97% gene similarity threshold). The metagenomic dataset was also processed using the MGnify platform version 5 (Mitchell *et al.*, 2019) resulting in the generation of a table of samples vs. metagenomic derived OTUs (mOTUs – 97% gene similarity threshold) based on 16S rRNA gene reads directly

retrieved from the metagenome sequences. Taxonomy was assigned using the Silva database version 132 (Quast *et al.*, 2012), for ASVs, OTUs and mOTUs.

5.2.4. Alpha and Beta diversity calculation

R was used for alpha and beta diversity analyses (R Core Team, 2023). The boxplots were produced using base R and the bar plots by using *ggplot2* R package v3.2.1 (Wickham, 2016). Rarefaction curves, hierarchical clustering and ordination analysis were performed using *vegan* R package v2.5.6 (Oksanen *et al.*, 2018). The ordination analysis of the structure of rare prokaryotic communities consisted of the unconstrained method of Correspondence Analysis (CA), which uses Chi-square transformed distance matrices and weighted linear mapping. The hierarchical clustering consisted of the unweighted pair group method with arithmetic mean (UPGMA) applied to Bray-Curtis distance matrices.

5.2.5. Rare Biosphere definition and calculation of types of rarity

A phylotype (ASV, OTU or mOTU) is rare in a given sample if its abundance is equal to or lower than the threshold of rarity. We used the algorithm from Gobet *et al.*, (2010), MultiCoLA, to decide specific thresholds of rarity in absolute and relative abundance, per sample, for the ASVs, OTUs and mOTUs datasets (Table 5.2). The selected thresholds were based on the interpretation of the MultiCoLA results (Supplementary Figure 10.15). To discard the possibility that the thresholds obtained were influenced by the differences in sampling size, we checked the threshold values obtained after rarefaction (Supplementary Figure 10.16). Furthermore, because the different datasets (ASV-, OTU- and mOTU-based) have different orders of magnitude, there is the possibility of overestimating the rare biosphere in one set of approaches relative to others. To avoid that problem, we also used the commonly used definition of 0.1% relative abundance (Pedrós-Alió, 2012), per sample, for the non-rarefied and rarefied datasets. Unless stated otherwise, whenever we refer to the rare biosphere, we are referring to the taxa with abundance lower than the MultiCoLA defined threshold, for the non-rarefied data.

Table 5.2. Selected thresholds of rarity obtained from the MultiCoLA algorithm, expressed as absolute values and respective relative abundance range and median values.

Dataset	Absolute abundance threshold (nº of sequences)	Range of threshold values in relative abundance	Median relative abundance threshold
16S rRNA amplicon derived ASVs	1148	0.65% to 3.52%	1.59%
16S rRNA amplicon derived OTUs	660	0.31% to 1.25%	0.92%
Metagenome derived OTUs	79	2.56% to 6.50%	5.18%

To calculate the different types of rarity in each group of samples, we used a set of logical statements: if the taxon is abundant in all samples, then it is named abundant; if it is rare in at least one sample and abundant in at least one sample, it is considered CRT; if the taxon is rare in all samples, then it is considered PRT. Within the PRT classification, if it is absent in at least one sample, it is classified as TRT instead. This classification used custom R commands. The taxonomic assessment of the types of rarity was processed and grouped using R custom commands, and visualized by means of Circos software (Krzywinski *et al.*, 2009), and *ggplot2* (Wickham, 2016).

5.3. Results

5.3.1. Rare prokaryotic alpha and beta diversity

The Arctic Ocean samples analyzed in this study were collected north of Svalbard, capturing the Arctic winter to spring transition (March, April and June) and different depths (5 m, 20 m or 50 m and 250 m). The samples from March and April are characterized by low light penetration due to the presence of ice with a thick snow cover and are divided into polar surface water (PSW) and modified Atlantic water (MAW), while the June samples were characterized by higher light penetration, after the beginning of the melting season, and the presence of Atlantic water (AW) masses (de Sousa *et al.*, 2019).

The rarefaction curves (Supplementary Figure 10.17) show that a higher sequencing effort could improve the diversity captured by the V4V5 16S rRNA gene amplicon and metagenomics derived OTUs. Conversely, the rarefaction curves built with ASV data were always observed to reach a plateau. The main analysis presented here derives from non-rarefied data, since we did not find a statistical improvement that justified the loss of information (McMurdie and Holmes, 2014). Notwithstanding, to discard possible effects of different sampling sizes on specific analyses, we provide results obtained with the rarefied data in the supplementary materials (indicated below).

A total of 799 968 high quality sequences from the V4V5 regions of the 16S rRNA gene were used to produce ASVs. The number of reads attributed to the rare biosphere was 313 117 (39% of all reads). Of a total of 4 253 ASVs, 4 138 were from rare prokaryotes (97% of all ASVs). A total of 871 719 reads from the V4V5 regions of the 16S rRNA gene were used to produce OTUs. The number of reads attributed to the rare biosphere was 96 274 (11% of all sequences). Of a total of 2 195 OTUs, 2 051 were from rare prokaryotes (93% of all OTUs). From the metagenomes, a total of 16 430 reads were used to produce mOTUs. The number of sequences attributed to the rare biosphere was 7 398 (82% of all sequences). Of a total of 1 379 mOTUs, 1 334 were from rare prokaryotes (97% of all mOTUs). These values are summarized in Supplementary Table 11.18, which also includes the values for the alternative rarity definition (0.1%, per sample), from rarefied and non-rarefied data.

The range of values of total reads per sample was similar for both 16S rRNA gene sequencing-based approaches (ASVs ranged from 32 662 to 176 670 reads, with a median value of 72 055 reads and the OTUs ranged from 52 679 to 213 069 reads, with

a median value of 72 115 reads), while the metagenomic sequences varied much less (from 1 216 to 3 083 reads and a median value of 1 525 reads) (Figure 5.1a). Despite that, when filtering for the rare biosphere, the number of rare reads was superior for ASVs and had a higher range of values between samples (from 15 691 to 50 951 reads, with a median value of 33 725 reads), followed by the OTUs (from 4 511 to 13 472 OTU reads and a median value of 11 173 reads) and mOTUs (from 633 to 12 882 reads, with a median value of 784 reads) (Figure 5.1b).

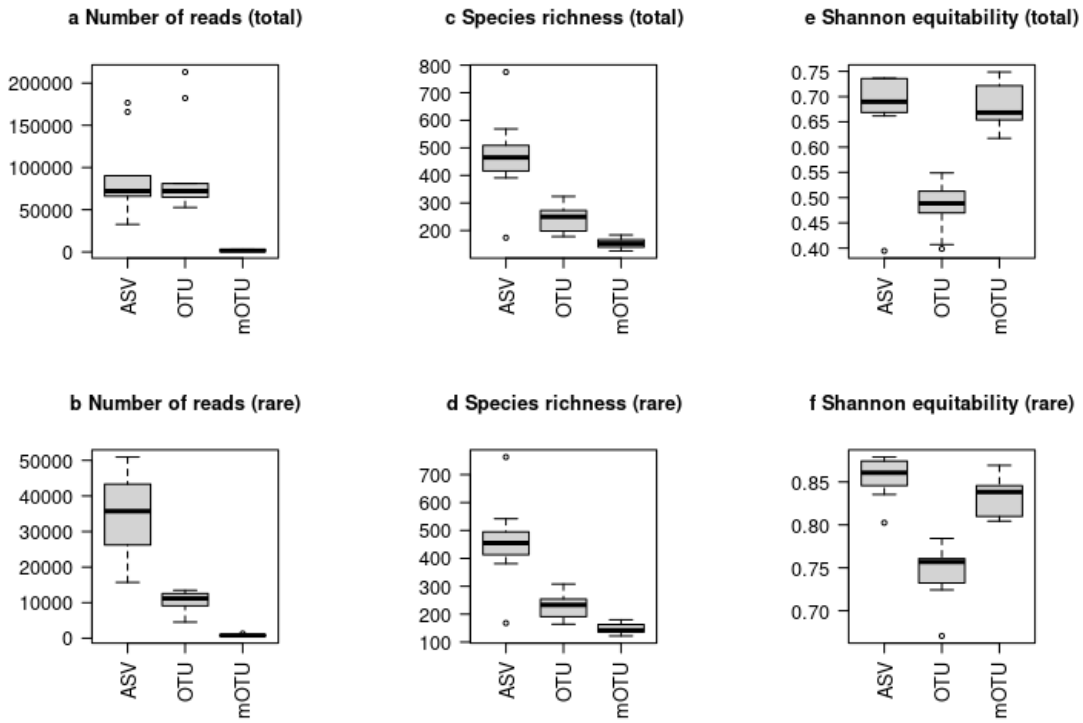


Figure 5.1. Alpha diversity for the total community (non-rarefied) and for the rare biosphere, as defined by the MultiCoLA algorithm, across ASVs, OTUs and mOTUs: **a**. Number of reads for the total community; **b.** Number of reads for the rare biosphere; **c.** Species richness for the total community; **d**. Species richness for rare biosphere; **e**. Shannon equitability for the total community; **f**. Shannon equitability for rare biosphere. n = 9 samples (each boxplot).

Species richness of the total community was higher for ASVs (from 174 to 775 ASVs, with a median value of 466 ASVs), than for OTUs (from 178 to 324 OTUs, with a median value of 250 OTUs) and mOTUs (from 126 to 184 mOTUs, with a median value of 153 mOTUs) (Figure 5.1c). The same pattern (number of ASVs > number of OTUs > number of mOTUs) was found for the rare biosphere, for the respective approaches (Figure 5.1d). Across the nine samples, for the Shannon equitability of the total community, the values were higher for ASVs (from 0.39 to 0.74, with a median value of 0.69), followed by mOTUs (from 0.62 to 0.75, with a median of 0.67) and the OTUs (from 0.40 to 0.55, with a median of 0.49) (Figure 5.1e). Similarly, the rare ASVs presented the highest

values (from 0.80 to 0.88, with a median value of 0.86), followed by the rare mOTUs (from 0.80 to 0.87, with a median value of 0.84) and rare OTUs (from 0.67 to 0.78, with a median value of 0.76) (Figure 5.1f). The equilibrium of phylogenetic units (equitability) in the total communities was higher for the ASVs and mOTUs approach than for the OTUs approach, as revealed by the range of the number of reads and the range of the number of ASVs/OTUs/mOTUS, and this pattern was maintained for the rare biosphere.

To verify if the differences in alpha diversity between the different methodologies were overestimated for the samples with higher sampling size, we applied rarefaction. Results were different for the sum of reads (Supplementary Figure 10.18a), a direct consequence of rarefaction. However, the remaining diversity metrics followed the same pattern revealed for the non-rarefied data (Supplementary Figure 10.18a-f). To discard the possible overestimation of the rare biosphere by the MultiCoLA algorithm in specific datasets (like ASVs), we compared the results with an arbitrary threshold of 0.1% relative abundance, per sample, for all datasets. Results showed again that alpha diversity was higher for the rare ASVs than OTUs, or mOTUs (Supplementary Figure 10.19a-f), while the rarefaction resulted in more reads for the OTUs than for the ASVs datasets (Supplementary Figure 10.20a) and in the abovementioned pattern for the species richness and Shannon equitability measures (Supplementary Figure 10.20b-f). An important drawback of using a threshold of 0.1% is that the rare mOTUs become virtually absent, as they are reduced to mOTUs with 1 or 2 reads. Despite that drawback, those sequences represent 48% of the mOTUs.

To understand how the rare biosphere prokaryotic communities were organized across different sampling points (Table 5.1) and from the point of view of each methodological approach (ASVs, OTUs and mOTUs), both hierarchical clustering (illustrated as dendrograms) and ordination analysis (CAs), were employed (Figure 5.2).

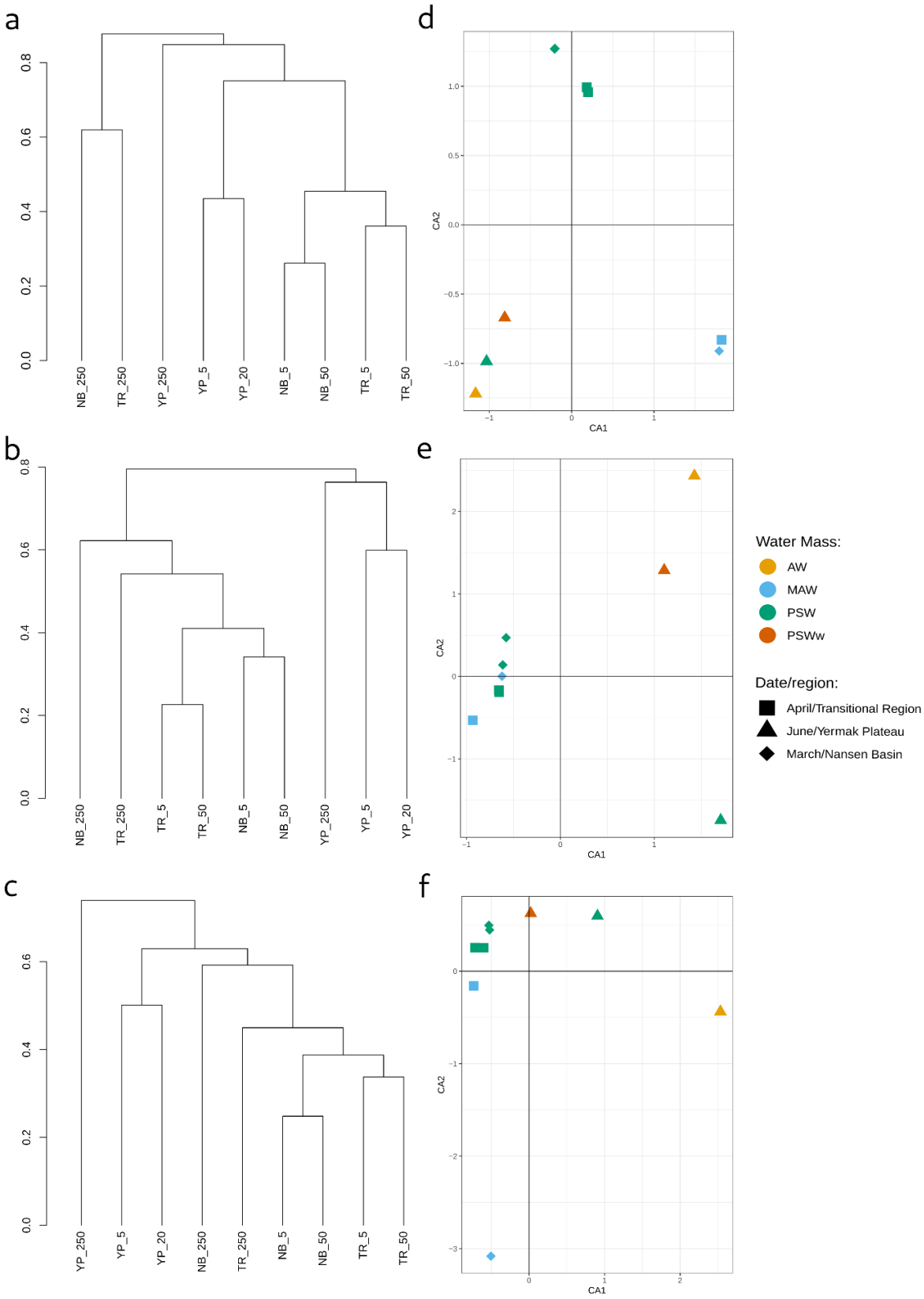


Figure 5.2. Hierarchical clustering (a-c) and ordination analysis (d-f) for the rare biosphere as defined by MultiCoLA and using non-rarefied data. **a**. Dendrogram of the rare ASVs; **b**. Dendrogram of the rare OTUs; **c**. Dendrogram of the rare mOTUs; **d.** CA for the rare ASVs; **e.** CA for the rare OTUs. **f.** CA for the rare mOTUs.

For all approaches (ASV, OTU, mOTU) the rare biosphere communities from the samples of April and March, at 5 m and 50 m, were grouped together and, collectively, they belonged to the PSW (Figure 5.2a-f). With the OTUs the 250 m samples from March and April became closer to the 5 and 50 m samples from the same months, but this did not happen in the ASV analysis, where the 250 m samples from March and April were outliers. One difference between the rare ASVs and rare OTUs was that for the rare OTUs the MAW and PSW samples were more similar with each other than for the rare ASVs (Figure 5.2d,e). The next cluster includes samples taken at the Yermak Plateau, in June, at 5 m, 20 m and 250 m depth and they were further subdivided into AW and the PSW and warmer PSW (PSWw) (Figure 5.2a-f). For the rare OTUs this pattern was found in the hierarchical clustering (Figure 5.2b), but not in the CA (Figure 5.2e), where those samples were separated. As for the rare mOTUs, the June samples were separated from each other and from the remaining samples (of March and April), in both the dendrogram and the CA (Figure 5.2c,f). Overall, the patterns were more accentuated for the ASVs, where the rare communities from distinct water masses clearly form different groups (Figure 5.2d), while for the OTUs and mOTUs the most evident pattern was the separation of the June samples from the March and April ones (Figure 5.2e,f). To verify if the clustering was a consequence of the differences in sampling size, we repeated the analysis for the rarefied data a similar clustering of samples (Supplementary Figure 10.21a-f). To verify if the clustering across different strategies was influenced by the different thresholds of rarity (obtained from MultiCoLA), we also applied the 0.1% relative abundance, per sample, threshold for the original, non-rarefied, data. Results showed the same clustering for the ASVs and OTUs (Supplementary Figure 10.21a,b,d,e), while the mOTUs did not form any specific clusters (Supplementary Figure 10.22c,f). However, when rarefaction was applied to the 0.1% relative abundance, per sample, threshold, a better resolution of clusters for the rare mOTUs was obtained (separating the June samples from the remaining ones) (Supplementary Figure 10.23f). In contrast, the patterns for the rare ASVs were similar with either the MultiCoLA or the 0.1% definition, rarefied (Supplementary Figures 10.21a,d and 10.23a,d). For the rare OTUs the separation between depths during March and April became more visible (Supplementary Figure 10.23b,e) when we used the 0.1% definition. We alert that any analysis of the rare mOTUs is not trustworthy when using 0.1% relative abundance (rarefied or not), because it translates, in this study, into communities of mOTUs represented by 1 or 2 reads (Supplementary Table 5.1).

5.3.2. Different types of prokaryotic rarity

A large proportion of the rare prokaryotic biosphere was transiently rare in all approaches (either ASVs, OTUs or mOTUs), which means that most taxa identified were always rare and disappeared in at least one sample (Figure 5.3a-c). The second most common designation was that of permanent rarity (always rare and never disappear), followed by conditional rarity (oscillate between rare and abundant) (Figure 5.3a-c). Strikingly, the proportion of phylotypes considered truly abundant was lower than those of any type of rarity in all approaches. Furthermore, when filtering specific variables (*e.g.*, samples at 5 m depth), the pattern was a higher proportion of transient than permanent and conditional rarity, in all approaches (Figure 5.3a-c). The variations in the proportion of conditionally rare taxa were more visible in the OTUs dataset (Figure 5.3b). Specific sets of samples (like different depths, in June) resulted in an increased number of permanently and conditionally rare taxa (Figure 5.3a-c). To verify if the variation in the types of rarity across different sets of samples was due to differences in the number of sequences, we compared the same analysis for the rarefied data and found that the highest variation was that of transient and permanently rare taxa (Supplementary Figure 10.24a-c). Furthermore, the proportion of conditionally rare taxa became more constant for OTUs (Supplementary Figure 10.24b) and decreased for ASVs and mOTUs (Supplementary Figure 10.24b,c). Independently of rarefaction, the approach with highest proportion of transient rarity was that based on ASVs (Figure 5.3a and Supplementary Figure 10.24a). We repeated the analysis for a common threshold of rarity (0.1%, per sample). With that alternative threshold the proportion of abundant taxa increased, as well as the conditionally rare taxa, while the proportion of permanently rare taxa decreased, but the transient rarity remained the type of rarity with the highest proportion (Supplementary Figure 10.25a-c). The rarefied results for a threshold of 0.1%, per sample, were similar, although there was a slight increase in the proportion of conditionally rare taxa (Supplementary Figure 10.25a-c). As noted previously, this definition of rarity (rarefied or not) is not suited for the mOTUs dataset, indeed half the proportion is relative to abundant or conditionally rare taxa, while the remaining reflect singletons and doubletons.

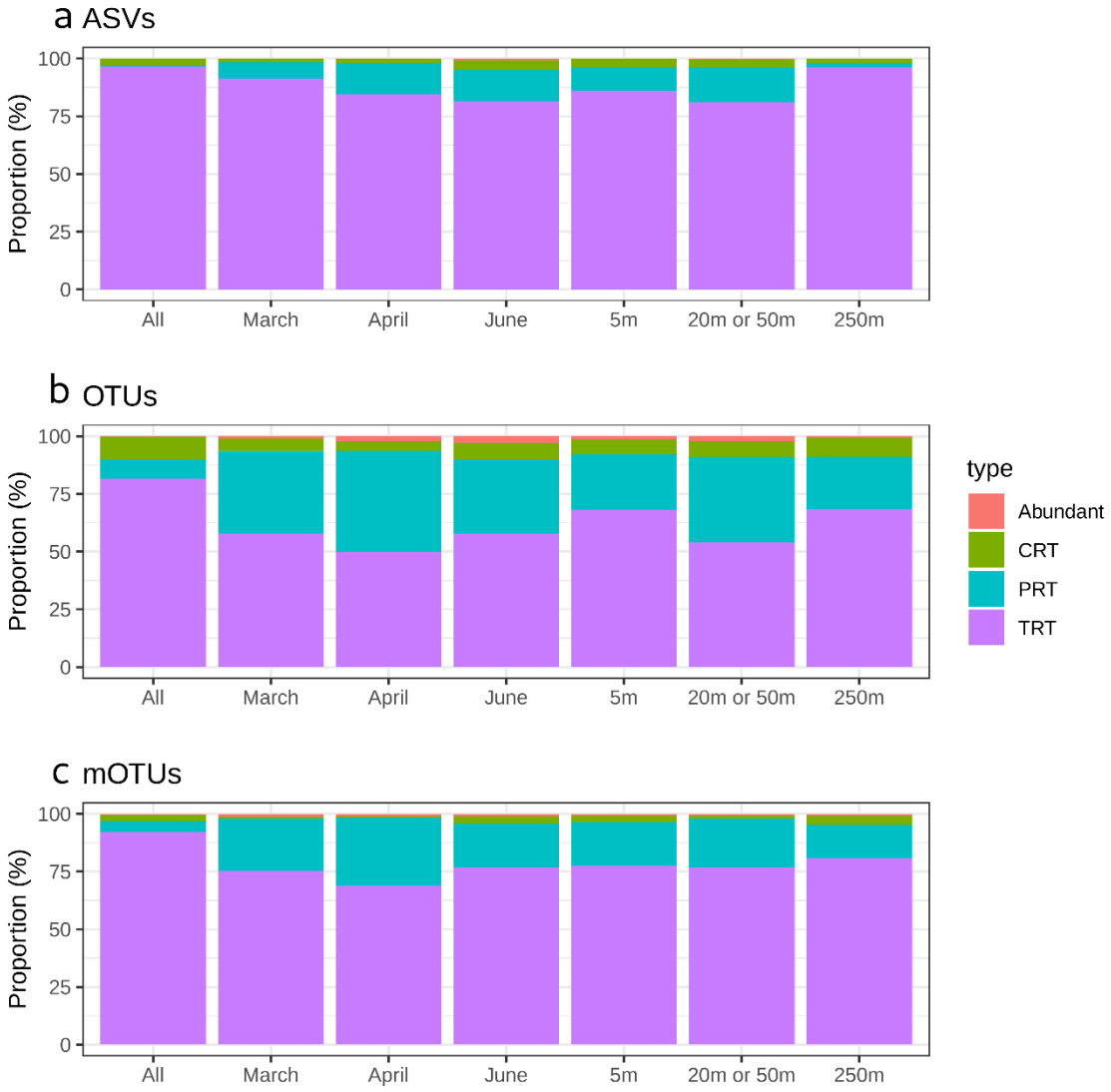


Figure 5.3. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. Rare biosphere was defined by the MultiCoLA algorithm using non-rarefied data. **a.** The proportions for the rare ASVs; **b.** The proportions for the rare OTUs; **c.** The proportions for the rare mOTUs. The stacked bar plots are colored with purple for transiently rare taxa (TRT), blue for permanently rare taxa (PRT), green for conditionally rare taxa (CRT) and red for the abundant taxa. Within each dataset, different groups of samples were selected, specifically: "All" is for all samples; "March" is the group of samples from the month of March and they represent different depths (5 m, 50 m and 250 m); "April" is the group of samples from the month of April and they represent different depths (5 m, 50 m and 250 m); "June" is the group of samples from the month of June and they represent different depths (5 m, 20 m and 250 m); "5 m" is the group of samples taken at 5 m of depth and they represent the different months (March, April and June); "20 m or 50 m" is the group of samples taken at 20 m or 50 m of depth and they represent the different months (March, April and June); "250 m" is the group of samples taken at 250 m of depth and they represent the different months (March, April and June).

The types of rarity were further attributed to the taxa identified (Figure 5.4-5.6). The inner circle of each circular visualization represents the types of rarity, thus, by comparing the coloring in the different Circos figures, it is possible to associate the previous information (that of Figure 5.3a-c) with the taxonomy and see how the types of rarity vary across methods within the same samples. Specifically, in the ASVs Circos, for each phylum, the

transiently rare dominated (Figure 5.4, filled with purple) with few conditional and permanent rare taxa (Figure 5.4, green and blue). The OTUs Circos (Figure 5.5) is also dominated by transiently rare taxa, although all kinds of rarity were found in most phyla. Interestingly, *Planctomycetes* presented more permanently than transiently rare taxa. For the mOTUs (Figure 5.6), most phyla were composed of many transiently and few permanently rare taxa, with only the major phyla, like *Proteobacteria*, having abundant mOTUs. Also, the candidate phyla were the most transient of all, since they were found in the fewest number of samples (Figure 5.6).

4.3.3. Rare prokaryotic taxonomy

From a taxonomic point of view, the rare ASV sequences were classified into 30 different phyla (Figure 5.4), of which 7 were candidate phyla (Table 5.3). The rare OTU sequences were classified into 34 different phyla (Figure 5.5), of which 6 were candidate phyla (Table 5.3), while the rare mOTU sequences were classified into 56 phyla (Figure 5.6), of which 28 were candidate phyla (Table 5.3).

Table 5.3. List of the presence or absence of candidate phyla from within and outside the candidate phyla radiation, using the phylogenetic tree from Hug *et al.*(2016) as reference, across the different bioinformatics approaches used in here (ASVs, OTUs, and mOTUs).

Name	ASV	OTU	mOTU
Candidate phyla radiation, using Hug *et al.* (2016) tree			
Candidatus Gracilibacteria	No	Yes	No
Candidatus Peregrinibacteria	No	Yes	Yes
Candidatus Saccharibacteria	No	Yes	Yes
Candidatus Uhrbacteria	No	Yes	Yes
Candidatus Adlerbacteria	No	No	Yes
Candidatus Amesbacteria	No	No	Yes
Candidatus Azambacteria	No	No	Yes
Candidatus Berkelbacteria	No	No	Yes
Candidatus Campbellbacteria	No	No	Yes
Candidatus Collierbacteria	No	No	Yes
Candidatus Falkowbacteria	No	No	Yes
Candidatus Kaiserbacteria	No	No	Yes
Candidatus Kerfeldbacteria	No	No	Yes
Candidatus Magasanikbacteria	No	No	Yes
Candidatus Nomurabacteria	No	No	Yes
Candidatus Peregrinibacteria	No	No	Yes
Candidatus Roizmanbacteria	No	No	Yes
Candidatus Wolfebacteria	No	No	Yes
Other non-candidate phylum radiation, uncultivated phyla			
Candidatus Marinimicrobia (SAR406)	Yes	Yes	Yes

Candidatus Hydrogenedentes	No	Yes	No
Candidate division NC10	No	No	Yes
Candidatus Diapherotrites	No	No	Yes
Candidatus Doudnabacteria	No	No	Yes
Candidatus Komelibacteria	No	No	Yes
Candidatus Lloydbacteria	No	No	Yes
Candidatus Omnitrophic	No	No	Yes
Candidatus Rokubacteria	No	No	Yes
Candidatus Woesearchaeota	No	No	Yes
Candidatus Yonathbacteria	No	No	Yes
Candidatus Zambryskibacteria	No	No	Yes
Candidatus Anck6	Yes	No	No
Candidatus Dadabacteria	Yes	No	No
Candidatus Dependentiae	Yes	No	No
Candidatus PAUC34F	Yes	No	No
Candidatus Schemanbacteria	Yes	No	No
Candidatus WPS-*2*	Yes	No	No

End of Table 5.3.

All datasets were able to identify several so-far uncultivated phyla (Table 5.3), but only the OTUs and mOTUs approaches were able to recover members of the Candidate Phyla Radiation (CPR), as defined in Hug *et al*. (2016), with the metagenomics approach identifying up to 17 distinct CPR phyla (Table 5.3). A candidate phylum identified by all approaches was *Candidatus* Marinimicrobia (previously SAR406), which was rare across the three methodologies used, although it did have a few conditionally rare members in the OTUs dataset, and a few abundant ones in the mOTUs dataset, while in the ASVs dataset they were always rare (mostly TRT and a few PRT) (see Figures 5.4a-c).

Interestingly, for each dataset the overall pattern of types of rarity was distributed across all the major taxonomic groups. Thus, our results suggest that there was not a connection between the taxonomy (at higher level, that is, phylum and class) and the distribution pattern of the type of rarity (that is the PRT, TRT or CRT designations). There were few exceptions to this pattern, for example, in *Planctomycetes* (from the OTUs dataset), the most prevalent type of rarity was PRT (Figure 5.5).

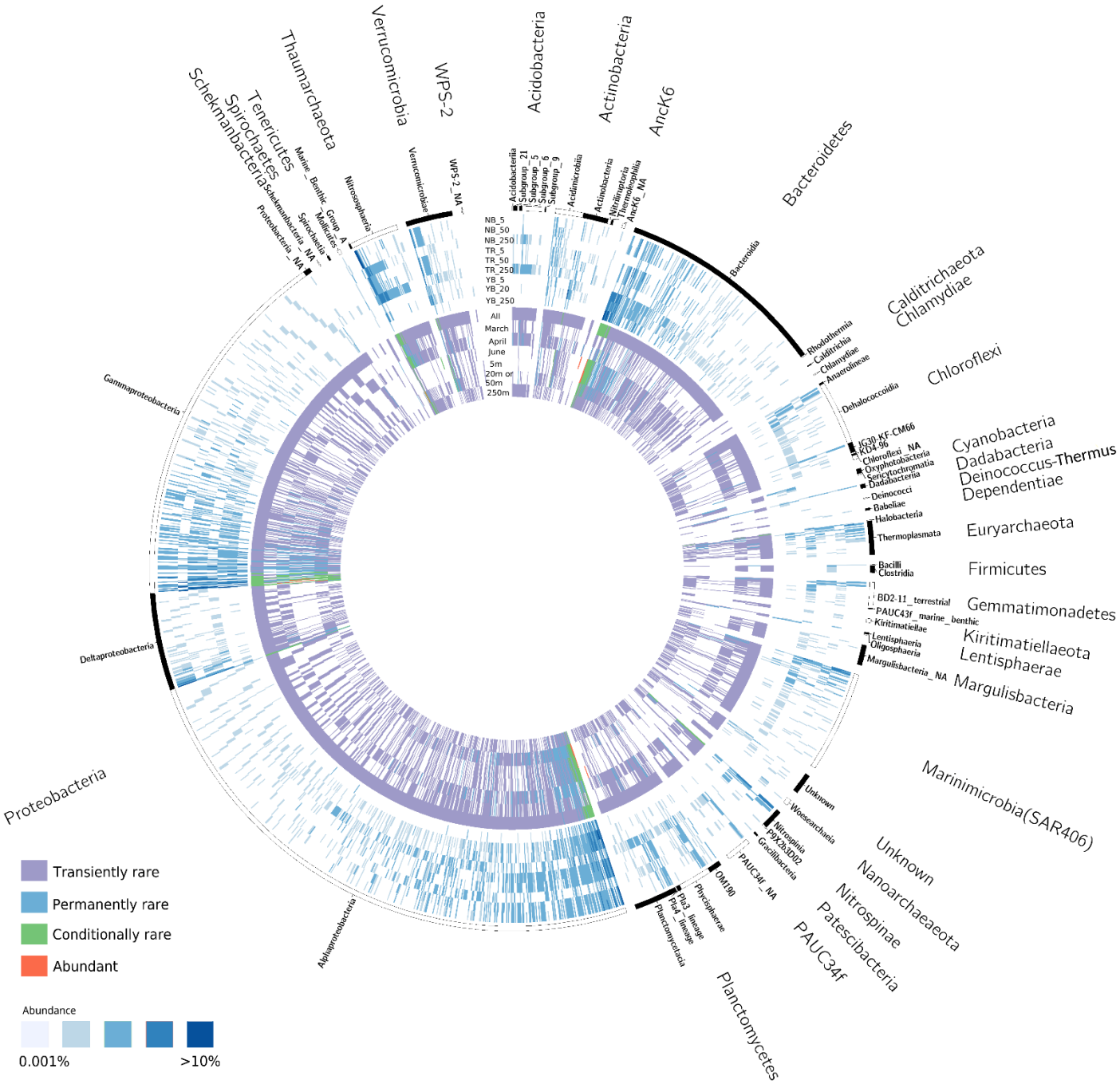


Figure 5.4. Circular visualizations of the ASVs taxonomy, abundance and types of rarity. The relative abundance is in the outer circle (blue heatmaps) and the types of rarity of each taxon (if rare), for each set of compared variables, are in the inner circle (multi-colored). Specifically: “All” is for all samples; “March” is the group of samples from the month of March and they represent different depths (5 m, 50 m and 250 m); “April” is the group of samples from the month of April and they represent different depths (5 m, 50 m and 250 m); “June” is the group of samples from the month of June and they represent different depths (5 m, 20 m and 250 m); “5m” is the group of samples taken at 5m of depth and they represent the different months (March, April and June); “20 m or 50 m” is the group of samples taken at 20 m and 50 m of depth and they represent the different months (March, April and June); “250 m” is the group of samples taken at 250 m of depth and they represent the different months (March, April and June). The types of rarity (if rare) are depicted in inner heatmap, with qualitative colors, specifically: red is for abundant taxa, blue is for permanently rare taxa (PRT), green is for conditionally rare taxa (CRT) and purple is for transiently rare taxa (TRT). The taxonomic identification used for the figures is that of the methodological output, that is why the *DADA2* derived phyla do not have the *Candidatus* designation in the Circos label.

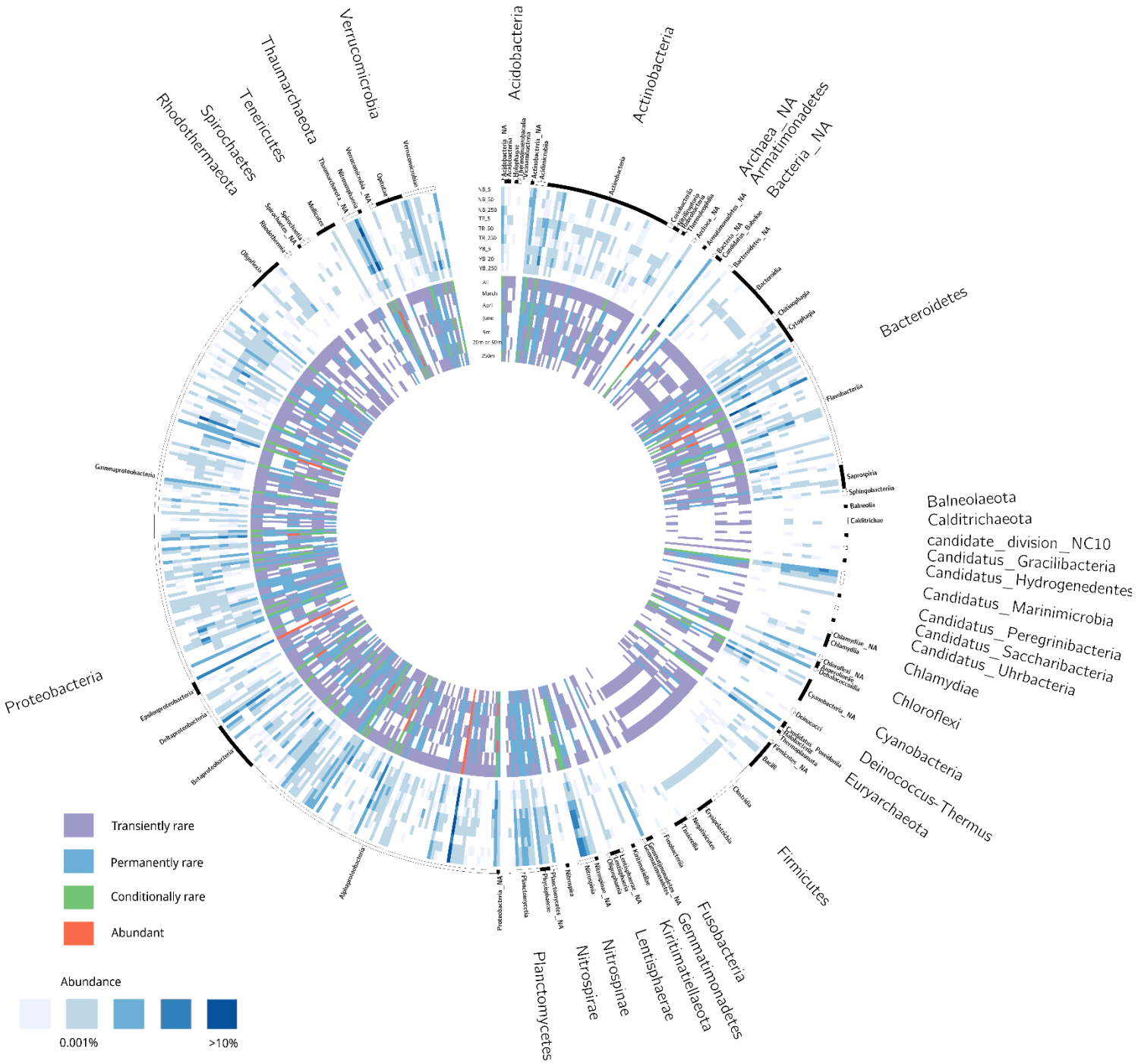


Figure 5.5. Circular visualizations of the OTUs taxonomy, abundance and types of rarity. For instruction on how to read this figure, see legend of Figure 5.4.

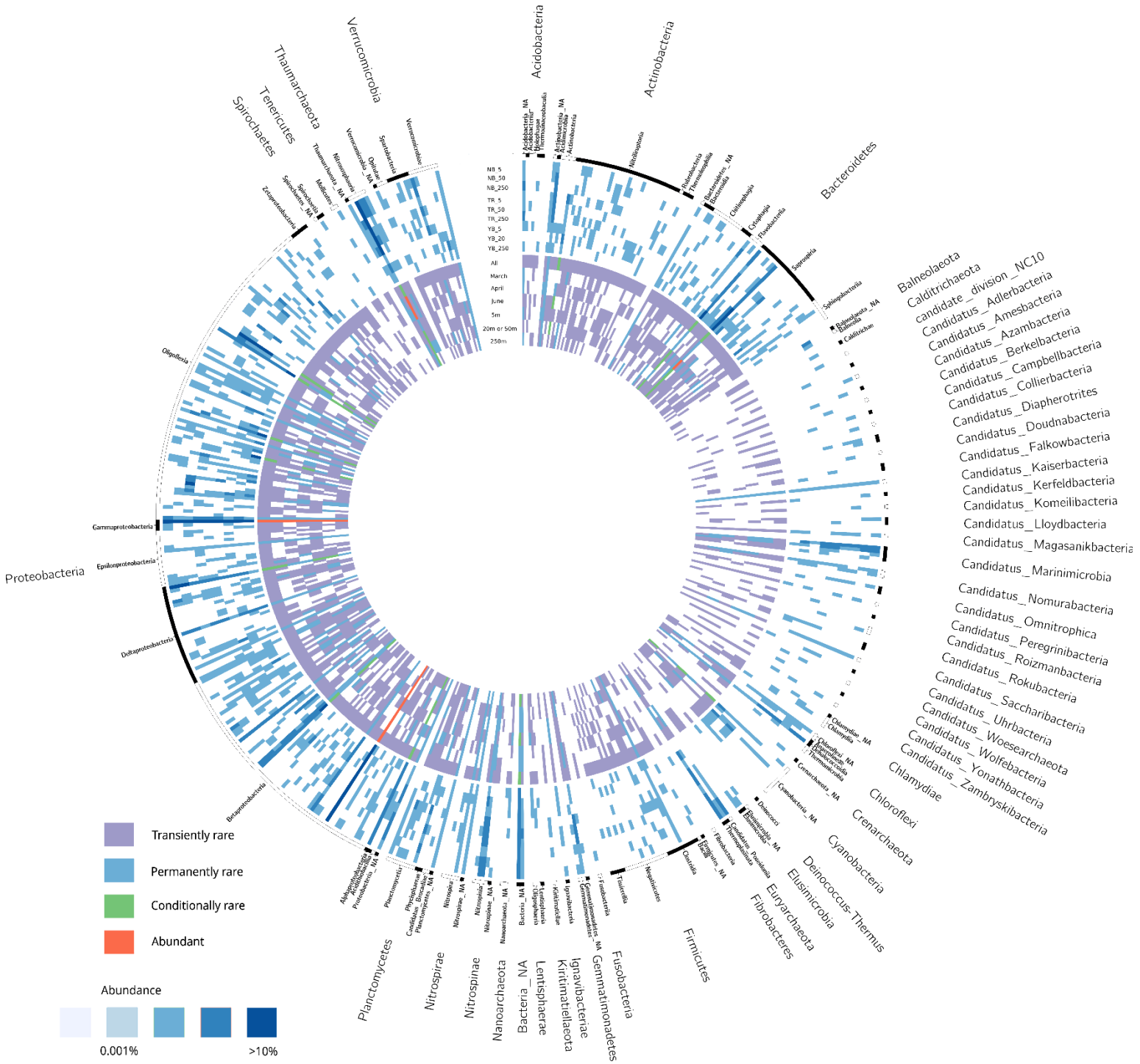


Figure 5.6. Circular visualizations of the mOTUs taxonomy, abundance and types of rarity. For instruction on how to read this figure, see legend of Figure 5.4.

5.4. Discussion

5.4.1. The prokaryotic rare biosphere of the Arctic Ocean is sensitive to environmental heterogeneity

Our analysis showed that very few taxa remain abundant in this heterogeneous environment (mOTUs and OTUs) and none were always abundant across the entire set of samples (ASVs). In fact, the environmental heterogeneity imposed by the different water masses of the region is one of the most important aspects of microbiome composition and distribution in the Arctic Ocean (Wilson *et al.*, 2017; Müller *et al.*, 2018; de Sousa *et al.*, 2019), including the rare biosphere (Galand *et al.*, 2009; Kirchman, Cottrell and Lovejoy, 2010). This was supported by our amplicon- and metagenome-derived 16S rRNA gene data, where the PSW and MAW formed a cluster in March and April, separated from the June samples. The clusters of rare biosphere community composition were clearer for the ASVs than for the OTUs and mOTUs approaches, which might be a consequence of clustering of similar sequences in both OTU based methods. Overall, alpha diversity metrics were higher using ASVs than OTUs or mOTUs (Figure 4.1). Although approximately the same number of sequences was processed for the retrieval of 16S rRNA gene derived OTUs and ASVs, after filtering for the rare biosphere the ASVs approach resulted in more rare sequences than did the OTUs approach. This is most likely because OTUs cluster sequences based on a cutoff of similarity and within a single OTU several distinct ASVs can be described. As shown previously, the clusters of sequences that form a single OTU, with as low as 1% identity cutoff, can be divided into several ASVs. This was not a consequence of an overestimation of the rare biosphere by the MultiCoLA algorithm, because we identified a similar relationship with a different, common threshold (0.1% per sample), independently of rarefaction (Supplementary Figures 10.19 and 10.20). It was also not a consequence of sampling size in the MultiCoLA-based definition since the same pattern remained after rarefaction (Supplementary Figure 10.18). The metagenomes, as expected, scored lower on the number of reads, because the ability to retrieve 16S rRNA genes from the environment was lower in comparison with amplicon sequencing of the marker genes. Because of that difference, the definition of 0.1% relative abundance proved adequate for amplicon-based approaches but did not for shotgun sequencing of metagenomes. Thus, it does not necessarily mean that the metagenomes did not identify the rare biosphere, but it means they did not identify the same level of rarity as in the amplicon-based approaches.

We suggest that there is a deeper problem, which regards the concept of rarity in biology, which remains very subjective and hard to tackle methodologically.

The distinction of types of rarity, using the terminology from Jia *et al*. (2018), further allows us to understand how the abundance of rare taxa behaves across a set of samples. Independently of the definition of rarity, rarefaction, and method (ASVs, OTUs or mOTUs), most taxa remained rare or disappeared and only a minority was conditionally rare. The best approach to detect variations in the conditional rarity was the OTUs from 16S rRNA gene sequencing, which is also the most used strategy in the field. Interestingly, the conditional rarity was higher when comparing all samples at once and decreased in specific sets of samples (for example, in June). This is because of the heterogeneous conditions reflected by all the examined samples (which range in depth, season, and water masses): in each condition there are few abundant types, but they persist as rare in the other conditions, thus they are catalogued as conditionally rare across all samples. If we subset samples for a limited range of conditions (for example, samples at 5 m of depth), then we can identify groups of taxa which are always rare, but if we add new conditions, the chance of the taxa growing into abundant increases. Thus, by adding all samples at once, we identify all taxa that oscillate between rare and abundant, but that will appear as permanently rare if we restrict the analysis to a smaller group of samples. Likewise, the same explanation holds for the transiently rare taxa, because the chance of the taxa disappearing in at least one sample grows with the growing number of samples.

All methods were able to identify distinct uncultivated rare phyla. For example, *Candidatus* Marinimicrobia (SAR406) was identified with all approaches. Interestingly, this phylum was divided into more distinct ASVs than into distinct OTUs or mOTUs. Most of the *Candidatus* Marinimicrobia (SAR406) ASVs were transiently rare and only a few were permanently or conditionally rare. The number of transiently rare taxa was lower for the OTUs and mOTUs datasets, where the same *Candidatus* Marinimicrobia phylotypes were identified more consistently across samples. Possibly, this is a direct reflection of the higher resolution of the ASV method in distinguishing different taxa, although this did not correspond to increased taxonomic accuracy since the *Candidatus* Marinimicrobia sequences retrieved could not be classified at the class level and below. *Candidatus* Marinimicrobia is a very diverse phylum which has been identified in several environments. It has been suggested that *Candidatus* Marinimicrobia from distinct water masses is characterized by genomic differences (Martinez-Gutierrez and Aylward, 2019). Thus, one hypothesis to explain why some members of this phylum were very

transient, while others persisted across different sampling points, is that the *Candidatus* Marinimicrobia types consistently found in the samples were adapted to the Arctic Ocean waters analyzed, while the transient *Candidatus* Marinimicrobia members resulted from random dispersal from other waters (for example, deeper layers).

In a previous biogeographical assessment of Arctic Ocean microbial communities, most taxa from the rare biosphere were permanently rare (Galand *et al.*, 2009) and the same was found for the western Arctic Ocean, during the summer to winter transition (Kirchman, Cottrell and Lovejoy, 2010). These studies used 16S rRNA gene derived OTUs and, in our 16S OTUs data, several OTUs were identified as always rare, across a gradient of environmental conditions, but the most prevalent type of rarity was TRT. Notwithstanding, the designation of transient rarity was not applied in Galand *et al.*, (2009) and Kirchman *et al.*, (2010) but included within the permanently rare taxa. Thus, our results on the types of rarity in the Arctic Ocean are coherent with the previous literature.

5.4.2. Methodological strategies and trade-offs

To date, studies on the rare biosphere ecology of the Arctic Ocean have used 16S rRNA gene OTUs (Galand *et al.*, 2009; Bowman *et al.*, 2012; Fernández-Gómez *et al.*, 2019; Gokul *et al.*, 2019). For broad-scale ecological interpretation of entire microbiomes, *i.e.*, either rare or not, it has been shown to be indifferent the utilization of OTUs or ASVs (Glassman and Martiny, 2018). However, a recent study showed that the higher taxonomic resolution of ASVs could improve the observation of biogeographic patterns of the rare biosphere (Bay *et al.*, 2020). It has long been known that the rare biosphere includes novel diversity (Elshahed *et al.*, 2008). Thus, even though high throughput sequencing of 16S rRNA genes is a suitable technique for rare biosphere assessments because of its high capacity to retrieve low abundance taxa (Pedrós-Alió, 2012), the mOTUs approach can offer a few underexplored advantages in the rare biosphere context (Pascoal, Magalhães and Costa, 2020). An example is the improved identification of CPR and other so-far uncultivated taxa, as supported by our data (compare Figures 5.4 to 5.6 and Table 5.3). In fact, independently of the alpha diversity scores for the rare biosphere, the rare taxonomic diversity found was more than double for the mOTUs compared to both amplicon sequencing approaches, independently of the binning method (ASVs or OTUs). CPR is a relevant topic for the understanding of biological diversity, but it has been overlooked from the perspective of the rare biosphere (Pascoal, Magalhães and Costa, 2020). Again, we note that the identification of the rare

biosphere with metagenomes is very dependent on the definition of rarity used. Despite this, several candidate phyla specific to the metagenomes had much lower abundance than the most common phyla and, in this case, that is not due to any PCR bias towards better described taxonomic groups.

The overall ecological patterns of the rare biosphere diversity are still better analyzed with the amplicon-based approaches as these are more sensitive to detect structural shifts across sample groups, owing to higher coverage of the diversity of the marker gene in use (Pedrós-Alió, 2012). This sensitivity, in our study, was even higher for ASVs (Figure 5.2). We argue that both amplicon and metagenome-based approaches can be concomitantly used to provide a complementary view on the rare biosphere diversity.

Regarding the definition of rarity, most studies pre-define a random threshold of 0.1% or 0.01% (per sample), but currently there is no objective methodology to calculate that threshold. Furthermore, those thresholds have been so far applied to SSU rRNA gene amplicon sequencing datasets, which contain sequence read numbers of orders of magnitude higher than metagenomic derived OTU datasets. Thus, because distinct methods were compared in this study, namely 16S rRNA gene amplicons processed into ASVs and OTUs, versus mOTUs, the utilization of 0.1% or 0.01% as rarity thresholds to all datasets would be problematic. For instance, if using 0.1% relative abundance per sample, as threshold of rarity for the mOTUs data, then the rare communities would be virtually composed of singletons and doubletons (per sample). To solve this problem, we took the approach suggested by Jia, Dini-Andreote and Falcão Salles (2018) and used the MultiCoLA algorithm (Gobet, Quince and Ramette, 2010) to define rarity thresholds appropriate to each dataset analyzed in this study. This approach resulted in the decision of an absolute abundance (per sample) threshold. The relative abundance equivalent to the absolute abundance used for each dataset ranged between values that were not far from those of the literature. Nevertheless, in 1 ASV sample and in 5 mOTU samples the threshold was above 5%, with thresholds going up to 5.7% (YP_250, ASVs) and 6.5% (TR_5, mOTUs) (Table 5.2), which are more relaxed rarity thresholds than those used in most studies (0.1% or 0.01%). From our perspective, these thresholds are suitable to define the rare biosphere in each sample within any given dataset, but the methodology used remains somewhat arbitrary, in that it depends on the interpretation of the MultiCoLA output, which is not obvious. In theory, this method assumes that the total and rare communities possess distinct structures (Gobet, Quince and Ramette, 2010). Thus, by progressively removing the most abundant taxa, the point where the similarity between the original and rarefied communities change drastically would correspond to

the threshold at which biological “rarity” would begin (Jia, Dini-Andreote and Falcão Salles, 2018). In this study, the MultiCoLA method was used to compare the beta diversity of the same community with one abundance threshold versus the next, for several thresholds, which were illustrated in the Procrustes plots (Supplementary Figure 10.15 and 10.16). One limitation in our exploration of the types of rarity was the transient rarity definition, because, with increasing numbers of samples, the chance of one rare phylotype to disappear (either ASV, OTU or mOTU) increases substantially. Is it also possible that, with increasing sequencing depth, the number of transiently rare taxa could decrease, but we did not identify that problem when comparing, for example, rarefied vs non-rarefied data.

5.5. Conclusion

Our results, across three different methodologies, were checked against several parallel statistical analyses and agreed with the previous literature, which used different methods and sampling points from the Arctic Ocean. Thus, we argue that the multiple studies and methodologies used in the survey of the rare biosphere in the Arctic Ocean, collectively, provide strong evidence that the rare biosphere remains rare, while only a small fraction oscillates between rare and abundant, with the cyclical climate conditions. Finally, our work highlights the need for a better understanding of the rare biosphere definition, it found that ASVs and OTUs are both valid to study the rare biosphere community patterns and identified the advantages of combining amplicon-based strategies with shotgun sequencing metagenomics.

5.6. Declarations

Author contribution

Conceptualization (FP, RC and CM); R code for all data analyses (FP); circular visualization with Circos software (FP); sampling campaign design (CM and PD); funding (all authors); writing – original draft (FP); writing – review and editing (all authors).

Conflicts of interest

The authors declare no conflict of interest.

Funding

FCT funded this study through the grants PTDC/CTA-AMB/30997/2017 and PTDC/CTA-AMB/4946/2020 to CM, 2020.03139 CEECIND to CM and the PhD grant 2020.04453.BD to FP. Further Arctic campaign logistic and traveling support was provided by PROPOLAR and by the former Centre for Ice, Climate and Ecosystems at the Norwegian Polar Institute, the Research Council of Norway (project no. 244646), the Norwegian Ministries of Foreign Affairs and Climate and Environment through the program Arktis 2030 (project ID Arctic). This study was also partially funded by “Programa Operacional Regional de Lisboa” (Project N. 007317) and the Strategic Funding UIDB/04423/2020, UIDP/04565/2020 and UIDB/04565/2020 through national funds provided by FCT and ERDF, in the framework of the “PT2020” program.

Availability of data

The raw sequences used for 16S rRNA gene analysis and metagenomes are publicly available at ENA under the accession numbers: PRJEB21950 and PRJEB15043. The sample vs 16S rRNA gene derived OTUs table is available in the MGnify platform (MGYS00001922). The sample vs. metagenomic derived OTUs (mOTU) table is available at the MGnify platform (MGYS00001869).

Code availability

The code used for processing raw 16S rRNA gene sequences into ASVs and the code for the diversity metrics used for all datasets (ASVs, OTUs and mOTUs) are available in a single R script in the Supplementary Information of the original publication page: https://link.springer.com/article/10.1007/s00248-021-01821-9#Sec16.

Custom commands were used for the Circos figures, based on the data obtained from the R script provided.

Part III. Development of an unsupervised machine learning approach to define the microbial rare biosphere.

Chapter 6. Definition of the microbial rare biosphere through unsupervised machine learning.

Based on materials published by Communications Biology journal, at time of the thesis submission:

Pascoal, F. *et al.* (2025a) “Definition of the microbial rare biosphere through unsupervised machine learning,” *Communications Biology*, 8(1), p. 544. Available at: https://doi.org/10.1038/s42003-025-07912-4.

See author contributions in section **6.6. Declarations**.

6. Definition of the microbial rare biosphere through unsupervised machine learning

Francisco Pascoal[1,2], Paula Branco[3], Luís Torgo[4], Rodrigo Costa[5,6]* and Catarina Magalhães[1,2]

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal; [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal; [3]School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa; [4]Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada; [5]Department of Bioengineering, Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico, University of Lisbon, Portugal; [6]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, University of Lisbon, Portugal.

Abstract

The microbial rare biosphere, composed of low-abundance microorganisms in a community, lacks a standardized delineation method for its definition. Currently, most studies rely on arbitrary thresholds to define the microbial rare biosphere (*e.g.*, 0.1% relative abundance per sample), hampering comparisons across studies. To solve this problem, we present *ulrb* (Unsupervised Learning based Definition of the Rare Biosphere), available as an R package. *ulrb* uses unsupervised machine learning to optimally classify taxa into abundance categories (*e.g.*, rare, intermediate, or abundant) within microbial communities. We show that *ulrb* is more consistent than threshold-based approaches and can be applied to data derived from common microbial ecology protocols and non-microbial studies. *ulrb* can be used to identify different types of rarity and is statistically valid for the analysis of various dataset sizes. In conclusion, *ulrb* discerns rare from abundant organisms in a user-independent manner, finding applicability across multiple data types and ecological contexts.

Keywords: rarity, machine learning, microbial ecology, abundance classification, molecular ecology.

6.1. Introduction

The standard computational measure to study the rare biosphere is to order all taxa from the most to the least abundant, in a Rank Abundance Curve (RAC). The shape of the RAC can be mathematically described by the power-law (Ser-Giacomi *et al.*, 2018), whereby a few taxa are abundant, but many are rare in the so-called long tail of the RAC. This is a common trend found in biological communities in general, including microbial communities (McGill *et al.*, 2007; Pedrós-Alió, 2012; Lynch and Neufeld, 2015; Jousset *et al.*, 2017; Pascoal, Costa and Magalhães, 2021). One of the first studies recognizing the general trend of a high number of rare species is *The Origin of Species*, where Darwin states: “rarity is the attribute of a vast number of species” (Darwin, 1859). In general ecology, it is important to identify and quantify rare species within biological communities to help policy makers on biodiversity conservation, as these species are often closer to extinction (Gaston and Fuller, 2008). The microbial rare biosphere (Sogin *et al.*, 2006) is considered a reservoir of genetic diversity (Pedrós-Alió, 2012; Lynch and Neufeld, 2015) maintaining the genetic potential for biogeochemical cycles of crucial relevance for the resistance and resilience of ecosystems (Jousset *et al.*, 2017), a source of symbionts shaping host-associated microbiomes (Taylor *et al.*, 2013), and a source of novel biosynthetic genes (Pascoal, Magalhães and Costa, 2020).

Most studies define the microbial rare biosphere using relative abundance thresholds such as 0.1% or 0.01% per sample (*e.g.,* (Quero and Luna, 2014; Fuentes *et al*., 2016; Dawson *et al*., 2017; Idris *et al*., 2017; Richa *et al*., 2017; Wang *et al*., 2017; De Anda *et al*., 2018; Gokul *et al*., 2019; Pascoal *et al*., 2022a; Tang *et al*., 2023)), based on early microbial ecology studies of the RAC (Pedrós-Alió, 2006; Sogin *et al.*, 2006). One of the problems with threshold-based approaches to define the rare biosphere is that the same threshold will necessarily have a different meaning according to the specific methodology used or the assembly features of the community, such as degrees of dominance and equitability among taxa. Moreover, different thresholds provide different interpretations of the RAC and most likely none provides consistent results across different methods or communities. Using more specific examples, the meaning of a 0.1% relative abundance, per sample, is different between using amplicon sequencing of a small region of the 16S rRNA gene or using shotgun metagenome sequencing. This is because the methods produce abundance tables with taxon abundance scores in different orders of magnitude. Thus, a definition of 0.1% relative abundance, per sample, might work well to describe the long RAC tail of a 16S rRNA sequencing dataset. However, this same threshold would yield a very different view of the rare biosphere from the shotgun metagenome

sequencing data from the same sample (Pascoal *et al.*, 2022a) (Chapter 5). This is a problem, because it impairs inter-comparability across studies and sequencing methodologies. Additionally, the utilization of relative abundance thresholds is fundamentally flawed, even if it is supported by the interpretation of a RAC, because different researchers could draw conclusions pointing to the utilization of different thresholds based on the same RAC, *i.e.*, threshold-based approaches are subjective methods to delineate the beginning of the "long-tail" of the RAC (see Supplementary Figure 10.27 for a schematic representation of the problem).

Here, we propose an unsupervised machine learning approach to solve the major issues of the threshold-based methods to define the microbial rare biosphere. We refer to our approach and respective methodology (using default parameters, unless stated otherwise) as Unsupervised Learning based Definition of the Rare Biosphere (*ulrb*).

ulrb clusters all taxa sampled from a biological community using the k-medoids model with the partitioning around medoids algorithm (pam) (Kaufman and Rousseeuw, 1987). The k-medoids model is an unsupervised learning model that partitions points of data into k clusters, minimizing the distance between the points and the centroid of the clusters (Kaufman and Rousseeuw, 1987). Within *ulrb*, the points are the taxa abundance score in a given sample, and the clusters represent their abundance classifications. The *ulrb* method allows for different number of classifications, which can be adapted to the experimental design of the user. As a default parameter, *ulrb* uses three clusters (k=3), corresponding to the classifications "rare", "undetermined" and "abundant". The "undetermined" classification can be interpreted as "intermediate", that is, a state of abundance between "rare" and "abundant". There are metrics that can be used to inspect what is the best number of classifications (Calinski and Harabasz, 1974; Davies and Bouldin, 1979; Rousseeuw, 1987) and there is an option to automatically decide the number of classifications in *ulrb* (see section 5.2.1.3).

The introduction of an intermediate classification is optional but recommended to avoid the existence of taxa with very similar abundance scores having opposite classifications ("rare" or "abundant"). Previous studies, using relative abundance thresholds, have also introduced intermediate classifications to provide more comprehensive information (Vergin *et al.*, 2013; Baltar *et al.*, 2015). The ecological implication of considering intermediate classifications is the acknowledgment that some taxa are neither rare nor abundant, for example, they might be transitioning between being rare and abundant, as conditionally rare taxa (Jones and Lennon, 2010; Shade *et al.*, 2014). The most important aspect of *ulrb* is that it automatically classifies taxa based solely on their abundance

score within a community. Furthermore, the method considers that a taxon is not rare/abundant by itself. Instead, a taxon is rare relative to another that is abundant, or vice-versa.

Previous studies have proposed alternative ways of defining the rare biosphere, for example, by calculating the impact of different thresholds on beta diversity with MultiCoLA (Gobet, Quince and Ramette, 2010; Jia, Dini-Andreote and Falcão Salles, 2018). However, in a previous study we showed that MultiCoLA did not resolve the arbitrary nature of threshold-based approaches to define the rare biosphere (Pascoal *et al.*, 2022a) (Chapter 5). Other studies have suggested evaluating several thresholds against the RAC (Jia, Dini-Andreote and Salles, 2022; Ramond *et al.*, 2023) and recalibrate according to sequencing depth, using the ratio between observed and expected taxa (with Chao index)(Jia *et al.*, 2022). Outside the scope of microbial ecology, the utilization of unsupervised learning to define rare and common species has been proposed with the FuzzyQ method (Balbuena *et al.*, 2021).

The objective of this study is to present an unsupervised machine learning approach for the definition of the rare biosphere and validate its applicability to a wide range of datasets. Our method can be used for the analysis of any biological community with the R package *ulrb* (Unsupervised Learning based Definition of the Rare Biosphere), which uses open-source code and is available in The Comprehensive R Archive Network (CRAN, https://cloud.r-project.org/web/packages/*ulrb*/index.html) and GitHub (https://github.com/pascoalf/*ulrb*) repositories. Additionally, the R package *ulrb* includes a dedicated website with several tutorials and extensive documentation on all functions (https://pascoalf.github.io/*ulrb*/). *ulrb* was tested against microbial communities obtained from different sequencing and bioinformatics strategies and compared against threshold-based methods for the description of the rare biosphere. Its statistical validity was evaluated against variations in the number of phylogenetic units, samples and sequencing depth. Further, the applicability of the *ulrb* for non-microbial (animal and plant) datasets was tested, while also applying the FuzzyQ method to the microbial datasets analyzed in this study. Finally, an *ulrb* extension to identify types of rarity in a host-microbiome context was illustrated.

6.2. Methods

6.2.1. The *ulrb* algorithm

6.2.1.1. *ulrb* algorithm description

The unsupervised learning method used was pam algorithm (Kaufman and Rousseeuw, 1987), based on k-medoids model (Vinod, 1969). In the context of *ulrb*, we apply the pam algorithm for a single feature, which is the abundance scores of taxa in a given sample. Thus, the result obtained in one sample is independent from the result obtained in another sample. The principle of the pam algorithm (Kaufman and Rousseeuw, 1987), in *ulrb,* is to divide all taxa into a predefined number of clusters (k), so that taxa within the same cluster are more similar with each other than what they are compared to taxa of other clusters. This is achieved by finding the centroids of clusters (medoids) and maximizing the objective function, which in this case minimizes the distance between taxa and their respective medoid. To do this, the algorithm randomly selects two candidate taxa as medoids, then it calculates the distance between them and all other taxa, attributing all taxa to the nearest medoid (Figure 6.1). Then, the algorithm enters in the swap phase, whereby the medoids are replaced and distances are calculated again (Figure 6.1). The swap phase is repeated until the total distances between taxa are minimized, and clusters are defined (Figure 6.1). For more detailed on the algorithm, we refer the reader to the *ulrb* package documentation (https://pascoalf.github.io/*ulrb*/), as well as to Kaufman and Rousseeuw (Kaufman and Rousseeuw, 1987) and *cluster* package documentation (Maechler *et al.*, 2023). Because *ulrb* explores one dimension of the abundance table (phylotype abundances in a sample), any data transformation will not change the relative distance between points for abundance classification, and thus the method works equally well for compositional and non-compositional data. For the same reason, for *ulrb*, it does not matter what kind of unit (*e.g.*, ASVs) is used as a proxy of taxa and it does not matter the abundance score used, *i.e.*, it does not matter the normalization and bioinformatics processing. However, note that the way a study defines taxa and calculates abundance scores will be relevant for the study design and research question. Thus, such considerations should be made before applying the *ulrb* method.

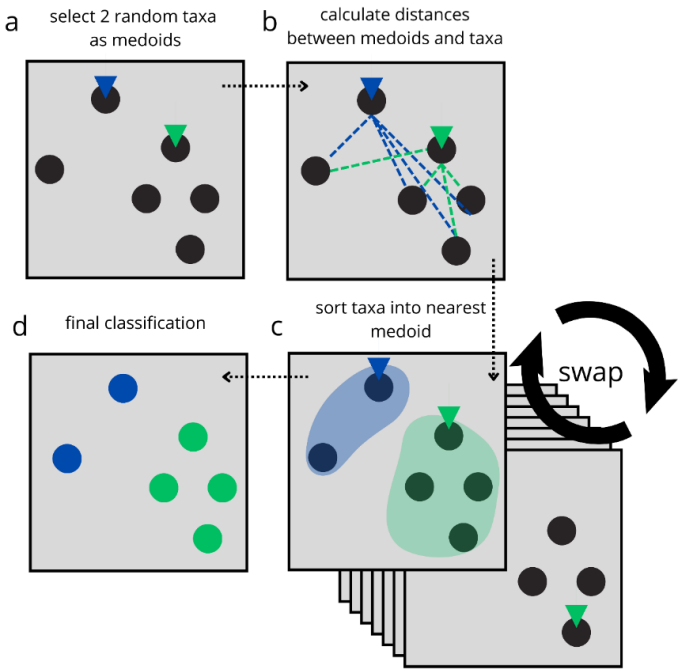


Figure 6.1. Schematic representation of k-medoids. This is a simplified illustration of k-medoids, using 2 dimensions (x and y axis), for two clusters (k = 2), represented by different colors.

6.2.1.2. *ulrb* R package construction and utilization

The *ulrb* R package was built using the functionalities of *devtools* (Wickham *et al.*, 2022). This package was built to include functions to prepare abundance tables, apply the pam algorithm, and helper functions to verify statistics and for visualization.

The main function in the *ulrb* package is called define_rb(), which will apply the *ulrb* method and automatically provide a classification of all taxa into “rare”, “undetermined” or “abundant”. The define_rb() function uses an abundance table as input. This table should include, at least, three columns, indicating the abundance, sample name and phylogenetic unit. Additional variables are allowed and unchanged by the function define_rb(). To apply the pam algorithm (Kaufman and Rousseeuw, 1987; Kaufman and Rousseeuw, 1990) we used the pam() function, from the *cluster* package (Maechler *et al.*, 2023). Besides the default parameters, it is possible to choose a specific number of abundance classifications, but in this case the user needs to manually name them. For example, if the user decides to use k = 4, then the abundance classifications will be named “1”, “2” and so on, but it is trivial to change those automatic names into user specified terms, *e.g.*, “very rare”, “rare”, “abundant”, and so on.

It is possible to automatically decide k in define_rb() function. For that purpose, we made an additional function, suggest_k(), which will calculate the best k possible, based on either the average Silhouette score (Rousseeuw, 1987), Davies-Bouldin index (Davies and Bouldin, 1979) or Calinski-Harabasz index (Calinski and Harabasz, 1974) (more details below). To calculate the average Silhouette score we used the pam() function from the *cluster* R package (Maechler *et al.*, 2023) and to calculate the Davies-Bouldin and Calinski-Harabasz indices we used the *clusterSim* package (Walesiak and Dudek, 2020). By default, suggest_k() will use the average Silhouette score. Independently of using default or user specified parameters, the define_rb() function will throw a warning for samples with low Silhouette scores. To do that, define_rb() identifies clusters, across all samples, where at least half the taxa correspond to a Silhouette score below 0.5. Even if this warning appears, the user can proceed normally, being aware that it might be possible to improve the clustering performance. However, the fact that a specific cluster got a bad average score does not imply that the structure of the entire clustering result is artificial. An artificial cluster is a cluster produced through a human method without prior assumptions on the data and that may have an unknown or currently unobservable meaning when looking at the properties of the data. We warn the user, however, that if different studies use different numbers of classifications, to accommodate the best Silhouette scores, then comparability is hindered.

The function suggest_k() provides the best k value for all samples used as input, by default. However, suggest_k() can alternatively return a detailed result, which provides a list with a report on the behavior of the three different indices (average Silhouette score, Davies-Bouldin and Calinski-Harabasz indices) across different values of k. The values of k that are tested by default range from 3 to 10. This range of k values can be changed, but more than 10 clusters might erode the purpose of using unsupervised learning methods to define the rare biosphere (and other domain-related abundance classifications, like “abundant”), because the more clusters there are, the less information they provide. For example, if a user applies the maximum k, then each taxon will be its own cluster, which does not provide new information. The user can use any range of allowed values of k, ranging from two up to the total number of different abundance score values in a given sample. The maximum value of k corresponds to the maximum number of different abundance scores in a sample. Note that if more than one sample is tested at the same time, then the maximum k will be the lowest maximum k across all samples tested. A tutorial is available on the *ulrb* R package website illustrating the impact of extreme k values on abundance classifications

(https://pascoalf.github.io/*ulrb*/articles/explore-classifications.html).

To help the users format their dataset for *ulrb* package functions, we provide the prepare_tidy_data() function, which can transform common abundance table formats into the required format. Specifically, taxa by rows, with samples as columns; or vice-versa.

Additional functions used within the major functions described in here were illustrated in the package tutorials, available online (https://pascoalf.github.io/*ulrb*/index.html).

6.2.1.3. Unsupervised learning statistics

The package *ulrb* includes three main statistics to evaluate the quality of the clustering, which are the average Silhouette score (Rousseeuw, 1987), Davies-Bouldin index (Davies and Bouldin, 1979) or Calinski-Harabasz index (Calinski and Harabasz, 1974). To evaluate the quality of the clusters obtained from *ulrb* results in this study, we relied on the Silhouette score (Rousseeuw, 1987). However, depending on the user needs, one of the other statistics might be more useful. Briefly, the average Silhouette score measures cluster definition and separation, the Calinski-Harabasz index measures cluster separation and density, and Davies-Bouldin measures cluster separation. Below, we describe the Silhouette score in more detail, because it was the index used to evaluate the results presented here. For more details on Calinski-Harabasz and Davies-Bouldin indices, see Supplementary Methods (section 6.7).

6.2.1.3.1. Silhouette score

The Silhouette score calculates how close a taxon is to its own cluster relative to the next closest cluster. The Silhouette score of a given taxon, $S(i)$, is given by Equation 6.1,

$$S(i) = \frac{(b - a)}{\max(a, b)}$$

Equation 6.1. Silhouette score formula.

where *a* is the mean distance between the *i*th taxon and all other taxa on the same cluster, and *b* is the mean distance between all taxa in the cluster of the *i*th taxa and the centroid of the next closest cluster. It follows that $-1 \leq S(i) \leq +1$. By convention, $S(i) = 0$ means that the *i*th taxon is as close to its own cluster as it is to the next closest cluster; $S(i) = -1$ means that the *i*th taxon is better positioned in the next closest cluster, instead of its own cluster; and $S(i) = +1$ means that the *i*th taxon is in the center of its own cluster (Rousseeuw, 1987). Note that a perfect score might indicate an artificial cluster in the

case of an outlier group (Rousseeuw, 1987), but we address this issue in the Discussion (section 6.4) and accept clusters of outliers as valid.

Based on Kaufman and Rousseeuw (Kaufman and Rousseeuw, 1990), we interpreted the average Silhouette score as: > 0.71 strong cluster; > 0.51 reasonable cluster; ≥ 0.26 weak cluster; and values below 0.26 indicate a potentially artificial cluster.

The Silhouette score is calculated for each taxon, but it can provide information on a specific cluster or all clusters (Figure 6.2). Thus, the average Silhouette score of all clusters provides a statistic of quality of the clustering method, which is comparable with other methods.

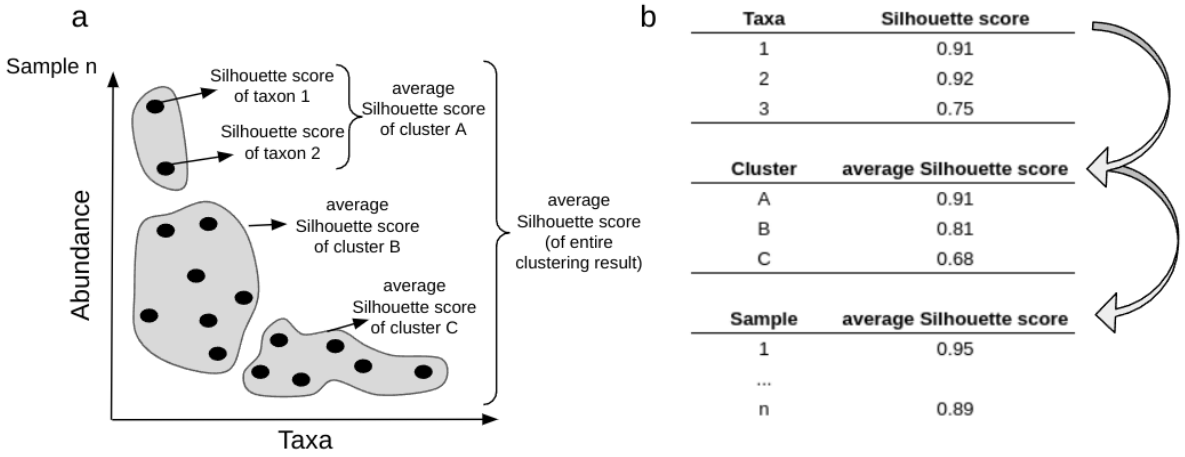


Figure 6.2. Schematic representation of the information that the Silhouette scores can provide. **a**. Visual representation of taxa grouped in three different clusters, each taxon with a Silhouette score. **b**. The average Silhouette scores of all taxa within a cluster can be calculated to deliver a single quality score regarding that specific cluster, or the Silhouette scores of the taxa from all clusters can also be calculated, to obtain a single average Silhouette score for the entire sample. The values shown are arbitrary, for the purpose of illustrating the possible ways of grouping the data.

6.2.2. Datasets

We used a dataset presented for the first time in this article (Environmental Monitoring of Svalbard and Jan Mayer, MOSJ 2016-2020), but we also used publicly available datasets to strengthen the analysis and cover a representative range of methodologies: Norwegian Young Sea Ice Expedition (N-ICE), MOSJ 2019, Ants, Barro Colorado Island Tree Counts (BCI), and coral microbiome. A summary of selected datasets and their major features is available in Table 6.1 (see also Data Availability in section 6.6). In the following subsections (6.2.2.1, 6.2.2.2, 6.2.2.3, 5.2.2.4, and 6.2.2.5), we provide a short description of the previously published datasets, with additional details on

Supplementary Methods (section 6.7), and we provide full details on the MOSJ 2016-2020 dataset (section 6.2.2.6).

Table 6.1. Summary of datasets used in this study. This table includes information on the relevant methodology used, it informs whether the dataset is microbial or not, the kind of phylogenetic unit used, the purpose of the dataset in this study, and the number of samples used.

Dataset name	Method	Microbial	Phylogenetic unit	Purpose in this study	Samples used	References
N-ICE	Amplicon sequencing of V4V5 16S rRNA gene	Yes	ASV	Validate *ulrb* for different kinds of phylogenetic units	9	(Granskog *et al.*, 2016; de Sousa *et al.*, 2019; Pascoal *et al.*, 2022a)
N-ICE	Amplicon sequencing of V4V5 16S rRNA gene	Yes	OTU		9	
N-ICE	Shotgun metagenome sequencing	Yes	mOTU		9	
MOSJ 2019	Amplicon sequencing of V4V5 16S rRNA gene	Yes	ASV	Validate *ulrb* for different amplicon sequencing approaches	6	(Pascoal *et al.*, 2024)
MOSJ 2019	Amplicon sequencing of full-length 16S rRNA gene	Yes	ASV		6	
MOSJ 2016-2020	Amplicon sequencing of V4V5 16S rRNA gene	Yes	ASV	Evaluate *ulrb* for different sample size, sequencing depth and number of ASV	117	This study.
Coral microbiome	Shotgun metagenome sequencing	Yes	mOTU	Identify types of rarity	20	(Keller-Costa *et al.*, 2021)
Ants	Counts	No	Species		98	(Arnan, Gaucherel and Andersen, 2011;

				Validate *ulrb* for non-microbial data		Calatayud *et al.*, 2019; Balbuena *et al.*, 2021)
BCI	Counts	No	Species		18	(Condit *et al.*, 2019)
		Total			292	

End of Table 6.1.

6.2.2.1. N-ICE dataset

The N-ICE dataset is composed of samples collected north of Svalbard in 2015 (Granskog *et al.*, 2016, 2018), which were used for V4V5 16S rRNA gene amplicon sequencing and shotgun metagenomic sequencing (de Sousa *et al.*, 2019). The sequencing results were previously processed, using distinct bioinformatics approaches (Pascoal *et al*., 2022a), resulting in ASVs (n = 9 samples), OTUs (n = 9 samples), and mOTUs (n = 9 samples). For extended details on sampling, sequencing and bioinformatics processing, see Supplementary Methods (section 6.7). A summary of sequencing statistics of N-ICE is available in Supplementary Table 11.19.

6.2.2.2. MOSJ 2019 dataset

The MOSJ 2019 dataset is composed of samples collected during an expedition in Svalbard, in the framework of the Environmental Monitoring of Svalbard and Jan Mayer (Renner, Dodd and Fransson, 2018) (MOSJ) in 2019. Samples were collected for two different amplicon sequencing approaches (Pascoal *et al.*, 2024) (Chapter 4), specifically: V4V5 16S rRNA gene amplicon sequencing, with Illumina technology, and full-length 16S rRNA gene amplicon sequencing, with CCS PacBio technology. Initially, there were 18 samples available per amplicon sequencing strategy, but after filtering for the samples with high quality in both sequencing strategies, this number was reduced to 6 samples, each. For extended details sequencing and bioinformatics processing, see Supplementary Methods (section 6.7). Sequencing statistics were summarized in Supplementary Table 11.20.

6.2.2.3. Coral microbiome dataset

The coral microbiome dataset comprises samples characterized by shotgun metagenomic sequencing to describe coral host associations (Keller-Costa *et al.*, 2021). Specifically, samples were collected within the coral tissue (n = 13 samples), and in the sediment (n = 3 samples) and seawater (n = 4 samples) surrounding the corals. The corals selected are within the group of octocorals and include the species *Eunicella gazella* (n = 3 samples of healthy tissue, and n = 3 of necrotic tissue), *Eunicella verrucosa* (n = 4 samples of healthy tissue), and *Leptogorgia sarmentosa* (n = 3 samples). The 16S rRNA gene reads from the shotgun metagenomic dataset included 93 589 high-quality reads and 1 041 mOTUs defined at a 97% similarity cut-off (Keller-Costa *et al.*, 2021).

For extended details on sampling, sequencing and bioinformatics processing, see Supplementary Methods (section 6.7).

6.2.2.4. Ants dataset

The Ants dataset includes 49 different species surveyed at 99 sites (Arnan, Gaucherel and Andersen, 2011). This dataset was made available in the *FuzzyQ* R package (Balbuena *et al.*, 2021) and originally described by (Arnan, Gaucherel and Andersen, 2011; Calatayud *et al.*, 2019). For this study, a site is equivalent to a sample. Prior to analysis with *ulrb*, one sample was removed from the Ants dataset (site 95) because of low sampling effort.

6.2.2.5. BCI dataset

The BCI dataset is publicly available and was described by Condit *et al*. (2002, 2019). The BCI dataset used 50 plots of 1 hectare, surveyed over 35 years. For this study, a subset of the full BCI census was used to make a species abundance table, filtering alive trees and counting the number of species found in each combination of plot and year of survey (sample for our purpose). Then, we filtered samples with, at least, more than two tree species. Our final species abundance table, derived from a BCI subset, includes 327 tree species and 18 samples in a species abundance table.

6.2.2.6. MOSJ 2016-2020 dataset

6.2.2.6.1. Sampling and sequencing details

Microbiome samples were collected from 2016 to 2020 (n = 119 samples) in a standardized way (Kopf *et al.*, 2015) in the framework of the MOSJ (Renner, Dodd and Fransson, 2018). Every year, during the summer season, the MOSJ campaign collects samples at several stations from the Kongsfjorden transect, covering the epipelagic, mesopelagic and bathypelagic layers. Details on sampling coordinates and depth for the samples that were used are available in Supplementary Table 11.21.

Seawater was filtered (0.2 L to 7.5 L) through cartridge filters (0.22 µm pore size; Sterivex units) and DNA was extracted following the DNeasy PowerWater Sterivex Kit (Qiagen) and best practices from OSD (Kopf *et al.*, 2015). For the amplification of V4V5 16S rRNA gene, the primers 515YF (5′-GTGYCAGCMGCCGCGGTAA-3′) and 926R (5′ - CCGYCAATTYMTTTRAGTTT- 3′) (Caporaso et al., 2011; 2012; Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016) were used. Sequencing was performed with

Illumina technology, on MiSeq platforms (2x300bp). This is the first study to integrate all 119 samples from MOSJ2016-2020 into a single dataset.

6.2.2.6.2. Processing of V4V5 16S rRNA gene amplicons

To produce ASVs from V4V5 16S rRNA gene sequencing, we used a bioinformatic protocol based on *DADA2* R package (Callahan *et al.*, 2016). Reads were trimmed at 249 nt (Forward) and 214 nt (Reverse) based on quality profiles of the entire MOSJ dataset (2016 to 2020) (Supplementary Figure 10.28). Thus, the trimming criteria were the same for all years, as a compromise to allow standardization of ASV creation. Default parameters were used for the remaining steps of *DADA2* (Callahan *et al.*, 2016), which include the creation of an error model for quality filtering, identification of ASVs (*i.e.*, the unique sequences), chimera removal and taxonomic assignment with Naive-Bayesian algorithm (Wang *et al.*, 2007) and the Silva v138 database (Quast *et al.*, 2012; Yilmaz *et al.*, 2014).

ASV tables were filtered to remove taxa attributed to unknown domain-level classifications, eukaryotes and organelles, if any; and singletons were removed, if any. ASV tables were rarefied at several rarefaction levels, considering the rarefaction curves (Supplementary Figure 10.29), always discarding samples below the rarefaction threshold (n = 117 samples after this step).

For a summary of raw read processing statistics, see Supplementary Table 11.22.

6.2.2.6.3. Combining sequencing data from different years (2016-2020) into a single dataset

The integration of the data from different years into a single dataset was possible, because an effort was made to standardize the five years as much as possible. Specifically, the OSD protocol was used for seawater sampling (Kopf *et al.*, 2015), with small variation in filtration volume (mean = 2.9 L, sd = 1.4 L and n = 117 samples, Supplementary Table 11.21). Based on previous work, variable filtration volume does not constitute a confounding variable (Pascoal *et al.*, 2023) (Chapter 3). The processing of raw reads into amplicon sequence variants was also performed using the same protocol for all years (described in detail in the previous section), which was optimized in a previous study (Pascoal *et al.*, 2024) (Chapter 4). The best phylogenetic unit to integrate different datasets into one are the ASVs, because they are sequences with explicit biological meaning (Callahan *et al.*, 2016, 2019).

6.2.3. Types of rarity

Depending on how the abundance classification changes, taxa can be grouped in types of rarity (Lynch and Neufeld, 2015). For example, if one taxon oscillates between being rare and abundant, it can be considered conditionally rare (Shade *et al.*, 2014). The current version of *ulrb* does not allow for the automatic calculation of the types of rarity. However, once the *ulrb* classification is obtained ("rare", "undetermined" and "abundant" classifications), it is possible to manually inspect how specific taxa change their classification across some variable.

6.2.4. Statistical analysis, visualization and reproducibility

All statistical analyses and plots were produced using R software (R Core Team, 2023). Several plots used the package *ulrb* (presented in here) together with *ggplot2* (Wickham, 2016) and *gridExtra* (Auguie, 2017). Rarefaction was done using the rrarefy() function from the *vegan* R package (Oksanen *et al.*, 2025), to standardize the total number of reads per sample. When necessary, centrality metrics were used to avoid overlapping of samples in the plots. The centrality metric used was the mean ± standard deviation (sd), with the number of samples indicated in the figure legend. For an alternative unsupervised approach to classify the rare biosphere, we used the *FuzzyQ* R package (Balbuena *et al.*, 2021), which also calculates Silhouette scores for statistical evaluation of results. For reproducibility, all source data and code are publicly available (see Source Code and Data, and Data Availability statements in section 6.6).

6.3. Results

6.3.1. Testing *ulrb* for different kinds of phylogenetic units

To test *ulrb* applicability across common kinds of phylogenetic units, we used the N-ICE dataset. Briefly, we considered ASVs (n = 9 samples) and pre-computed OTUs (OTUs, n = 9 samples) from V4V5 16S rRNA gene amplicon sequencing, and metagenomic derived OTUs (mOTUs) (n = 9 samples) from full-length 16S rRNA genes obtained via shotgun metagenomic sequencing. For ASVs, OTUs, and mOTUs, *ulrb* provided a RAC description of the microbial communities consistent with the classical view of the rare biosphere as the long tail of the RAC (Figure 6.3a), showing that it can be used for either kind of phylogenetic unit.

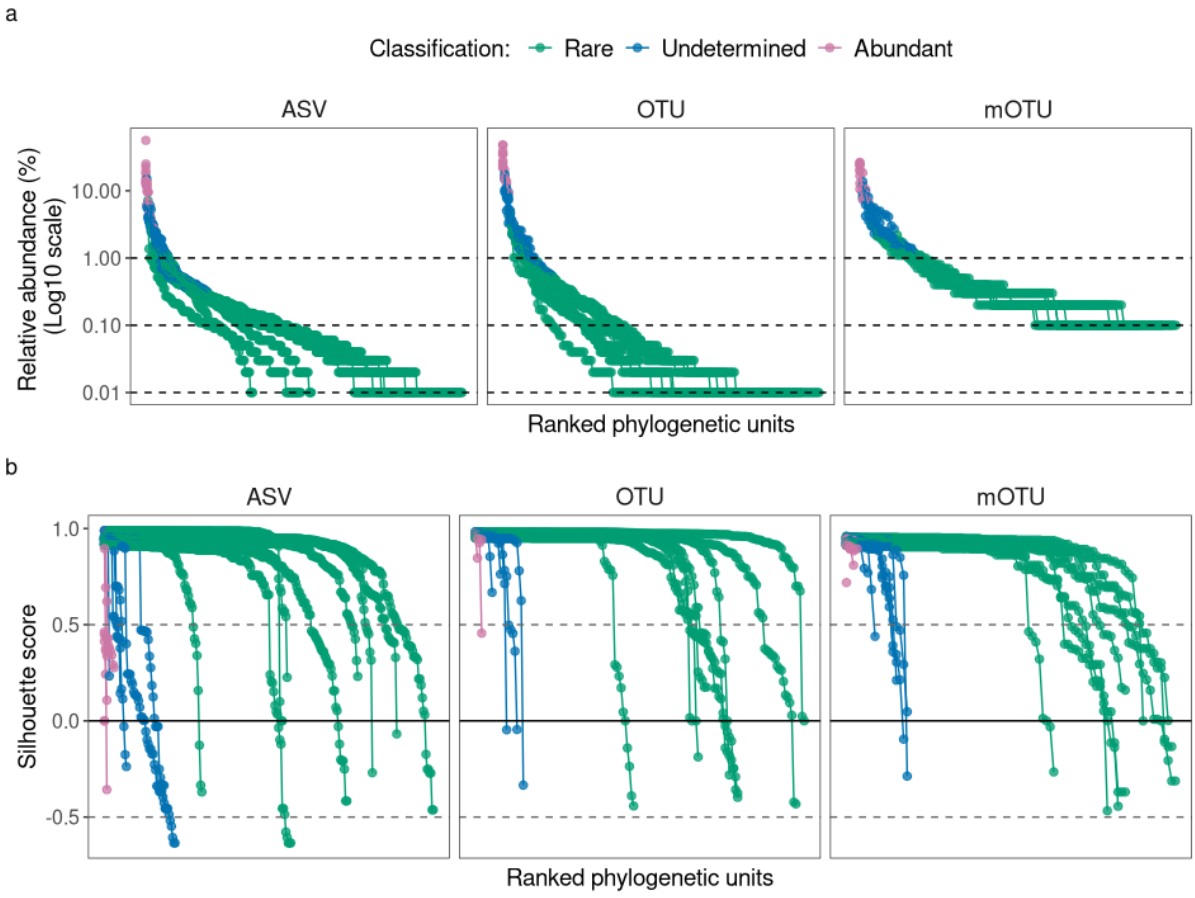


Figure 6.3. RAC and Silhouette score plots for N-ICE dataset. **a**. Relative abundance of each phylogenetic unit (either ASV, OTU or mOTU) from all samples. For each sample, phylogenetic units were ordered in the x axis from the most to the least abundant. The y axis shows the abundance score of each phylogenetic unit in Log10 scale. For context, three relative abundance thresholds are highlighted with dashed lines (0.01%, 0.10% and 1.00%). **b**. Silhouette scores obtained for each phylogenetic unit from all samples. Phylogenetic units are ordered from highest to lowest Silhouette score. For (a) and (b), lines were used to group phylogenetic units in the same sample and abundance classification. Default classifications ("rare", "undetermined" and "abundant") were colored coded.

To determine the statistical support of the unsupervised learning results, we calculated Silhouette scores for the datasets obtained with each phylogenetic unit (see 6.2.1.3.1.). The Silhouette scores were higher for OTUs and mOTUs than ASVs (Figure 6.3b), meaning that clustering of phylogenetic units into abundance classifications by *ulrb* was overall more robust for OTUs and mOTUs than for ASVs. More than 75% of OTUs and mOTUs formed strong or reasonable clusters (Supplementary Figure 10.30). However, 58% of the abundant ASVs formed weak clusters (Supplementary Figure 10.30). OTUs formed strong clusters in all samples; the mOTUs formed either strong or reasonable clusters in all samples; and ASVs formed strong or reasonable clusters for “rare” and “undetermined” classifications, except for one sample (Supplementary Figure 10.30). Although some phylogenetic units and clusters had lower Silhouette scores, the average Silhouette score indicated that the clustering structure across the entire dataset was strong or reasonable in all samples. This was consistent for all tested phylogenetic units, including ASVs (Supplementary Figure 10.30). The fully automatic alternative of *ulrb* selected three clusters for ASVs and OTUs, but four clusters for mOTUs. Thus, based on average Silhouette scores (default settings), the ASV clustering could not be improved any further by using any other value of k. One possible reason why abundant ASVs were more difficult to cluster with *ulrb*, might be that it included more different abundance values than OTUs (216 ASVs *vs* 192 OTUs) and more extreme values (ASV maximum abundance = 5 613 reads; OTU maximum abundance = 4 825 reads). Note that this comparison refers to the *ulrb* statistical robustness against using different phylogenetic units (OTUs, ASVs, and mOTUs). It does not, however, imply any recommendation regarding which phylogenetic unit should be used in specific studies.

To verify if *ulrb* can provide consistent results across methods, we compared the alpha diversity (number of ASVs/OTUs/mOTUs) within each classification obtained (Figure 6.4). Results showed a consistent trend for all phylogenetic units tested, revealing, in all cases, that the rare biosphere consisted of a larger richness of phylogenetic units than that of undetermined or abundant phylogenetic units (Figure 6.4). Thus, *ulrb* reflected the shape of the RAC with better consistency than threshold-based definitions, which presented distinct patterns for each phylogenetic unit approach (Figure 6.4). Note that the absolute values of the response variable (number of ASVs/OTUs/mOTUs) are different, because the methodology is different, but they are consistent in that they have the same relationship. Thus, when using *ulrb*, the definition of rarity (and, by extension, the definition of “abundant” and “undetermined” classifications) had the same meaning, *i.e.*, the same interpretation, across phylogenetic units.

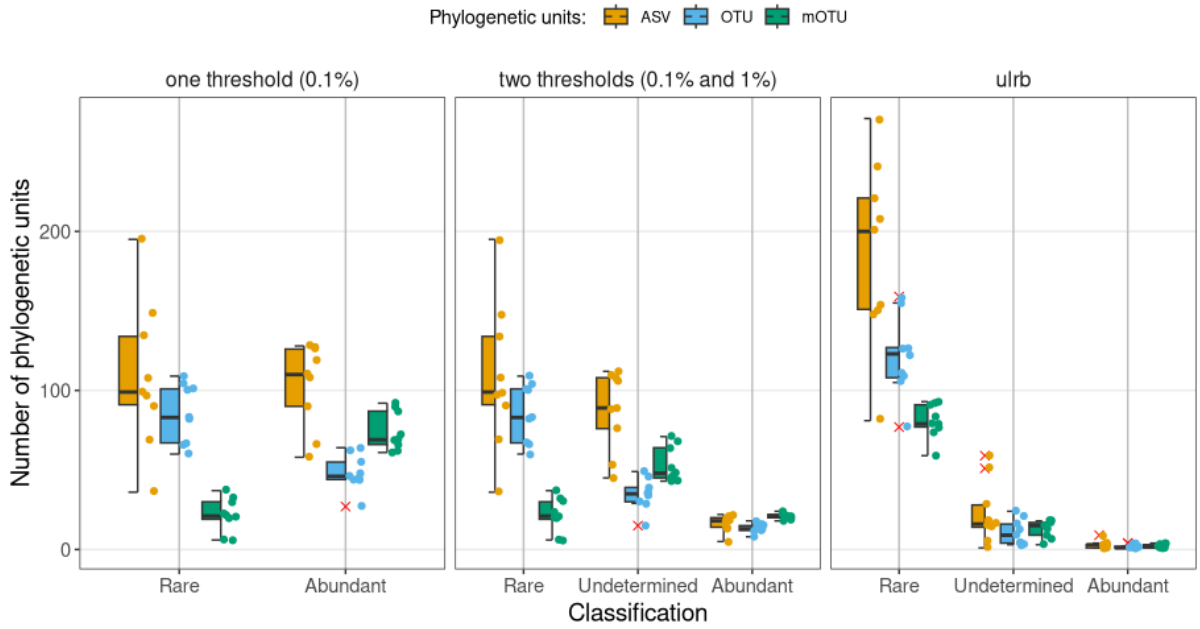


Figure 6.4. Comparison of number of rare, undetermined (if applicable) and abundant phylogenetic units from ASVs, OTUs, and mOTUs. Abundance classifications were estimated with different methods: one single relative abundance threshold (0.1%, per sample), two relative abundance thresholds (0.1% and 1%, per sample) and *ulrb* using default parameters (k = 3). The number of ASVs was illustrated with boxplots and points for each sample (n = 6 samples). The outliers were marked with red cross.

For perspective, we applied the same analysis using an alternative unsupervised learning approach to define the rare biosphere, using *FuzzyQ* (Supplementary Figure 10.30). The *FuzzyQ* method worked as expected (Supplementary Figure 10.31), presenting generally good quality clusters (Supplementary Figure 10.31). Similar to *ulrb*, the phylogenetic unit with worse quality clusters was the ASVs (lower Silhouette scores, Supplementary Figure 10.31).

6.3.2. Testing *ulrb* for different amplicon sequencing strategies

To test *ulrb* applicability to different amplicon sequencing strategies, we used the MOSJ 2019 dataset, which includes samples from short-reads (V4V5 region of the 16S gene, n = 6 samples), and long-reads (full-length 16S rRNA gene, n = 6 samples). For either sequencing approach, *ulrb* was able to characterize the classical RAC in a way that shows a long-tail of rare ASVs, followed by an intermediate region of undetermined ASVs and a few ASVs with very high abundance (Figure 6.5a).

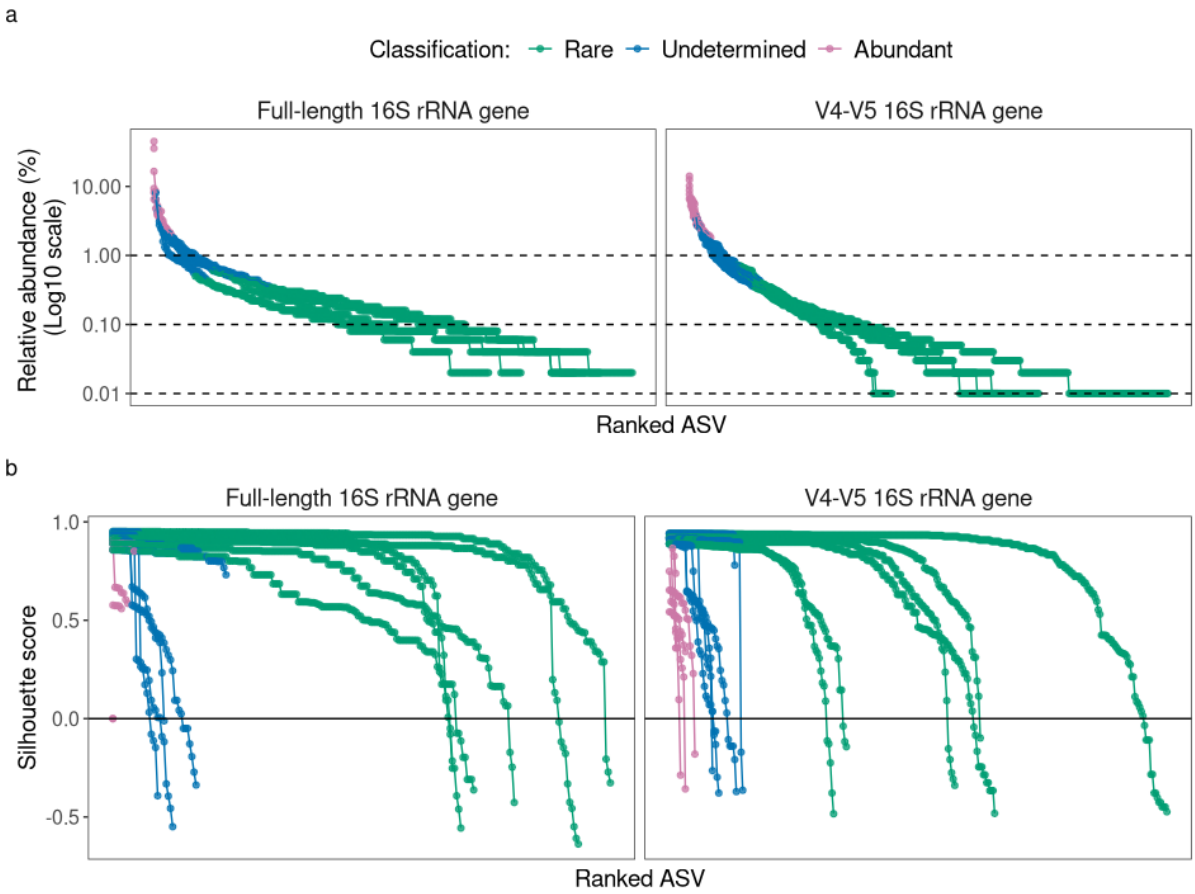


Figure 6.5. Comparison of ASVs derived from V4V5 and full-length 16S rRNA gene. **a**. relative abundance of each ASV in Log10 scale, ordered in the x axis from the most to the least abundant. For context, three relative abundance thresholds are highlighted with dashed lines (0.01%, 0.10% and 1.00%). **b**. Silhouette score of each ASV. For both (a) and (b), lines were used to group ASVS in the same sample and abundance classification. ASVs were colored coded by abundance classification ("rare", "undetermined" and "abundant").

In terms of statistical quality of the unsupervised learning results, the Silhouette scores below 0.5 were more often attributed to abundant and undetermined than to rare ASVs, especially in the analysis of the full-length 16S rRNA gene approach (Figure 6.5b). More than 75% of the rare and abundant ASVs from the full-length 16S rRNA gene approach formed strong or reasonable clusters, in contrast with the undetermined ASVs (with up

to 34.5% weak and potentially artificial clustering) (Supplementary Figure 10.32). For the V4V5 16S rRNA gene approach, abundant ASVs presented the weakest Silhouette scores (Supplementary Figure 10.32b). For the full-length 16S rRNA gene approach, half of samples got either strong or reasonable clusters for any abundance classification (Supplementary Figure 10.32). Regarding the V4V5 rRNA gene approach, half of the samples displayed weak clustering for the “abundant” classifications (Supplementary Figure 10.32). When all clusters were considered, both approaches (V4V5 and full-length 16S rRNA gene sequencing) had strong or reasonable clustering results (Supplementary Figure 10.32). Thus, the average Silhouette score never fell below 0.5, which means that the clusters found were not artificial. We attempted an improvement by using the automatic option of *ulrb*, but the automatic result (relying on average Silhouette scores) also selected three clusters.

We compared the consistency of different definitions of rarity between V4V5 and full-length 16S rRNA gene sequencing (Figure 6.6). The most common approach to delineate the rare biosphere (0.1% relative abundance, per sample) resulted in a higher number of rare ASVs than abundant ASVs with the V4V5 region of the 16S rRNA gene, but the opposite was observed for the full-length 16S rRNA gene (Figure 6.6). Using two thresholds also resulted in different patterns, with the number of ASVs going up and down from rare to undetermined to abundant for the full-length 16S rRNA gene, but always decreasing when the V4V5 region of the 16S rRNA gene was used (Figure 6.6). Finally, the *ulrb* approach was the only one to provide the same pattern with both molecular methods, showing in each case a clearly higher richness of ASVs classified as rare than undetermined or abundant (Figure 6.6). Thus, *ulrb* was able to provide a consistent definition of rarity between the two sequencing strategies, while the other two definitions, relying on relative abundance thresholds, failed to do so.

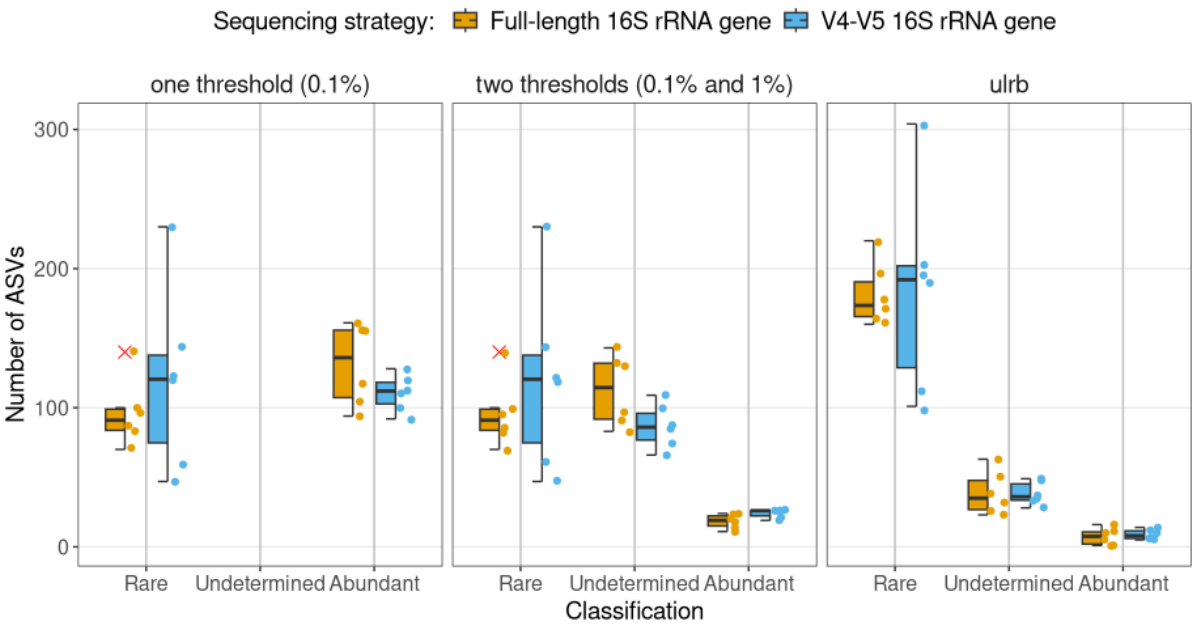


Figure 6.6. Comparison of number of rare, undetermined (if applicable) and abundant ASVs from V4V5 and full-length 16S rRNA gene sequencing. Abundance classifications were estimated with different methods: one single relative abundance threshold (0.1%, per sample), two relative abundance thresholds (0.1% and 1%, per sample) and *ulrb* using default parameters (k = 3). The number of ASVs was illustrated with boxplots and points for each sample (n = 6 samples). There were no outliers.

We tested the applicability of *FuzzyQ* in comparing amplicon V4V5 with full-length 16S rRNA gene sequencing (Supplementary Figure 10.33). The method was able to classify taxa into common and rare but using three classifications (instead of two) would have been better for the full-length 16S rRNA gene sequencing data, because some ASVs were grouped near the threshold of 0.5 commonality index (Supplementary Figure 10.33). Regardless of the number of clusters, the clustering quality was good for both sequencing strategies, except for a few common ASVs obtained from V4V5 16S rRNA gene sequencing approach (Supplementary Figure 10.33).

6.3.3. Verifying robustness of *ulrb* against sample size, sequencing depth and number of taxa

To verify the robustness of *ulrb* we used the MOSJ2016-2020 dataset, which includes up to 117 high-quality samples (after quality filtering). We tested the quality of clustering (measured by average Silhouette score) as a function of three variables that distinguish datasets: 1) number of samples (n); 2) number of taxa (ASVs in this context); and 3) sequencing depth, per sample.

To test the effect of sample size (n), we locked the sequencing depth at 10 000 reads (using rarefaction), resulting in a total pool of 114 high-quality samples. Then, we subsampled n random samples, without replacement, from the pool of 114 high-quality samples. At each step (from n = 6 to n = 114), we applied *ulrb* to all the samples and then calculated the average Silhouette score of each sample and plotted the mean ± sd (Figure 6.7a). Results showed that *ulrb* provided high quality clustering (average Silhouette scores > 0.75) for the rare biosphere with low (n < 30) and high (n > 30) sample size. The “undetermined” and “abundant” classifications, similarly to the previous sections (Figure 6.3 and Figure 6.5), presented lower quality. However, the “undetermined” classification presented mostly reasonable clusters, and the “abundant” classification varied between weak and reasonable clusters (Figure 5.7a). Importantly, the average Silhouette scores presented more random variation at the “undetermined” and “abundant” classification at low sample size (n < 30) than at large sample size (n > 30). In fact, above 30 samples, *ulrb* results were very robust for all abundance classifications (Figure 6.7a).

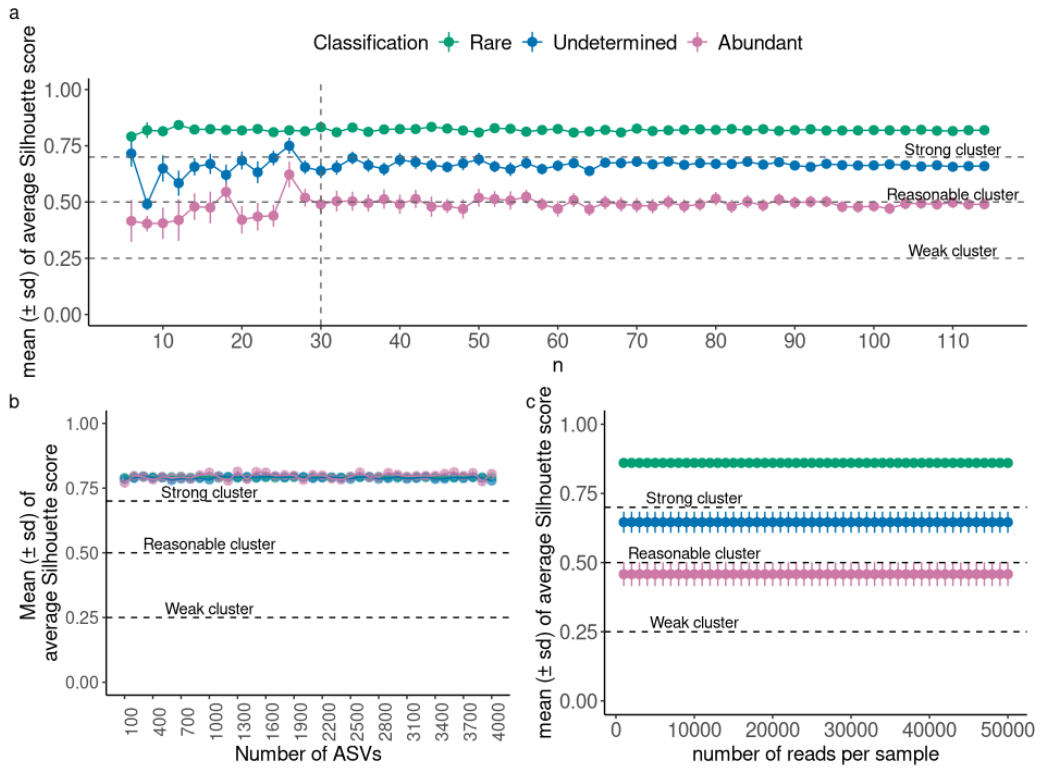


Figure 6.7. Quality of *ulrb* clustering measured by the average Silhouette score as a function of number of samples, ASVs, and sequencing depth. **a**. Number of samples (n), ranging from 6 samples up to 114 samples, in increments of two samples; **b**. Number of ASVs, ranging from 100 to 4000 taxa per sample, in increments of 300 taxa; **c**. Number of reads per sample, in increments of 1000 reads. The samples (a), ASVs (b) and reads (c) were randomly selected in each increment, without replacement. For the response variable, the mean (± sd) of the average Silhouette score was used and the classifications were grouped by different colors, as illustrated in the figure legend. The MOSJ2016-2020 dataset was used for this figure.

To test the robustness of *ulrb* against different number of taxa (ASVs in this context), we selected 34 samples and rarefied them to 50 000 reads, to have as many ASVs as possible and at least n > 30 samples. Then, we collected random ASVs (from 100 ASVs to up to 4000 ASVs) per sample, without replacement. Figure 6.7b shows that, for this set of samples, all abundance classifications obtained very good scores (average Silhouette score > 0.75). Importantly, the number of ASVs clearly had no effect on the quality of the clustering obtained by *ulrb*. To show that the random selection of ASVs was able to keep the RAC shape and was not exclusively obtaining ASVs of one single abundance classification, we illustrate the RAC obtained by a random selection of 100, 1 000 and 3 000 ASVs in a random sample (Supplementary Figure 10.34).

To test the impact of sequencing depth, we selected the 34 samples with more reads and applied different rarefaction levels to them (from 1 000 reads to up to 50 000 reads). Remarkably, *ulrb* was extremely robust for variations in sequencing depth, since the average Silhouette score was almost perfectly constant as a function of sequencing

depth (Figure 6.6c). As in the sample size analysis, the “rare” classification presented better quality than the “undetermined” and “abundant” classifications (Figure 5.6c).

In summary, by applying variations in specific features of a large dataset, we showed that *ulrb* presented robust results for variations in sample size, number of taxa (ASVs in here) and sequencing depth. Since any abundance table will ultimately vary because of a combination of different numbers of taxa, samples and order of magnitude of the abundance score, we present evidence that *ulrb* is robust for virtually any abundance table (Figure 6.6).

Finally, we verified the impact of the same variables on the application of *FuzzyQ* and found that this method generally presented high quality clustering for the “rare” classification, but potentially artificial clusters for the “common” classification (Supplementary Figure 10.35). However, *FuzzyQ* results improved for larger datasets (n > 30 and ASVs > 700) and were not limited by sequencing depth (Supplementary Figure 10.35).

6.3.4. Validating *ulrb* for non-microbial datasets

To test if *ulrb* can be applicable for non-microbiome datasets, we applied *ulrb* to animal and plant datasets that were publicly available, the Ants (Arnan, Gaucherel and Andersen, 2011; Balbuena *et al.*, 2021) and the BCI (Condit *et al.*, 2002, 2019) datasets. *ulrb* was able to classify all ant species into abundance categories, depicting a few species that were abundant, undetermined or rare in different samples and a long tail of rare species (Figure 5.8a). The clustering quality was also good for most species (average Silhouette score > 0.75), with very few species in low quality clusters (Figure 5.8b).

Similarly, *ulrb* was able to classify rare, undetermined and abundant tree species in the BCI dataset (Figure 6.8c,d). Furthermore, the classifications obtained showed a reasonable division between abundance scores, illustrating the applicability of *ulrb* (Figure 6.8a). The Silhouette plot reveals that *ulrb* provided robust classifications for most species, with only a few presenting low average Silhouette scores (Figure 5.8d).

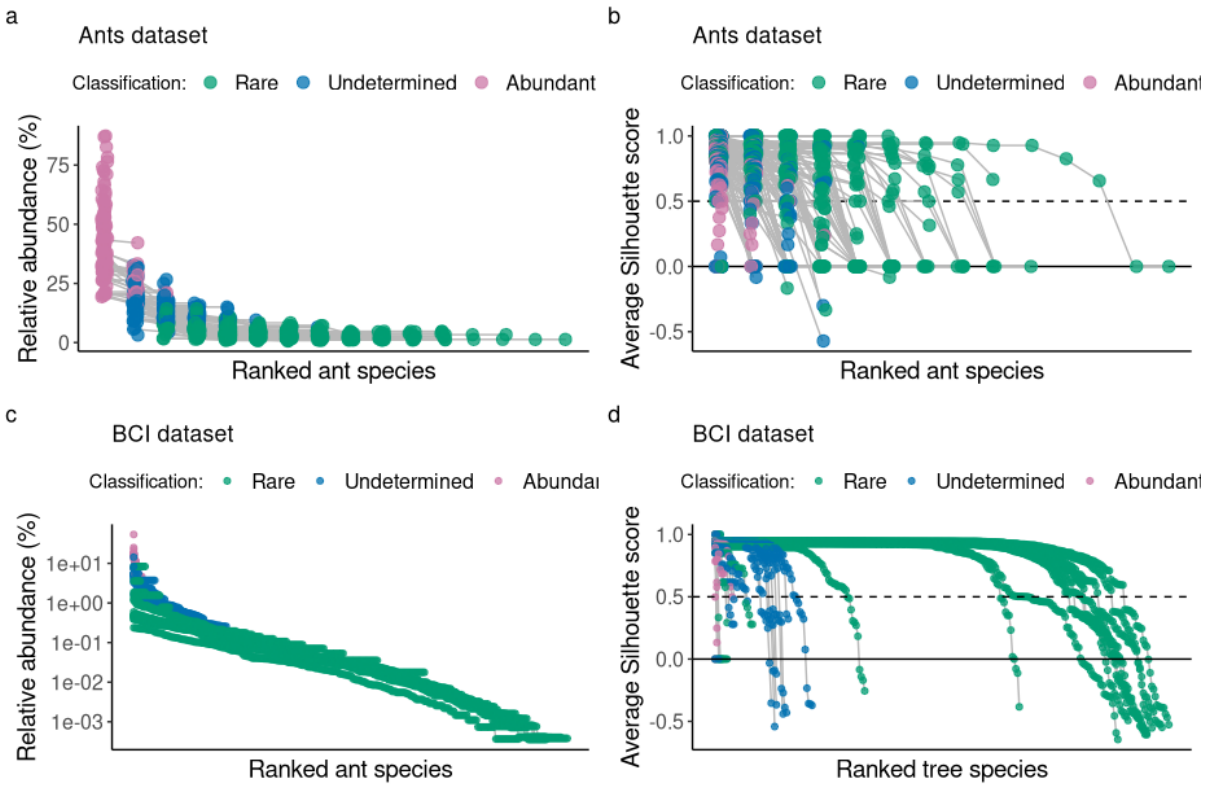


Figure 6.8. Analysis of *ulrb* applicability to Ants and BCI datasets. **a**. Mean ± sd of relative abundance of ranked species from the Ants dataset (n = 98 samples), colored by classification; **b**. Mean ± sd of average Silhouette score of ranked species from the Ants dataset (n = 98 samples), colored by classification; **c**. Mean ± sd of relative abundance (in Log10 scale) of ranked species from the BCI dataset (n = 18 samples), colored by classification; **d**. Mean ± sd of average Silhouette score of ranked species from the BCI dataset (n = 18 samples), colored by classification.

6.3.5. Using *ulrb* to establish types of rarity in a coral microbiome

To show that *ulrb* can be used to monitor taxa and, therefore, describe types of rarity, we used a coral microbiome from a shotgun metagenomic sequencing dataset (Keller-Costa *et al.*, 2021). This dataset was selected because the original experimental design was adequate to identify different roles of prokaryotes from the coral microbiome, which can be used to identify different types of rarity. Specifically, we monitored the classifications obtained for a selected group of mOTUs (561, 559 and 866, based on (Keller-Costa *et al.*, 2021)) across different coral species, health status and surrounding environment, in a way that effectively described different types of rarity (Figure 6.9). The mOTUS (561, 559, and 866) were selected based on previous knowledge of their ecology and adequacy to describe types of rarity. OTU 561 (genus *Anaerospora*) was absent in healthy coral tissue and in sediment but was a member of the seawater rare biosphere and colonized the necrotic coral tissue, becoming rare or undetermined (Figure 5.9). Thus, *ulrb* helped identify OTU 561 as a potential necrotic tissue colonizer and established its origin in the seawater rare biosphere. Another example of a necrotic tissue colonizer was OTU 559 (family *Rhodobacteraceae*), which was abundant in necrotic tissue and seawater, but rare or undetermined in healthy octocoral tissue. This result indicates that specific, abundant members of the seawater microbiome (in this case, a *Rhodobacterales* phylotype) may belong to the rare biosphere of healthy, host-associated microbiomes and rapidly colonize decaying host tissue, transitioning from rare to abundant while the symbiotic microbiome enters the dysbiosis state (Figure 5.9). A contrary example is OTU 866 (family *Endozoicomonadaceae*), which was abundant in healthy coral tissues, but became rare or undetermined under necrosis (except for EG18_N) and was rare or absent in the sediment and seawater samples (Figure 6.9). Thus, *ulrb* indicates that this phylotype in the family *Endozoicomonadaceae* represents a coral symbiont enriched in healthy while depleted in necrotic tissues.

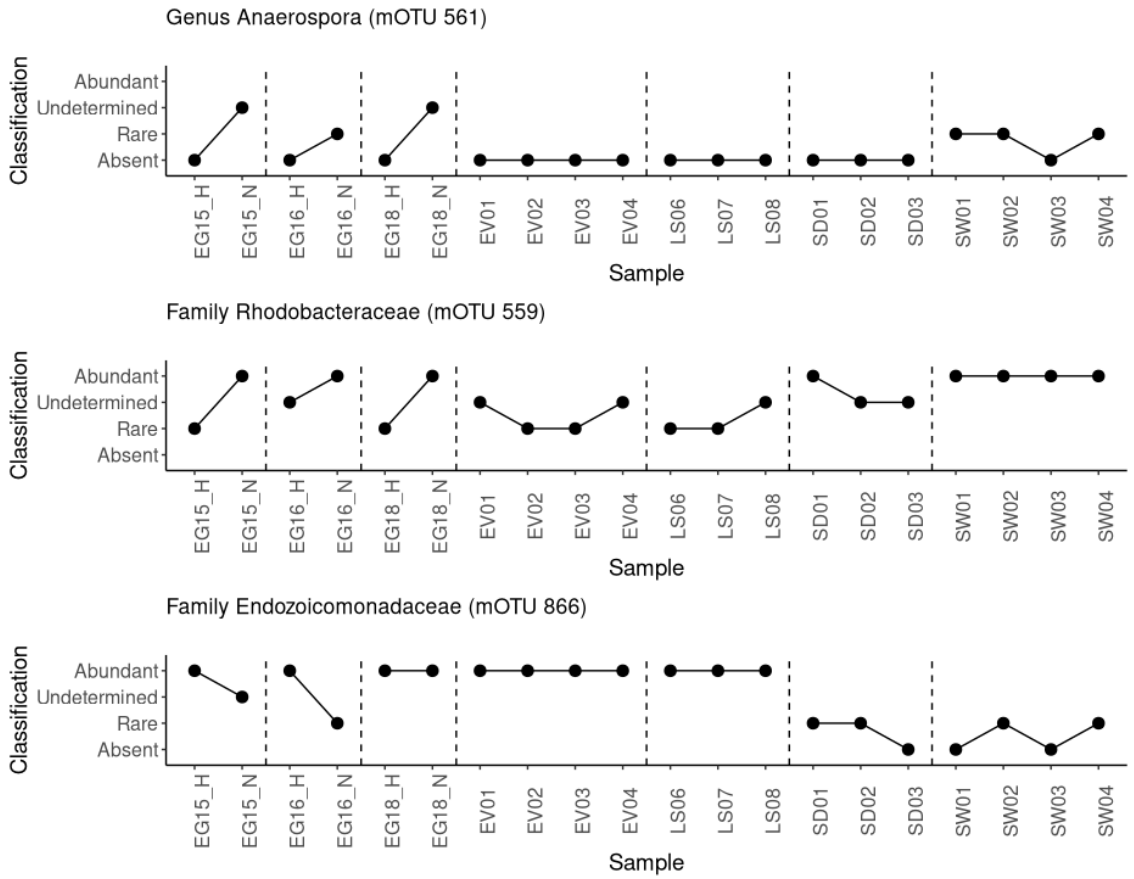


Figure 6.9. Monitoring of the abundance classifications of selected OTUs. Samples are represented on the x axis and the abundance classification on the y axis, with lines connecting samples that are related. For *Eunicella gazella* species, lines connect healthy to necrotic tissue (EG15_H, EG16_H, EG18_H and EG15_N, EG16_N, EG16_N), while *Eunicella verrucosa* (EV01, EV02, EV03, EV04) and *Leptogorgia sarmentosa* (LS06, LS07, LS08) only presented healthy tissue. Additionally, sediment (SD01, SD02, SD03) and seawater (SW01, SW02, SW03, SW04) samples were also represented in distinct groups. The different coral species, sediments and seawater were separated by vertically dashed lines. Relevant taxonomic information of the selected OTUs is indicated at the top of each plot. To indicate that an OTU was absent in a sample, we added the classification “absent”, which is not obtained from *ulrb*.

6.4. Discussion

Microbial ecology studies usually delineate rare from abundant taxa based on relative abundance thresholds (Pedrós-Alió, 2012; Lynch and Neufeld, 2015). Here we propose the *ulrb* method, which automatically clusters taxa based on the relationship between their abundances in a given sample, without the need of a threshold selection. Thus, the common observation that most taxa are “rare” means that these taxa are within a small range of low abundance values, while the few “abundant” have a disproportionately higher abundance. In microbial ecology, the killing-the-winner hypothesis (Thingstad, 2000; Thingstad *et al.*, 2014), in which lower cell abundance decreases the probability of encountering bacteriophages, is often evoked to explain a possible ecological strategy underlying the existence of so many rare taxa within a community. In addition, some microorganisms might be dormant but keep the ability to grow and become abundant under conditions that are more favorable (Shade *et al.*, 2014). Other microorganisms can keep high metabolic activity, even though at low abundance (Pester *et al.*, 2012). Ecological effects, such as dispersion and drift might also contribute for the emergence of some rare taxa (Jia, Dini-Andreote and Falcão Salles, 2018; Pascoal, Costa and Magalhães, 2021). Finally, some rare taxa might be decreasing their abundance towards local extinction (Pedrós-Alió, 2012). Previous reviews have summarized evidence for these mechanisms (Pedrós-Alió, 2012; Lynch and Neufeld, 2015; Jia, Dini-Andreote and Falcão Salles, 2018; Pascoal, Costa and Magalhães, 2021) (Chapter 2). Since *ulrb* is an unsupervised machine learning method, it makes no assumptions about metabolic or ecological mechanisms shaping community composition, *i.e.,* the classification solely depends on the abundance table provided. However, the resulting classifications can be explained by such ecological mechanisms. For example, an abundant taxon that becomes rare could indicate the existence of a top-down factor, while the sudden emergence of a rare taxa previously unreported in a specific environment could indicate dispersion effects.

The most used methods to define the rare biosphere are problematic, because they are based on arbitrary thresholds of relative abundance (Jia, Dini-Andreote and Falcão Salles, 2018; Pascoal *et al.*, 2022a) (Chapter 5). To provide concrete numbers, we summarize the literature on the microbial rare biosphere, from January 2006 to 2024 (Supplementary Table 11.23). Of 181 articles, 37% did not provide a clear definition of rarity and, among those that defined the rare biosphere explicitly, 84% used relative abundance thresholds (Supplementary Table 11.23). Within the studies that relied on relative abundance thresholds, approximately 60% used a single threshold, while the

remaining used two or more thresholds (Supplementary Table 11.23). Approximately half of the studies used 0.1% relative abundance to distinguish rare from abundant phylogenetic units within communities, with approximately 70% applying the threshold per sample, instead of applying it to the whole dataset at once (Supplementary Table 11.23).

We compared two of the most common approaches to define the rare biosphere against *ulrb*, specifically, the utilization of a single threshold of 0.1% relative abundance per sample, and the alternative including an intermediate level of relative abundance ranging from 0.1 to 1%. To do this comparison, we applied the different definitions of rarity to environmental replicates assessed by different methods (V4V5 and full-length 16S rRNA gene sequencing and metagenomics), showing that threshold-based definitions have patterns of diversity that are method-dependent, while *ulrb* provided the same pattern across all methods. More specifically, using threshold-based definitions, the number of rare and abundant taxa was inconsistent across methodologies, but it was very consistent when using *ulrb*. This is because any threshold-based method is inherently dependent on the nature of the dataset, which in turn is dependent on methodology, including the sequencing strategy, marker gene and bioinformatics processing. *ulrb* captures the rarity concept without the need for arbitrary thresholds in a way that is consistent across datasets, because it solely depends on the relative distance between taxa abundance scores. This ability to capture connections between the abundance of taxa, independently of the order of magnitude of the abundance scores, provides classifications that are non-random and both biologically and ecologically informative.

The clustering results from *ulrb* were generally stronger for OTUs and mOTUs than for ASVs (based on Silhouette scores, Supplementary Figure 10.29 and 10.31). ASVs may be harder to cluster into abundance classifications, because they are more prone to extreme values, which will affect the clustering result, for example, by creating a single cluster for outliers. This problem can be solved by removing outliers (Rousseeuw, 1987), but in this context abundant taxa are outliers that must be kept, because they represent real taxa. Therefore, we propose that taxa that are outliers relative to the remaining taxa should be considered “abundant” taxa, even if it means that very few taxa are defined as “abundant”. Another factor that might contribute to the difficulty of clustering ASVs by *ulrb* is that such datasets usually consist of diverse, highly similar phylogenetic units (*e.g.*, sequences diverging in few nucleotides from one another), each of which possessing its own abundance score, but frequently representing one single microbial species (or subpopulations within one species) (Pascoal *et al.*, 2024) (Chapter 4).

The *ulrb* method proved to be statistically robust for any variation of the main variables shaping an abundance table. Collectively, the classification of taxa into “rare” was usually of better statistical quality than the “undetermined” and “abundant” classifications. A reason for this is that the “rare” classification includes more taxa than the “undetermined” and “abundant” classification, which in turn gives the “rare” classification stronger clusters. Additional evidence for this assertion is that if we randomly select a certain number of taxa, the clustering quality becomes equivalent for all classifications. Another reason for the observation of stronger clusters in the rare biosphere is that the variability of abundance among rare taxa is much lower, thus contributing to better defined clusters. The robustness of *ulrb* was not affected in any way by the sequencing depth, which explains why *ulrb* was able to provide consistent results for microbial datasets derived from different sequencing methodologies. In fact, the reason why we cannot use the same relative abundance threshold for 16S rRNA gene metabarcoding (amplicon sequencing) and 16S rRNA gene data derived from shotgun metagenome sequencing is precisely the different order of magnitude of the datasets, which is an issue that is not solved by the compositional nature of the data. Furthermore, *ulrb* was also statistically robust independently of the number of samples, even though the Silhouette scores of the “undetermined” and “abundant” classifications varied more for datasets containing less than 30 samples. This was expected, because a low sample size ($n < 30$) may not be enough to characterize the mean value of a distribution of data (Krzywinski and Altman, 2014). Since *ulrb* is applied to a specific sample and its result is not impacted by the existence of other samples, a perfect result would be a horizontal line, representing no variation in the clustering quality. However, because we are selecting random samples from the dataset, those samples will have random variation between them. As the sample size increases, the random variation decreases and approaches the true average.

ulrb was designed for handling large abundance tables derived from molecular analyses of microbial communities. Yet we showed that it can also be applied to non-microbial data, using the ants and plants datasets (Figure 6.7). This was expected, because *ulrb* relies on the relative distance between the abundance scores of the taxa within a sample. Furthermore, the microbial and non-microbial abundance tables have the same underlying structure, with differences in the number of taxa and the abundance score of those taxa. Thus, since we showed that *ulrb* was consistent for any variation in taxa numbers, sequencing depth and number of samples (Figure 6.6), *ulrb* was expected to work properly for non-microbial data.

Outside the scope of microbial ecology, a previous study has suggested the utilization of unsupervised learning to define rare and common taxa, using *FuzzyQ* (Balbuena *et al.*, 2021). *FuzzyQ* has an analogous framework to *ulrb*, because both can define rare taxa without the introduction of arbitrary thresholds, *i.e.,* they provide automatic classifications. However, there are several differences between both methods, because of the number of features used (Supplementary Figure 10.36), making it unreasonable to directly compare them. However, we applied *FuzzyQ* to similar data in parallel to provide perspective and identify potential advantages and disadvantages. Briefly, both methods can be used to define the rare biosphere for microbial and non-microbial data, but *ulrb* provides information at sample level, while *FuzzyQ* provides information at the whole dataset level. Thus, one disadvantage of *FuzzyQ* relative to *ulrb* is that it is not clear if a taxon is common/rare due to the frequency of occurrence or to its underlying abundance in the study. Consequently, it provides little information on the transition between rare and abundant states for a given taxon across samples. Some advantages of FuzzyQ include the commonality index, which indicates how common or rare a taxon is with a particular dataset. Additionally, the automatic inclusion of frequency of occurrence provides information on commonality, showing how often a taxon appears across samples. This information about the “rare” and “abundant” classification can be useful in certain experimental settings.

We showed that *ulrb* can be adapted to manually inspect the types of rarity within a given dataset, using data derived from a coral microbiome study (Keller-Costa *et al.*, 2021). Such an analysis supported the identification of likely mutualistic octocoral symbionts, such as members of the family *Endozoicomonadaceae*, which were abundant in healthy coral tissues, but rare or absent in most necrotic tissues, sediment and seawater. The same approach also allowed the identification of the genus *Anaerospora* as a seawater rare biosphere member with the ability to colonize necrotic corals, but not healthy ones. Such examples demonstrate that *ulrb* can be easily adapted to ascertain different types of rarity by monitoring selected taxa across relevant variables. However, *ulrb* is currently unable to automatically calculate types of rarity for any dataset, which means that the user must manually do such monitoring. We foresee the implementation of such capabilities in future versions of *ulrb*.

On the microbial side, this study focused specifically on prokaryotes, but *ulrb* should work equally well for other microbial groups (*e.g.*, fungi and protists) obtained with high-throughput sequencing methods, because the data will have similar characteristics. Furthermore, any variation in such datasets will necessarily be within differences in

number of taxa, samples and sequencing depth, which we showed did not have any impact on *ulrb* robustness and applicability.

The identification of types of rarity across a set of samples, the optimal number of abundance clusters to be used, and the eventual occurrence of clusters represented by outliers are all challenges that need to be met in current rare biosphere research. For each case, the present version of *ulrb* offers possible solutions. We show that types of rarity can be defined by manually inspecting target taxa; we suggest a standard number of clusters (k = 3), and the existence of clusters composed of a single outlier taxon can be explained by the extremely high abundance of such taxon relative to the remaining, but this problem can be solved by testing a different number of clusters. In terms of computational power, since *ulrb* applies its calculations on a single dimension, it is quite fast. It is worth noting that if different studies select different numbers of clusters, then inter-comparability across studies might be compromised.

6.5. Conclusion

This study presents the *ulrb* R package, with a new methodology to define the rare biosphere across microbial communities. This R package is open-source and includes a dedicated website (https://pascoalf.github.io/*ulrb*/), with tutorials explaining how to use *ulrb* functions and extensive documentation.

We show that *ulrb* provides a more consistent interpretation of the microbial rare biosphere across different sequencing strategies and bioinformatic protocols than threshold-based definitions, because it is statistically robust against variations in taxa counts, sequencing depth and number of samples. We demonstrate that *ulrb* can also be used for non-microbial data, because it depends only on the relative differences between the abundance scores of taxa within a community. Thus, *ulrb* is effectively independent of the methodology used to produce the abundance table.

Finally, we show that *ulrb* results can be used to manually monitor specific taxa and ascertain types of rarity (Shade *et al.*, 2014; Lynch and Neufeld, 2015). However, future work is necessary to implement an automatic classification of types of rarity in the *ulrb* R package.

Owing to the features mentioned above, *ulrb* is readily applicable to discern rare from abundant organisms across various scenarios, showing great potential to standardize microbial rare biosphere analysis. *ulrb* can be used, but is not limited, to studying transitions from eubiosis to dysbiosis states in host-associated microbiomes, emerging microbial diseases because of climate change, biological invasions, community gradient analyses and landscape ecology data, to name a few possible applications.

6.6. Declarations

Author contributions

Conceptualization (all authors); Data curation (FP and CM); Formal Analysis (FP); Funding acquisition (all authors); Methodology (FP); Resources (CM); Software (FP, PB and LT); Supervision (LT, PB, RC and CM); Writing - original draft (FP); Writing - review and editing (all authors).

Funding

FCT funded this study through one grants to CM (2022.02983.PTDC) and through a PhD grant to FP (2020.04453). This research has also been supported by FCT through the projects UIDB/04565/2020 and UIDP/04565/2020 of iBB, UIDB/04423/2020 and UIDB/04565/2020 of CIIMAR, and the project LA/P/0140/2020 of i4HB. The work of Paula Branco was undertaken, in part, thanks to funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

The FASTQ files from the N-ICE dataset are available at ENA, with project ID PRJEB21950 (V4V5 16S rRNA gene amplicon sequencing) and PRJEB15043 (shotgun sequencing of metagenomes). The FASTQ files from the MOSJ dataset are all available in ENA under the projects PRJEB24517 (2016), PRJEB72025 (2017), PRJEB72030 (2018), PRJEB60815 (2019) and PRJEB72034 (2020). The Ants dataset used is available in the R package *FuzzyQ*. The BCI dataset used was derived from original data made publicly available in DRYAD (10.15146/5xcp-0d46). The octocoral microbiome dataset shotgun metagenomic sequencing data is available in ENA, under project PRJEB13222.

Source code and data

The source code for *ulrb* R package is available at GitHub (https://github.com/pascoalf/*ulrb*) and CRAN (https://cran.r-project.org/web/packages/*ulrb*/index.html). All the code to process raw reads and reproduce the figures and tables in this paper are available in a GitHub repository

(https://github.com/pascoalf/Unsupervised-machine-learning-definition-of-the-microbial-rare-biosphere).

Acknowledgments

We acknowledge the contributions of Pedro Duarte and Philipp Assmy to ongoing support on the participation and organization of the expeditions for the MOSJ and N-ICE dataset. We also acknowledge logistic support from the NPI, regarding the MOSJ and N-ICE dataset. We also acknowledge the original authors of the publicly available datasets used: Richard Condit, Stephen Hubbell, *et al* (BCI); Xavier Arnan, Alan Anderson, *et al*. (Ants dataset); Tina Keller-Costa *et al*. (coral microbiome); and the Microbiome Ecology and Biogeochemistry team, from CIIMAR (MOSJ and N-ICE dataset).

6.7. Supplementary Methods

6.7.1. Unsupervised Learning Statistics

6.7.1.1. Calinski-Harabasz index

The Calinski-Harabasz index (CH) measures the separation and density of clusters by calculating the ratio between inter-cluster dispersion and intra-cluster dispersion and can be calculated by Equation 6.2. (Calinski and Harabasz, 1974),

$$CH = \frac{\sum_{k=1}^{k} n_k |C_k - C|^2}{\sum_{k=1}^{k} \sum_{i=1}^{n_k} |X_{ik} - C_k|} \frac{(N-k)}{(k-1)}$$

Equation 6.2. Calinski-Harabasz index formula.

where k is the cluster number, n_k is the number of taxa in cluster k, C is the centroid of the dataset, C_k is the centroid of cluster k, N is the number of taxa in the dataset and X_{ik} is the ith taxon of cluster k.

The higher the CH value, the higher the separation and density of the clusters. Thus, the best k maximizes CH. To calculate CH, *ulrb* used the function index.G1() from the *clusterSim* R package (Walesiak and Dudek, 2020).

6.7.1.2. Davies-Bouldin index

The Davies-Bouldin index (DB) measures cluster separation and can be expressed as Equation 6.3. (Davies and Bouldin, 1979),

$$DB = \frac{1}{N} \sum_{i=1}^{N} R_{i,j}$$

Equation 6.3. Davies and Bouldin index formula.

where N is the number of taxa in the dataset and $R_{i,j}$ is the maximum similarity between possible combinations of cluster i and j, with ij. Thus, the best k minimizes DB. To calculate DB, *ulrb* uses the function index.DB() from the *clusterSim* R package (Walesiak and Dudek, 2020).

6.7.2 Datasets

6.7.2.1. N-ICE

The N-ICE (2015) sampling campaign was previously described by Granskog *et al*. (2016, 2018) and the seawater filtering, DNA extraction, preparation and sequencing was previously described by de Sousa *et al*. (2019). Briefly, a vessel was fixed on ice and drifted along from March to July. Samples were taken at 5 m, 50 m and 250 m in the region of the Nansen Basin (83°10.002′ N, 22°01.998′ E) on the 9th of March 2015 (n = 6 samples); then at the same depths in a transition region between the Eurasian Basin and Yermak Plateau (80° 30.775´N; 07° 52.428´E) on the 24th of April 2015 (n = 6 samples); and at 5 m, 20 m and 250 m in the Yermak Plateau (80° 30.775´N; 07° 52.428´E) on the 16th of June 2015 (n = 6 samples). For each site and depth, one replicate was taken for V4V5 16S rRNA amplicon sequencing (n = 9 samples) and the other for shotgun metagenome sequencing (n = 9 samples).

Between 3 L and 11 L of seawater were filtered through cartridge filters (0.22 µm pore size; Sterivex units) and DNA was extracted using the PowerWater DNA Isolation Kit protocol (MO BIO Laboratories, Inc.), following standardized guidelines from OSD (Kopf *et al.*, 2015). For amplicon sequencing, the V4V5 region of the 16S rRNA gene was selected based on primers that were improved to ensure wide coverage of marine prokaryotes (Apprill *et al.*, 2015; Parada, Needham and Fuhrman, 2016). More specifically, we used the primers 515YF (5′-GTGYCAGCMGCCGCGGTAA - 3′) and 926R (5′ - CCGYCAATTYMTTTRAGTTT- 3′) (Caporaso et al., 2011; 2012; Apprill *et al*., 2015; Parada, Needham and Fuhrman, 2016). Both the V4V5 16S rRNA amplicons and the shotgun metagenomes were sequenced by Illumina technology on a MiSeq (2 x 300bp) platform (de Sousa *et al.*, 2019).

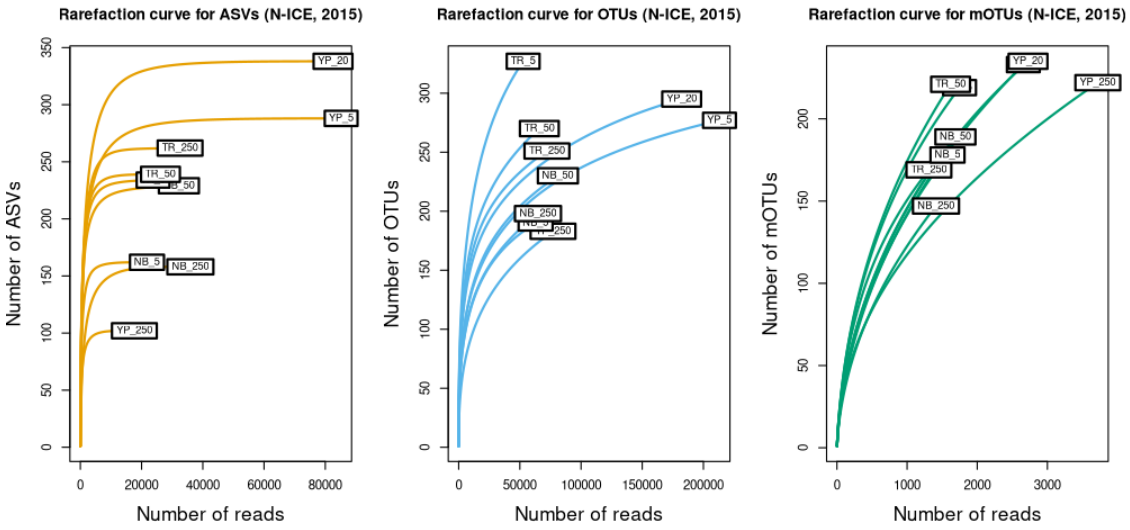
Raw reads from V4V5 16S rRNA gene amplicon sequencing were processed by two different approaches (Pascoal *et al.*, 2022a) (Chapter 5): into OTUs, using the MGnify platform v5 (Mitchell *et al.*, 2019); and into ASVs, using *DADA2* R package (Callahan *et al.*, 2016).

Briefly, to produce OTUs, the MGnify platform merged forward and reverse reads (SeqPrep v1.2), filtered high quality reads (Trimmomatc v0.36) and attributed reads to hierarchical OTUs (MAPSeq v1.2.3 (Matias Rodrigues *et al.*, 2017), using Silva v132 for taxonomy (Quast *et al.*, 2012; Yilmaz *et al.*, 2014)). The OTU picking step was not restricted to taxonomically classified sequences from the reference database but used

reference-mining with an approach that included unclassified sequences, using several similarity thresholds (with MAPSeq). This OTU picking approach provides OTUs hierarchically related at several similarity thresholds, making them fast to compute (Matias Rodrigues *et al.*, 2017). For simplicity, we refer to these “hierarchical OTUs” as “OTUs”.

Raw reads from shotgun metagenomic sequencing were processed into mOTUs by the MGnify platform v5 (Mitchell *et al.*, 2019). The bioinformatics steps were similar to those described previously for amplicon sequencing, but additional steps were necessary to identify SSU rRNA genes (Infernal v1.1.2 (Nawrocki and Eddy, 2013) and Rfam 13.0 (Kalvari *et al.*, 2018)). Reads were attributed to mOTUs at a 100% similarity cut-off (mOTU2 (Milanese *et al.*, 2019)) and taxonomically classified using the Silva database v132 (Quast *et al.*, 2012; Yilmaz *et al.*, 2014). mOTU tables were filtered to remove taxa attributed to unknown domain-level classifications, eukaryotic and organelles, if any; and singletons were removed, if any.

The ASV, OTU, and mOTU tables were filtered to remove taxa attributed to unknown domain-level classifications, eukaryotes and organelles, if any; and singletons were removed, if any. Based on rarefaction curves (Supplementary Methods Figure 6.1), the ASV and OTU abundance tables were rarefied to include 10 000 reads per sample, excluding any samples below that number of reads (n = 9 samples after this step). And the mOTU abundance tables were rarefied to include 1 000 reads per sample, excluding any samples below that number of reads (n = 9 samples after this step).



Supplementary Methods Figure 6.1. Rarefaction curves for N-ICE dataset samples (n = 9), grouped by phylogenetic units (ASVs, OTUs and mOTUs).

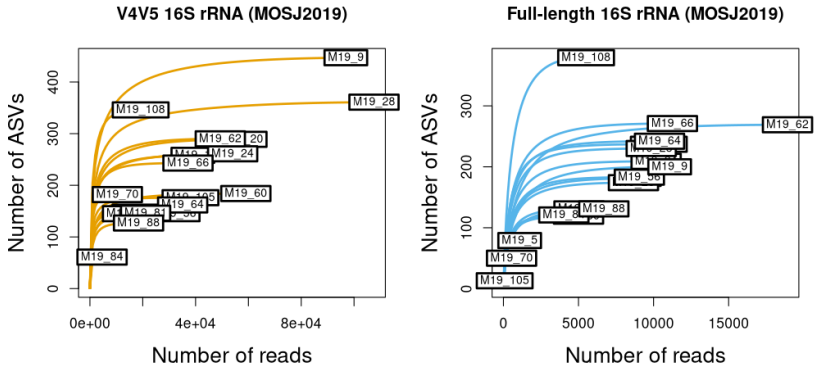
6.7.2.2. MOSJ 2019 dataset

Both V4V5 and full-length 16S rRNA amplicon sequencing samples were processed previously by Pascoal *et al.* (2024) (Chapter 4).

For the full-length 16S rRNA gene, the primers 27F (5'- AGRGTTYGATYMTGGCTCAG-3') and 1492R (5'-RGYTACCTTGTTACGACTT-3') (Paliy *et al.*, 2009) were used and amplicons were sequenced by PacBio CCS technology. Raw reads in bam files were cleaned and provided in FASTQ format by the sequencing provider. Read length was trimmed between 1000 bp and 1600 bp. Trimmed reads were then processed into ASVs following the *DADA2* protocol with adaptations for long-read sequencing by CCS PacBio (Callahan *et al.*, 2019; Kumar *et al.*, 2019).

For the V4V5 16S rRNA gene sequencing data, the *DADA2* (Callahan *et al.*, 2016) protocol was used, following the same steps described previously (Pascoal *et al.*, 2024) (Chapter 4).

For both V4V5 and full-length 16S rRNA gene sequencing, abundance tables were filtered to remove taxa attributed to unknown domain-level classifications, eukaryotes and organelles, if any; and singletons were removed, if any. For the full-length 16S rRNA gene sequencing, abundance tables were rarefied to 5 000 reads, based on rarefaction curve profiles (Supplementary Methods Figure 6.2), and samples below 5 000 reads were removed (n = 6 samples after this step). For the V4V5 16S rRNA gene sequencing data, the sequencing depth was rarefied to 10 000 reads per sample, based on the analysis of rarefaction curves (Supplementary Methods Figure 6.2), and any sample below that number of reads was discarded (n = 6 after this step).



Supplementary Methods Figure 6.2. Rarefaction curves for all MOSJ 2019 dataset samples (n = 18), grouped by V4V5 and full-length 16S rRNA gene amplicon sequencing.

6.7.2.3. Coral microbiome dataset

Sampling and sequencing details of the coral microbiome dataset used in this study were described previously (Keller-Costa *et al.*, 2021). Briefly, samples were taken at the Atlantic coast of Faro, Algarve, Portugal (36º 58' 47.2 N; 7º 59' 20.8W), on the 17th of July 2014, at 17 m depth. Branches of 10 to 20 cm were collected from replicate specimens of the octocoral species *Eunicella gazella* (EG, n = 6 samples), *Eunicella verrucosa* (EV, n = 4 samples) and *Leptogorgia sarmentosa* (LS, n = 3 samples). Additionally, 2 L of seawater surrounding the octocorals (n = 4 samples) and 100 g of surface sediment near the octocorals (n = 3 samples) were collected. To extract total community DNA from all samples, the UltraClean® Soil DNA isolation kit (MO BIO, Carlsbad, CA, USA) was used. Sequencing for shotgun metagenomic sequencing was performed using the Illumina HiSeq 2500 device at MR DNA (Shallowater, Texas, USA). For more details, see Keller-Costa *et al*. (2021).

The processing of shotgun metagenomic sequencing reads from the coral microbiome dataset was described previously (Keller-Costa *et al.*, 2021), and used the MGnify platform v2 (Mitchell *et al.*, 2019). Reads were processed following similar steps to the more recent v5 version, but resulting in the retrieval of mOTUs defined at a 97% similarity cut-off instead. The fact that the N-ICE and coral microbiome shotgun sequencing datasets were processed with different versions of the MGNify pipeline does not interfere with the analyses performed in this study, since we do not compare these datasets directly. Instead, it further adds to testing the robustness of *ulrb* at handling datasets representing different contexts and data processing scenarios. The 16S rRNA gene reads from the shotgun metagenomic dataset included 93 589 high-quality reads and 1 041 mOTUs defined at a 97% similarity cut-off (Keller-Costa *et al.*, 2021).

Chapter 7. *ulrb*: an R Package to Automatically Define Rare Taxa.

This chapter used materials from an article accepted for publication for Ecological Informatics journal at time of thesis submission:

Pascoal, F. et al. (2025b) “Architecture and implementation of ulrb algorithm in R,” Ecological Informatics. Article in press. Available at: https://doi.org/10.1016/j.ecoinf.2025.103229.

See author contributions in section **7.5 Declarations**.

7. *ulrb*: an R Package to Automatically Define Rare Taxa

Francisco Pascoal[1,2], Rodrigo Costa[3,4], Luís Torgo[5], Catarina Magalhães[1,2] and Paula Branco[6]

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, rua do Campo Alegre s/n, 4169– 007 Porto, Portugal; [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal; [3]Department of Bioengineering, Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico, University of Lisbon, Portugal; [4]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, University of Lisbon, Portugal; [5]Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada; [6]School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, Ottawa, Ontário, Canada.

Abstract

The microbial rare biosphere is important for ecosystems resistance and resilience but challenging to study. One of the main difficulties lies in the lack of an appropriate definition of rare taxa. Most studies use relative abundance thresholds (*e.g.*, 0.1% relative abundance, per sample) to discern rare from abundant taxa within a microbial community. This is inappropriate because such thresholds are arbitrary and lack biological meaning. To solve this problem, we have proposed the utilization of unsupervised machine learning, through the *ulrb* ("Unsupervised Learning based Definition of the Microbial Rare Biosphere") R package (v0.1.6). This R package uses the partition around medoids (pam) algorithm to cluster taxa based on their abundance, in a given community. Based on the clusters, *ulrb* automatically classifies taxa into "rare", "undetermined" or "abundant", by default. *ulrb* includes functions for all analytical steps necessary to define the rare biosphere. Specifically, we include four groups of functions: 1) transform common types of wide abundance tables into long format data; 2) classify taxa with the pam algorithm; 3) helper functions to evaluate detailed statistics of the clustering steps; and 4) visualization functions, focused on rank abundance curves and Silhouette scores, for assessment of clustering quality. In this article, we describe the *ulrb* R package architecture, coding organization, and strategy. Furthermore, we use a 16S rRNA gene amplicon sequencing dataset from the Arctic Ocean to provide illustrative examples, with code, on how to use and explore *ulrb* capabilities.

Keywords: unsupervised machine learning, rare biosphere, microbial ecology, rarity, k-medoids, partition around medoids.

7.1. Introduction

Most taxa in natural ecosystems are rare, a recognizable pattern in both macro- (McGill *et al.*, 2007) and microbial ecology studies (Pedrós-Alió, 2012). However, it is not clear how to delineate rare from abundant taxa. One common approach is to manually select a relative abundance threshold to discriminate between rare and abundant taxa in a community. However, such an approach lacks inherent biological meaning and is not comparable across studies. Even if two studies use a threshold of 0.1% relative abundance per sample, if the abundance tables are derived from different methods, with abundance scores differing by orders of magnitude, then the meaning of the 0.1% threshold is inherently different, making it very difficult to compare studies (Pascoal, Costa and Magalhães, 2021; Pascoal *et al.*, 2022a; Pascoal *et al.*, 2025a) (Chapters 2, 5 and 6). Furthermore, if two studies use different abundance thresholds, then it becomes even harder to establish comparisons.

To solve this problem, we proposed a new methodology based on unsupervised machine learning to cluster taxa within microbial communities, using their abundance score (Pascoal *et al.*, 2025a) (Chapter 6). For that purpose, we built the *ulrb* ("Unsupervised Learning based Definition of the Microbial Rare Biosphere", v0.1.6) package in R (R Core Team, 2024). The *ulrb* approach is based on k-medoids and can use several clusters, which can then be assigned to meaningful classifications. By default, three clusters are used: "rare", "undetermined" and "abundant". The inclusion of one intermediate group ("undetermined") avoids the existence of rare and abundant taxa with very similar abundance scores, *i.e.* near the limits of each cluster; and accounts for the possibility that some taxa are neither rare nor abundant. Moreover, in *ulrb* the names and number of classifications can be changed by the user, according to the needs of the experimental design, enabling researchers to approach the microbial rare biosphere across multiple biological contexts.

The *ulrb* R package is useful for ecologists, because it automatically classifies taxa based on their abundance scores, without the need for any arbitrary threshold. To the best of our knowledge, the most similar R package to *ulrb* is *FuzzyQ* (Balbuena *et al.*, 2021), which also uses unsupervised machine learning to define rare taxa. However, *FuzzyQ* applies a different clustering approach, using both frequency of occurrence and relative abundance to infer whether taxa are rare or abundant across a set of biological communities (samples), and it was not envisioned for microbial data. The *ulrb* package was developed with microbial communities in mind and delivers abundance

classifications for all taxa in each biological community (sample) within a set of samples. *ulrb* also works for macroorganismal data, if there is an abundance table (Pascoal ***et al.***, 2025a)(Chapter 6).

In here, we describe the *ulrb* R package (v0.1.6), explaining its architecture, functions, and strategy for algorithm implementation. The *ulrb* R package is freely available in CRAN (https://cran.r-project.org/web/packages/*ulrb*/index.html) and GitHub (https://github.com/pascoalf/*ulrb*). Additionally, *ulrb* includes a dedicated website with extensive documentation and learning materials (https://pascoalf.github.io/*ulrb*/index.html).

7.2. The *ulrb* R package

7.2.1. Description of the rare biosphere definition problem in machine learning terms

Consider an abundance table, with any number of samples and an abundance score for each taxon. In this context, each sample represents a community of taxa, and the observations are the abundance scores of each taxon. It is possible to have additional features attributed to each taxon and sample metadata. However, for the problem of the definition of rarity, we only consider the abundance scores within each sample.

Assume that samples are independent of each other and that the target classification is unknown, *i.e.,* we do not know which taxa are truly rare or otherwise. Thus, if we need to classify an unknown target, based on known features of the target, it immediately follows that we need to apply an unsupervised machine learning strategy.

7.2.2. Unsupervised machine learning algorithm and features selection

A previous study has proposed unsupervised machine learning with two features to solve the definition of rare and common taxa, the abundance score and the frequency of occurrence (Balbuena *et al.*, 2021). However, since *ulrb* assumes that samples are independent from each other, we cannot account for frequency in the algorithm for *ulrb*. Thus, we decided to use a single feature in *ulrb*, the abundance score.

Among the unsupervised learning algorithms, we focused on clustering algorithms with explicit classifications, thus excluding fuzzy classifications. A previous study has used a fuzzy approach (Balbuena *et al.*, 2021) so that taxa are given a probability of being rare. However, the *ulrb* approach obtains a specific classification, thus replacing the utilization of thresholds. Some of the models considered were the k-means (MacQueen, 1967) and k-medoids (Kaufman and Rousseeuw, 1987) since they are easy to interpret and very fast for a single feature. Even though both can work for this purpose, we selected k-medoids, implemented by pam algorithm (Kaufman and Rousseeuw, 1987), because it centers clusters around real observations, while k-means can place the center in intermediate points between observations; and because it is less sensitive to outliers (Kaufman and Rousseeuw, 1990). This latter point is relevant, because most taxa have very low abundance scores, making the few abundant taxa like outliers.

7.2.3. Number of clusters (k) used

The number of clusters is the number of classifications obtained by *ulrb*. The name of each classification can be dependent on the study design (default is “rare”, “undetermined” and “abundant”). For example, for two clusters we would have the traditional distinction between rare and abundant taxa.

Overall, we can look at the problem of the number of clusters, or k, in three ways: 1) we may standardize the number of clusters; 2) we can decide k automatically, based on some quality metric, like the average Silhouette score; and 3) the user can specify the number of clusters.

Considering that using the same abundance classifications across studies allows them to be comparable, independently of sample size, the default approach used by *ulrb* is the (1), *i.e.*, a predefined number of clusters. However, we allow the user to automatically decide the value of k and/or to use its own k value.

The predefined classifications are “rare”, “undetermined” and “abundant”, which correspond to k = 3. We consider it necessary to include an intermediate classification between rare and abundant taxa so that the output does not include rare and abundant taxa with very similar abundance scores, *i.e.*, those that would generally lay around the rarity threshold limits. If the user prefers the traditional dichotomy between rare and abundant, it is possible to do so, corresponding to k = 2. Note that k = 1 is mathematically possible, but self-contradictory, because it clusters all observations in the same cluster. Finally, the maximum value of k for each sample is the number of distinct observations. Using such values of k, however, is self-contradictory, because each possible abundance score becomes its own cluster.

For the automatic decision of k, we bound the possible range of values from 3 to 10, because we consider that above 10 clusters the clustering result is no longer useful for delineating rare taxa. We made a tutorial on the *ulrb* website explaining some of the extreme examples regarding k selection, including k = 1 and maximum k (https://pascoalf.github.io/*ulrb*/articles/explore-classifications.html).

7.2.4. *ulrb* architecture

ulrb (v0.1.6) was implemented as an R package, built using the *devtools* R package (Wickham *et al.*, 2022) and it includes functions, data, examples, and vignettes with tutorials. In general, the *ulrb* functions are divided into four groups:

1. Data preparation.
    a. prepare_tidy_data().
2. Classification task;
    a. define_rb().
3. Unsupervised clustering statistics.
    a. check_CH();
    b. check_DB();
    c. check_avgSil();
    d. evaluate_sample_k();
    e. evaluate_k();
    f. suggest_k().
4. Visualization.
    a. plot_*ulrb*();
    b. plot_*ulrb*_clustering();
    c. plot_*ulrb*_silhouette().

The flow of data input and output is illustrated in Figure 7.1. We describe each group of functions in *ulrb* next.

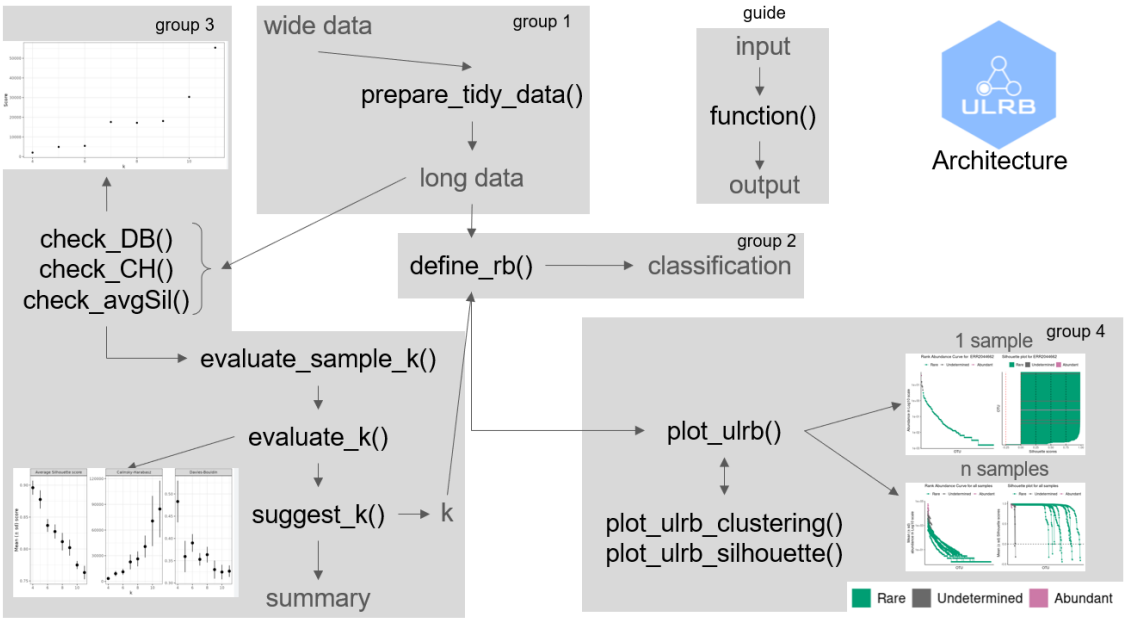


Figure 7.1. Schematic representation of *ulrb* architecture. The *ulrb* R package is divided into four groups of functions, highlighted in the figure by the gray shade. Each function is represented in black text and data input/output are in gray text. The arrows indicate the flow of data.

7.2.4.1. Dependencies

The *ulrb* R package depends on *rlang* (Henry and Wickham, 2024), *dplyr* (Wickham *et al.*, 2023) and *tidyr* (Wickham, Vaughan and Girlich, 2023) for data manipulation; *cluster* (Maechler *et al.*, 2023) to perform the pam algorithm, *purrr* (Wickham, Henry and RStudio, 2023) to iterate calculations over any number of samples, without the need of for loops; *stats* (R Core Team, 2023) and *clusterSim* (Walesiak and Dudek, 2020) for statistical calculations; *ggplot2* (Wickham, 2016) and *gridExtra* (Auguie, 2017) for visualization.

7.2.4.2. Functions' syntax and parameters

The function prepare_tidy_data() has the following syntax:

prepare_tidy_data(

data,

sample_names,

samples_in = "cols",

...)

The parameter *data* is a data frame in wide format, containing the abundance table. The parameter *sample_names* is a vector indicating the name of the samples as they are in

the abundance table, and the parameter *samples_in* indicates if the samples are in columns (set to "cols") or in rows (set to "rows"). By default, prepare_tidy_data() assumes samples are in columns and taxa in rows. The output of prepare_tidy_data() is a data frame in long format with, at least, one column for abundance and another column for samples. The … parameter represents extra arguments from other functions used within the first (the same goes for all other functions of *ulrb*).

The function define_rb() has the following syntax:

define_rb(
data,
classification_vector = c("Rare", "Undetermined", "Abundant"),
samples_col = "Sample",
abundance_col = "Abundance",
simplified = FALSE,
automatic = FALSE,
index = "Average Silhouette Score",
...)

Where the parameter *data* is a data frame in long format, including at least one column for the abundance scores and another for the sample names. Any number of additional columns are allowed, and they will not be altered. The parameter *classification_vector* is a vector with the names for the classification of the clusters obtained from the pam algorithm. By default, the classification names are: "Rare", "Undetermined" and "Abundant". Note that the size of the *classification_vector* parameter corresponds to the number of clusters and the order of the elements of the vector matters, because the classifications are attributed from the cluster with lowest median abundance to highest. The parameters *samples_col* and *abundance_col* explicitly indicate the name of the columns with sample and abundance information, respectively. By default, *ulrb* assumes those column names to be "Sample" and "Abundance". The parameter *simplified* can be TRUE/FALSE and it indicates if the define_rb() function output is the classification result (*simplified = TRUE*) or if it includes additional statistics on the clustering algorithm (*simplified = FALSE*). The parameter *automatic* can be TRUE/FALSE and it determines if the define_rb() function will apply the default value of k (*automatic = FALSE*) or if it will decide the number of k based on the optimization of some clustering quality metric (*automatic = TRUE*). If *automatic = TRUE*, the parameter *index* specifies which quality metric to use to automatically decide k, using the average Silhouette score by default.

The output of define_rb() is the same long format table used in the *data* parameter but with additional columns (which are controlled by the parameter *simplified*).

The functions check_CH(), check_DB and check_avgSil() share the same syntax:

check_CH/check_DB/check_avgSil(
data,
sample_id,
samples_col = "Sample",
abundance_col = "Abundance",
range = 3:10,
with_plot = FALSE,
...)

Where the parameter *data* is a data frame in long format, containing, at least, a column for the classification, abundance score, and sample information of each taxon. The parameter *sample_id* indicates to which sample the function will be applied. Thus, the *data* parameter can include any number of samples, but it will calculate the index on a selected sample. The parameters *samples_col* and *abundance_col* indicate the name of the columns with sample and abundance information, respectively. The parameter *range* indicates the values of k to be compared, by default, the index will be calculated for three to ten clusters. The *range* parameter can extend from two clusters up to the maximum number of different observations. The *with_plot* argument specifies the output. If *with_plot = TRUE*, the default, then the output is a plot with the index value for each value of k, for the selected sample. Otherwise, a vector is returned, indicating the index score for each number of clusters.

The function evaluate_sample_k() has the following syntax:

evaluate_sample_k(
data,
sample_id,
samples_col = "Sample",
abundance_col = "Abundance",
range = 3:10,
with_plot = FALSE,
...)

The parameters are similar to those used by the functions *check_CH(), check_DB(), and check_avgSil().* However, the output is different. The *with_plot = FALSE* setting (default)

returns a data frame, indicating the score for all quality metrics (Davies-Bouldin, Calinski-Harabasz, and average Silhouette score) for each number of clusters in the *range* selected, for the specified sample. If *with_plot* = *TRUE*, then it returns a plot with the scores for all the metrics, divided by grids.

The function evaluate_k() has the following syntax:

```
evaluate_k(
data,
range = 3:10,
samples_col = “Sample”,
abundance_col = “Abundance”,
with_plot = FALSE,
...)
```

The arguments of evaluate_k() are the same as those of evaluate_sample_k(), except that there is no parameter for *sample_id* in evaluate_k(), because evaluate_k() applies the function evaluate_sample_k() to all samples present in the data input. The output is controlled, in the same way, by *with_plot*. However, the data frame output (*with_plot* = *FALSE*) is nested and includes all quality metrics for all samples; and the plot output (*with_plot* = *TRUE*) provides the mean and standard deviation of each quality metric across all samples, for each k value.

The function suggest_k() has the following syntax:

```
suggest_k(
data,
range = 3:10,
samples_col = “Sample”,
abundance_col = “Abundance”,
index = “Average Silhouette Score”,
detailed = FALSE,
… )
```

The parameters *data, range, samples_col,* and *abundance_col* are the same as the ones used for the evaluate_k() function. The parameter *index* indicates which quality metric to be optimized to select the number of clusters. By default, *index = “Average Silhouette Score”*. The other alternatives for the parameter *index* are “Davies-Bouldin” and “Calinski-Harabasz”. The parameter *detailed* indicates the output options. If *detailed* = *TRUE*, then a list is returned, including extensive information on all available quality

metrics. If *detailed* = *FALSE*, then suggest_k() returns an integer with the number of clusters, based on the quality metric selected by the *index* parameter.

The function plot_*ulrb*() has the following syntax:

```
plot_ulrb(
data,
sample_id = NULL,
taxa_col,
plot_all = TRUE,
silhouette_score = "Silhouette_scores",
classification_col = "Classification",
abundance_col = "Abundance",
log_scaled = FALSE,
colors = c("#009E73", "#F0E442", "#CC79A7"),
...)
```

The parameter *data* is a data frame with, at least, the classification, abundance, sample information, and silhouette score for each taxon. The *sample_id* refers to the sample name of the sample to plot. If *sample_id* is not specified (default), then all samples are plotted, using a centrality metric (mean and standard deviation). The parameter *taxa_col* indicates the name of the column with taxa information and the *plot_all* parameter indicates if a single sample is to be plotted or all at once (note that this parameter must be consistent with *sample_id*). The parameters *silhouette_scores*, *classification_col* and *abundance_col* indicate the columns with Silhouette scores, classification, and abundance score, respectively. The *log_scaled* parameter indicates if the plot is on a log10 scale, and the colors parameter indicates the colors for each classification. Note that the vector in the *colors* parameter must be the same size as the vector in the *classification* parameter.

The functions plot_*ulrb*_clustering() and plot_*ulrb*_silhouette() share similar syntax. While the first plots the clustering result (as a RAC), the second plots the Silhouette scores. They have the following syntax:

```
plot_ulrb_clustering/plot_ulrb_silhouette(
data,
sample_id = NULL,
taxa_col,
plot_all = FALSE,
```

samples_col = "Sample",

classification_col = "Classification",

abundance_col = "Abundance",

log_scaled = FALSE,

colors = c("#009E73", "#F0E442", "#CC79A7"),

...)

The parameters for plot_*ulrb*_clustering() are the same as the ones for plot_*ulrb()*.

7.2.4.2. Data preparation

If the user has an abundance table in wide format, which is quite common in ecology, this table is transformed into an abundance table in long format, or tidy, by the function prepare_tidy_data(). The function will only change the organization of abundance information, keeping any additional features of the dataset intact. Any number of features can be maintained from beginning to end of the pipeline, without affecting the classification results. The user might decide to remove such features, because they are unnecessary, but might want to keep them to allow immediate subsequent domain-specific analysis, without the need of re-joining those features to the data. Alternatively, the user might directly use a table in the correct long format, without needing the prepare_tidy_data() function. In Figure 7.2 we illustrate the two accepted wide formats and the long format resulting from the prepare_tidy_data() function.

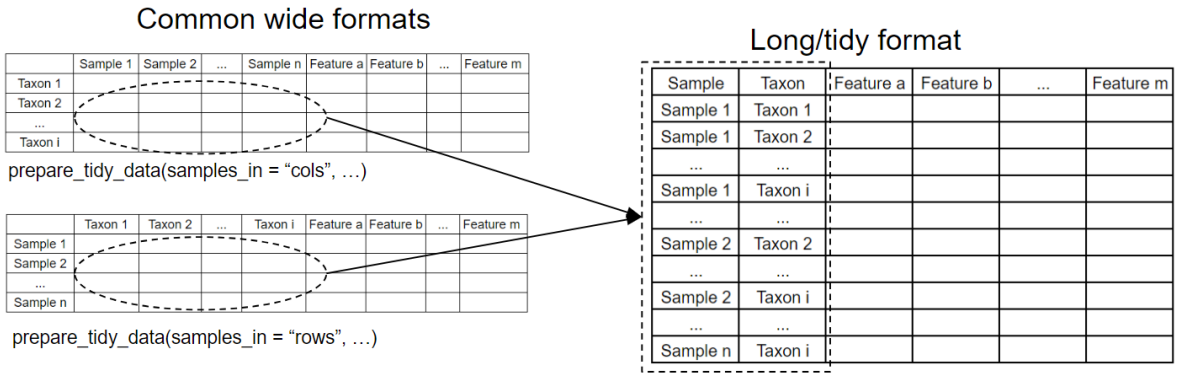


Figure 7.2. Schematic representation of wide and long table format in the context of abundance tables. On the left, the most common wide formats used in ecology are represented, whereby the samples and taxa are separated into distinct columns/rows. On the right is the long format, in which a single column exists for the sample and taxa variable. Additional variables are not modified by the prepare_tidy_data() function.

7.2.4.3. Classification task

The classification task is done by the define_rb() function, standing for “define rare biosphere”. This is the central function of the *ulrb* R package and requires data in long format. By default, the define_rb() function classifies taxa into rare, undetermined, and abundant (k = 3), but this can be changed in two ways: 1) by setting a different “classification_vector” argument; or 2) by “setting automatic = TRUE”, which will automatically calculate the best k. Note, however, that the automatic option will give numbers to the classification results, instead of meaningful names. Thus, it is up to the user to decide the meaning of each cluster. For example, if the automatic classification sets k = 5, we might interpret the classification result as “very rare”, “rare”, “undetermined”, “abundant” and “very abundant”. Such an example could be very informative and useful in several ecological settings. The algorithm for the automatic selection of k value is described in the next section (7.2.4.4).

The classification task uses the pam() function from the *cluster* R package (Maechler *et al.*, 2023), which returns a list object with the clustering result and relevant metrics. The function define_rb() keeps the pam() output in a dedicated column (“pam_object"), with a list per sample, because it applies the algorithm for each sample at a time. Additionally, one column with the classification result (“Levels”) and several columns with metrics to evaluate the clustering result: “Evaluation”, “median_Silhouette”, “Shilhouette_scores” and “Cluster_median_abundance”. Any additional features in the input are not changed. Alternatively, it is possible to include only the classification result in the table (setting *simplified* = *TRUE*). Note that the original table is always returned in a long format, even if the simplified option is set to true, guaranteeing easy piping with follow-up functions in the user analysis pipeline. Setting simplified to TRUE is not a faster option, it only reduces the number of columns, which might be a useful option if such columns are not of interest to the user.

Finally, the function define_rb() can trigger a warning related to the quality of the clustering. This warning indicates which samples obtained a bad Silhouette score, which was defined as a clustering where half the observations obtained less than 0.5 Silhouette score (do not confuse with average Silhouette score). Generally, this warning can be ignored, but it can indicate room for improvement of the clustering result, for example, by changing the value of k. If the automatic option is set to true, then, by default, the selected k will be the one with fewer warnings, if any. This is because the warning is tied to the Silhouette scores, which are used for the automatic selection of k. This, however,

can be changed by the “index” argument (to either Calinski-Harabasz or Davies-Bouldin index).

7.2.4.4. Automatic selection of k

If the user sets *automatic* = *TRUE* in the define_rb() function, the number of clusters is automatically obtained. To do so, define_rb() calls suggest_k() with default parameters. Then, suggest_k() returns an integer with the optimal k value, which is used by define_rb(). The user can specify optional parameters for the automatic selection of k, by changing the parameters of suggest_k() directly inside the define_rb() functions. To obtain the best k value, the function suggest_k() uses the evaluate_k() function, which uses the function evaluate_sample_k(), to obtain clustering statistics on each possible k value, for all samples. When multiple samples are used, different optimal k values might be obtained for different samples. Thus, suggest_k() chooses a consensus k value, by taking the average of k values obtained and rounding to the closest integer.

The user can optimize the k value based on the following metrics: average Silhouette score, Calinski-Harabasz, and Davies-Bouldin index. More detail on these indices and how they are calculated in the next sections (7.2.4.5 and 7.2.6).

7.2.4.5. Unsupervised clustering statistics

Besides the default statistics provided by define_rb() while doing the classification task, there are additional functions that provide a deeper understanding of the clustering result. If the default parameters are used, *i.e.*, with k = 3, then these functions are unnecessary. If, however, the user is considering another k value, then we recommend the utilization of these functions. Briefly, this group includes functions to calculate the three indices: Calinski-Harabasz index, by check_CH(); Davies-Bouldin Index, by check_DB(); average Silhouette score, by check_avgSil(). These three functions have similar arguments and options, changing only the specific metric they calculate. They take as input the same long format table that define_rb() accepts and return a vector with the score for each k, of a given sample. Thus, the index is relative to the entire clustering result, not the score of a specific observation. Instead of the vector, the output can be a plot, showing the score as a function of k. By default, the functions compare all k values between 2 and 10, but the user can test all possible k values.

The function evaluate_sample_k() wraps around the functions check_CH(), check_DB() and check_avgSil(), for a single sample. Like each one of its components, it can return

a plot with each index in a grid, or it can return a data frame with a vector for each index. The default settings are similar, thus comparing from k = 2 up to k = 10. To make this comparison across all samples, the function evaluate_sample_k() can be repeated any number of times by the function evaluate_k(). The evaluate_k() function returns the same as evaluate_sample_k(), but the data frame includes the results of all samples and plots centrality metrics (mean and standard deviations) of the scores obtained for each index in grids, as a function of k (same defaults).

Finally, the suggest_k() function uses the evaluate_k() function to obtain all indices for a range of k values (from 2 to 10, by default) and suggests, as the name implies, what is the best possible k value, based on one of the three indices (average Silhouette score, by default). This function returns an integer with the number of clusters (the k value), if *detailed=FALSE*, or a list with full details on the indices for all samples and for all k values tested, including some explanation of each index.

7.2.4.6. Visualization

After the classification task, independently of the method used to select k, *ulrb* provides functions to view the clustering result in its domain context. Specifically, the function plot_*ulrb*() plots both a RAC, colored by the classification obtained, and a traditional Silhouette score for a given sample. Alternatively, it is possible to plot all samples at once (or a specific group of samples), using centrality metrics to represent each observation along those samples (mean and standard deviations). Thus, plot_*ulrb*() provides a simple analysis tool for both the ecologist (with the RAC) and the data scientist (the Silhouette plot). The plot_*ulrb*() function wraps around the functions plot_*ulrb*_clustering(), which plots the RAC, and the plot_*ulrb*_silhouette(), which plots the Silhouette plot. The plots used the *ggplot2* R package (Wickham, 2016).

7.2.5. Implementation of pam to any number of samples by define_rb() function

The pam algorithm (Kaufman and Rousseeuw, 1987) was used for the clustering in the define_rb() function, using the function with the same name, pam(), from the *cluster* R package (Maechler *et al.*, 2023). The pam algorithm (Kaufman and Rousseeuw, 1990) is divided into two phases, build and swap. At the build phase, k clusters are formed, by selecting one cluster representative (that is, an observation that minimizes the sum of dissimilarities with the remaining observations). At the swap phase, the clustering result is improved by swapping the cluster representative and calculating the dissimilarity

between each observation and the closest centroid. More details of the implementation of this algorithm in R are available in the *cluster* R package (Maechler *et al.*, 2023) and the documentation page of define_rb().

To apply the pam() function to any number of samples, the function define_rb() groups and nests the long format table by sample, using *dplyr* (Wickham *et al.*, 2023) and *tidyr* (Wickham, Vaughan and Girlich, 2023). Then, define_rb() runs the pam() function for each sample, using the map() function from the *purrr* R package (Wickham, Henry and RStudio, 2023). The relevant pam() outputs are unnested into new columns.

The levels obtained by the pam algorithm are randomly ordered. To reorder them meaningfully, define_rb() calculates the median abundance of each level. Thus, the first level is the “rare” classification, and so forth. If the k value is not the default, the user must name the classifications manually. However, they are still ordered by their median abundance.

7.2.6. Explanation of evaluation metrics

In *ulrb*, three different metrics are implemented to evaluate the statistical strength of the clustering algorithm: average Silhouette score (Rousseeuw, 1987), Calinski-Harabasz index (Calinski and Harabasz, 1974) and Davies-Bouldin index (Davies and Bouldin, 1979).

To implement these metrics in *ulrb*, the average Silhouette score is calculated using the pam() function from the *cluster* R package (Maechler *et al.*, 2023). The Calinski-Harabasz index is calculated using the index.G1() function from the *clusterSim* R package (Walesiak and Dudek, 2020), and the Davies-Bouldin index is calculated using the index.DB() function from the *clusterSim* R package (Walesiak and Dudek, 2020).

7.2.6.1. Silhouette score and average Silhouette score

Consider *a(i)* to be the mean distance between taxon *i* and all other taxa from its own cluster, and that *b(i)* to be the distance between taxon *i* and the centroid of the closest cluster. Then, we can describe the Silhouette score, *S*, of a taxon, *i,* with Equation 7.1 (Rousseeuw, 1987).

$$S(i) = \frac{b(i) - a(i)}{max\{a(i),b(i)\}}$$

Equation 7.1. Silhouette score formula.

The Silhouette score of a taxon can vary between -1 and +1. Silhouette scores closer to +1 indicate that the taxon is closer to the center of the respective cluster, while scores close to zero indicate that the taxon is as close to its own cluster as it is to the next closest cluster, and scores close to -1 indicate that the taxon is closer to the center of a different cluster than its own. By taking the average of the Silhouette scores obtained by all taxa in a specific cluster, we obtain a metric of the overall quality of that specific cluster. Accordingly, it is possible to calculate the average Silhouette score over all taxa and all clusters, providing information on the quality of the entire clustering result. Thus, by comparing the average Silhouette score obtained by different numbers of clusters (different k), the higher the average Silhouette score, the better.

7.2.6.2. Calinski-Harabasz index

The Calinski-Harabasz index (CH) divides inter-cluster dispersion by intra-cluster dispersion, providing information on separation and density (Calinski and Harabasz, 1974).

$$CH = \frac{inter\ cluster\ dispersion}{intra\ cluster\ dispersion}$$

Equation 7.2. Calinski-Harabasz index (simplified).

From Equation 7.2, it follows that if inter-cluster dispersion is high, reflecting high cluster separation, then CH increases; if intra-cluster dispersion is high, reflecting low cluster density, then CH decreases. Thus, high values of CH indicate high separation and density, while low CH values indicate the opposite (low separation and low density). If CH values are compared as a function of k, then the best k maximizes CH.

7.2.6.3. Davies-Bouldin index

The Davies-Bouldin index (DB) calculates the mean similarity between clusters, *i.e.*, provides cluster separation information and can be expressed by Equation 7.3 (Davies and Bouldin, 1979).

$$DB = \frac{1}{N}\sum_{i=1}^{N} R_{i,j}$$

Equation 7.3. Davies-Bouldin index.

Where *N* is the number of taxa and $R_{i,j}$ is the maximum similarity between any combination of cluster *i* and *j,* with $i \neq j$. Lower values of DB indicate lower similarity between clusters. Thus, if DB values are compared as a function of k, then the best k minimizes DB.

7.2.7. Data available in *ulrb* R package

The *ulrb* R package includes a small dataset for testing functions and provides examples. This dataset contains an abundance table and environmental information, which have been previously processed and described (de Sousa *et al*., 2019; Pascoal *et al.*, 2022a) (Chapter 5). In the *ulrb* R package, the N-ICE dataset (see details below) is available in different stages of processing in the following objects: *nice*, *nice_env*, *nice_raw*, and *nice_tidy.*

Briefly, the *nice* and *nice_tidy* objects correspond to the same underlying abundance table, with OTU as a species proxy, OTU abundance scores, and sample information. The *nice* object is a table in wide format, while the *nice_tidy* is the same table, but in long format (also known as tidy format). The *nice_env* object is a data frame with environmental information about each sample in the dataset. Finally, *nice_raw* is the original abundance table obtained from the public repository (MGnify reference: MGYS00001922).

7.3. Results

7.3.1. Standard example of *ulrb* application

To illustrate the applicability of *ulrb*, we used the N-ICE dataset, which consists of nine samples of V4V5 16S rRNA gene sequencing data, collected in the Arctic Ocean and previously described (de Sousa *et al*., 2019; Pascoal *et al.*, 2022a) (Chapter 5). The original dataset is available in a domain-specific platform, MGnify (Mitchell *et al.*, 2019), which provides an OTU abundance table in wide format, with OTUs in rows and samples in columns, including additional features, such as taxonomic information. This dataset is available in the *ulrb* R package in both wide and long format (section 7.2.7). The following code can be used to load and transform the original wide data into the required long format:

```
> library(ulrb)
> library(dplyr)
> # load wide data
> data("nice")
> # make a vector with sample names
> nice_samples <- names(nice)[1:9]
> # transform wide into long format data
> nice_long <- prepare_tidy_data(nice, sample_names = nice_samples)
```

Then, we can use define_rb() with default parameters to divide OTUs into rare, undetermined, and abundant. This can be done with the following code:

```
> # classify OTUs
> nice_classified <- define_rb(nice_long)
```

This step adds new columns with the classifications obtained for each OTU, per sample, as well as other metrics, like the Silhouette score of each OTU. Then, it is possible to directly inspect the result with the plot_*ulrb*() function, with the following code:

```
> # verify results
> plot_ulrb(nice_classified,
>         plot_all = TRUE,
>         taxa_col = "OTU",
>         log_scaled = TRUE)
```

To plot all the samples together, we used the argument *plot_all* = *TRUE* and to set the log10 scale we used the argument *log_scaled* = *TRUE*. The standard output is presented in Figure 7.3.

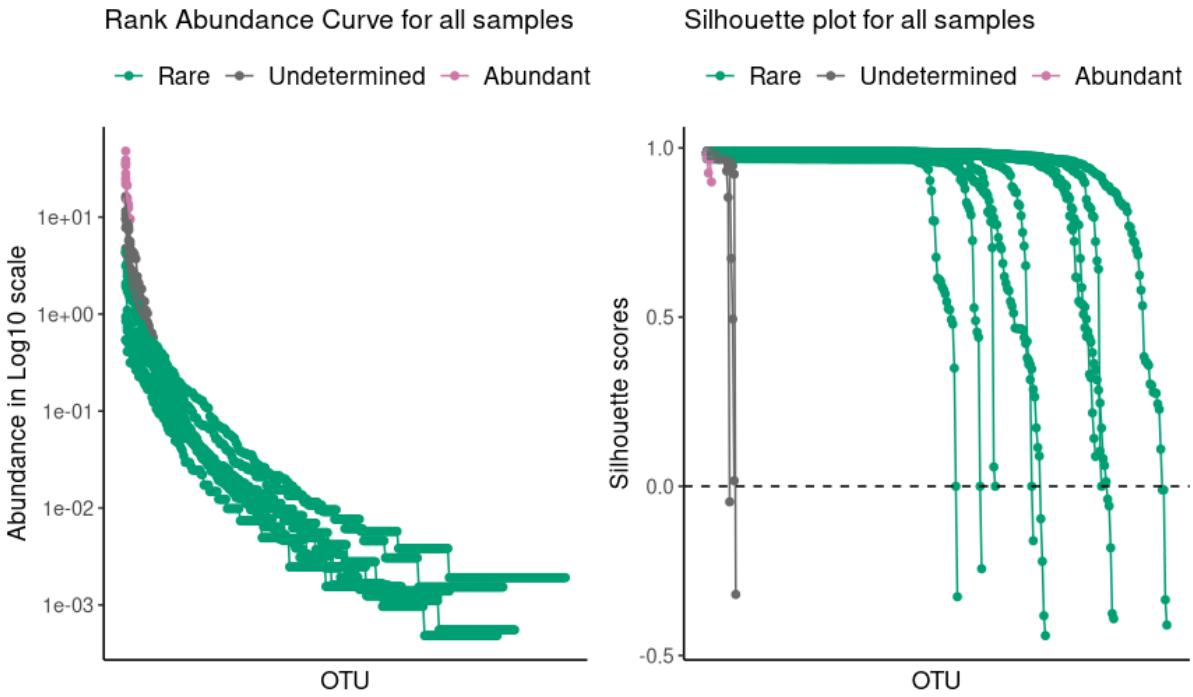


Figure 7.3. Default plot from plot_*ulrb*() function. The left grid shows the OTU relative abundance, on a Log10 scale, ordered from the most to the least abundant. The right grid shows the Silhouette score for each OTU, ordered from highest to lowest Silhouette score. OTUs are colored by classification and grouped by sample, with lines.

The overview plot (Figure 7.3) is intended to help the user evaluate the result in general, but different analyses can be made by the user, if necessary. Furthermore, for illustrative purposes, we are not concerned with normalization aspects of the underlying data, which could be important for the user. Additionally, note that, by default, the plot_*ulrb*() function uses a centrality metric, because it is intended for a dataset of any size. Naturally, for low sample sizes, the centrality metric is not very informative per se, but we still think it is necessary to give an outlook on all samples together. If the user wants to inspect a specific sample, that is possible with the following code:

```
> # inspect sample 1 (ERR2044662)
> plot_ulrb(nice_classified,
>           sample_id = nice_samples[1],
>           plot_all = FALSE,
>           taxa_col = "OTU",
>           log_scaled = TRUE)
```

The above code resulted in Figure 7.4.

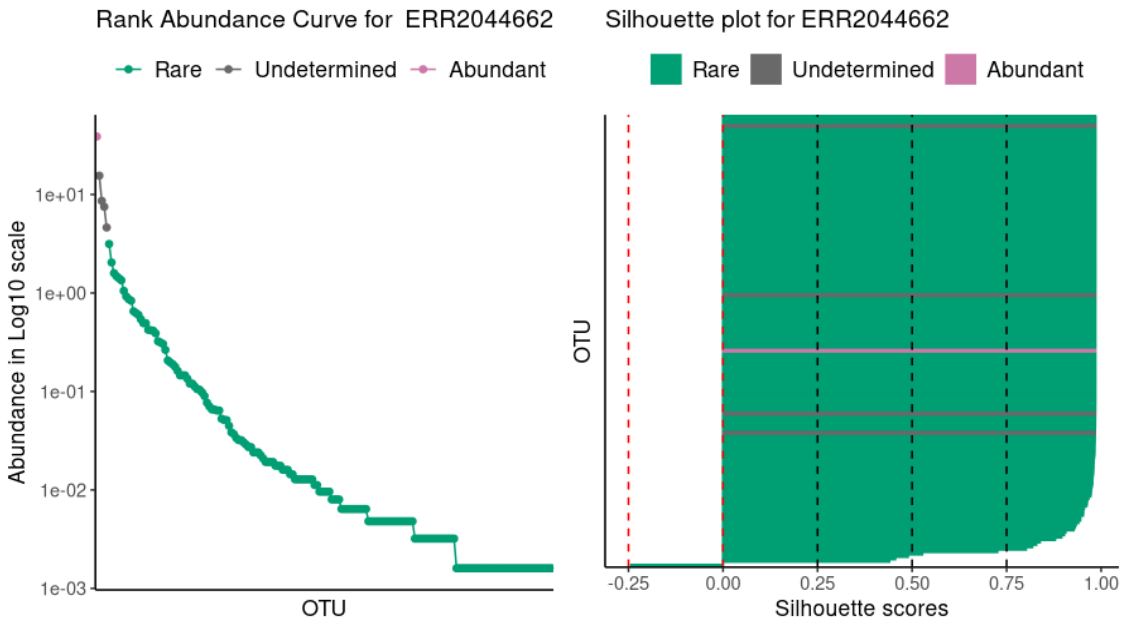


Figure 7.4. Example of plot_*ulrb*() output for one single sample. The left grid shows the abundance of each OTU, on a Log10 scale, ordered from most to least abundant. The right grid shows the Silhouette score of each OTU, ordered from highest to lowest Silhouette score. Each OTU was colored based on its classification. The sample ERR2044662 was used as an illustrative example.

By expecting a single sample, it is possible to see more clearly that most OTUs are clustered in the rare biosphere, a few are clustered in a grey area where they are neither rare, nor abundant (undetermined classification) and very few, only one in this example, are considered abundant relative to the others. The abundance is illustrated in a logarithmic scale, to allow visualization of very small relative abundance values. Overall, the RAC obtained by *ulrb* for a single sample and all samples was consistent with the standard ecological interpretation, because most OTUs were classified as rare and only very few were classified as abundant.

The quality of the clustering was also inspected in Figures 7.3 and 7.4. Both looking at all samples together and looking at a specific sample (as an example), most of the OTUs obtained a high Silhouette score (> 0.75), showing that *ulrb* presented reliable clusters.

7.3.2 Inspecting alternative number of classifications

The default number of classifications in *ulrb* can be modified by the user. To do so, the user might decide on a pre-set number of classifications that are useful for the experimental design; or they might analyze statistical metrics to optimize certain characteristics of the underlying clusters (see section 7.2.3).

Considering the first option, let us assume that a research group is interested in a fine-grained analysis of abundance, so they decide on the following classifications: “extremely rare”, “somewhat rare”, “undetermined”, “somewhat abundant”, “extremely abundant”. In this example, there are 5 different classifications, thus the underlying k will be 5. To apply this change, the user needs only to specify those classifications and *ulrb* will automatically assume k = 5, like so:

```
> # make new classification vector
> # new classifications are ordered from lowest to highest abundance
> new_classifications <- c(“extremely rare”,
>                          “somewhat rare”,
>                          “undetermined”,
>                          “somewhat abundant”,
>                          “extremely abundant”)
> # classify OTUs, based on new classifications
> nice_new_classification <- define_rb(nice_long, classification_vector = new_classifications)
```

Again, we can inspect the result with a standard plot (Figure 7.5), like so:

```
> # make vector of colors for new classifications
> colors_new_classifications <- c("darkgoldenrod4", "darkgoldenrod3",
>                                 "grey41",
>                                 "deepskyblue1", "deepskyblue4")
> # inspect new classifications (all samples)
> plot_ulrb(nice_new_classification,
>           plot_all = TRUE,
>           taxa_col = "OTU",
>           log_scaled = TRUE,
>           colors = colors_new_classifications)
```

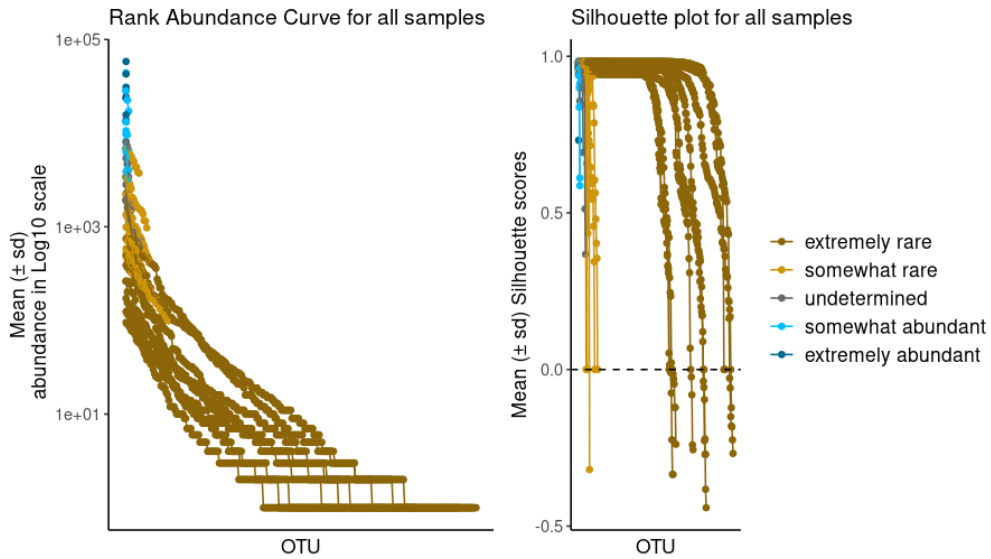


Figure 7.5. Standard plot_*ulrb*() output for user-defined classification. The left grid shows the OTU relative abundance, on a Log10 scale, ordered from the most to the least abundant. The right grid shows the OTU Silhouette score, ordered from highest to lowest. OTUs are colored by their classifications, from 1 to 5, and grouped by sample with lines.

The RAC obtained for the user-selected classifications (Figure 7.5) was not as well defined as the one obtained with default classifications (Figure 7.3) but allowed for a more fine-grained analysis of the abundance variation. For instance, the rare biosphere was divided into two groups with many OTUs; the undetermined OTUs were in the area where “somewhat rare” and “somewhat abundant” OTUs overlap and very few OTUs were divided into “somewhat abundant” and “extremely abundant”. Based on the silhouette scores obtained, the alternative classifications (Figure 7.5) were of lower quality than the default classifications (Figure 7.3). Thus, the addition of new classifications made the underlying clusters less well-defined. Additionally, note that adding so many labels into the legend complicates plot readability in Figure 7.5. For the sake of completeness, we also made the same plot for a single sample (Supplementary Figure 10.37). This analysis makes it clear that the “original” rare biosphere (defined using the default settings) was divided into two new classifications, “extremely rare” and “somewhat rare”, and some of the undetermined OTUs migrated to the “somewhat abundant” classification (Supplementary Figure 10.37).

Another alternative to deciding the number of classifications is to apply a specific quality metric to find the optimal k value. This is possible to do automatically with define_rb(), using the average Silhouette score as a parameter, but the Davies-Bouldin and Calinski-Harabasz indices are also available. However, for illustrative purposes, we will show the code that could be used to manually inspect those metrics, for a standard range of k (from 3 to 10, see section 7.2.3):

```
> # calculate quality metrics for each k
> evaluate_k(nice_long, with_plot = TRUE)
```

The plot obtained by the evaluate_k() function is in Figure 7.6. Accordingly, within the range of 3 to 10 clusters, the average Silhouette score presented a contrary pattern to Davies-Bouldin and Calinski-Harabasz indices (Figure 7.6), because the average Silhouette score favored lower values of k while Davies-Bouldin and Calinski-Harabasz favored higher number of clusters (Figure 7.6). Recall that for the average Silhouette score and the Calinski-Harabasz index the best k maximizes the score value while for Davies-Bouldin the best k minimizes the score (see section 7.2.6). The original values could be obtained by setting *with_plot = FALSE* (Supplementary Figure 10.38).

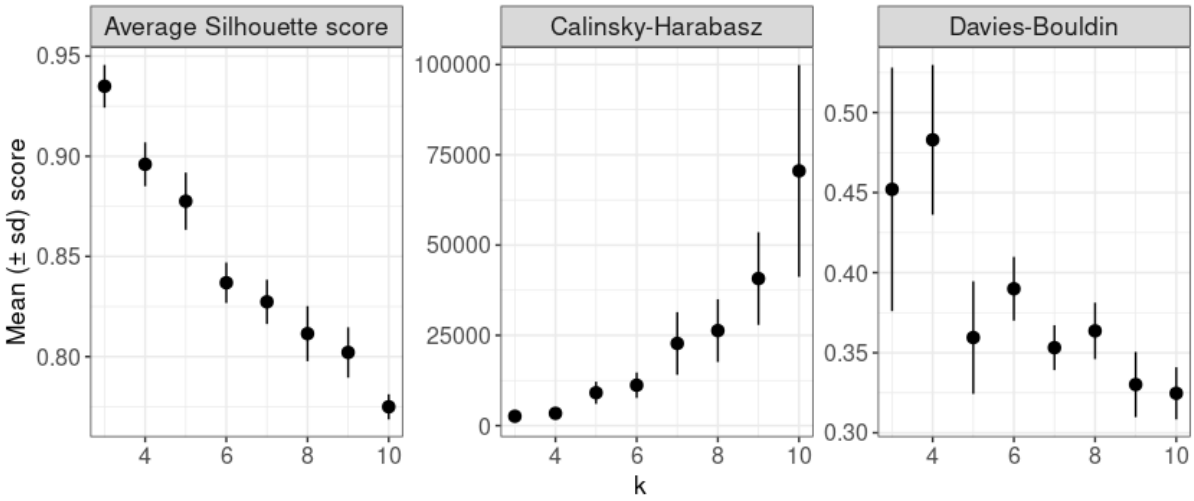


Figure 7.6. Manual evaluation of k using the function evaluate_k(). Each grid represents an index (from left to right: average Silhouette score, Calinski-Harabasz, and Davies-Bouldin) as a function of k. For each index, the mean and standard deviation are calculated (n = 9 samples).

The analysis of Figure 7.6 does not suffice to decide the best k, because it groups several samples with centrality metrics. Considering that each sample will have a different optimal value, the suggest_k() function can identify a consensus value of k for a given index (recall section 7.2.4.4). This consensus value is the average k of all optimal k's and can be obtained as follows:

```
> # get consensus k for all samples, based on average Silhouette score
> suggest_k(nice_long)
> # get consensus k for all samples, based on Davies-Bouldin index
> suggest_k(nice_long, index = "Davies-Bouldin")
> # get consensus k for all samples, based on Calinski-Harabasz index
> suggest_k(nice_long, index = "Calinski-Harabasz")
```

According to the suggest_k() function, the consensus k for each index would be: k = 3 (average Silhouette score); k = 6 (Davies-Bouldin); and k = 9 (Calinski-Harabasz). Arguably, using nine classifications is harder to interpret than 3 classifications, but the user can experiment with such options. To verify the results, the user could attribute the following nine classifications as an example: "extremely rare", "very rare", "rare", "somewhat rare", "undetermined", "somewhat abundant", "abundant", "very abundant" and "extremely abundant". To verify the results of this example, it could be done like so:

```
> # make a vector with nine classifications
> # ordered from least to most abundant
> nine_classifications <- c("extremely rare",
>                           "very rare",
>                           "rare",
>                           "somewhat rare",
>                           "undetermined",
>                           "somewhat abundant",
>                           "abundant",
>                           "very abundant",
>                           "extremely abundant")
> # Repeat classification for the new classifications (k = 9)
> nice_nine_classifications <- define_rb(nice_long,
>                                        classification_vector = nine_classifications)
> # Make a color vector for the new classifications
> nine_colors <- c("darkgoldenrod4", "darkgoldenrod3",
>                  "darkgoldenrod2", "darkgoldenrod1",
>                  "grey41", "deepskyblue1", "deepskyblue2",
>                  "deepskyblue3", "deepskyblue4")
> # plot results for nine classifications
> plot_ulrb_clustering(nice_nine_classifications,
>                           plot_all = TRUE,
>                           taxa_col = "OTU",
>                           colors = nine_colors,
>                           log_scaled = TRUE)
```

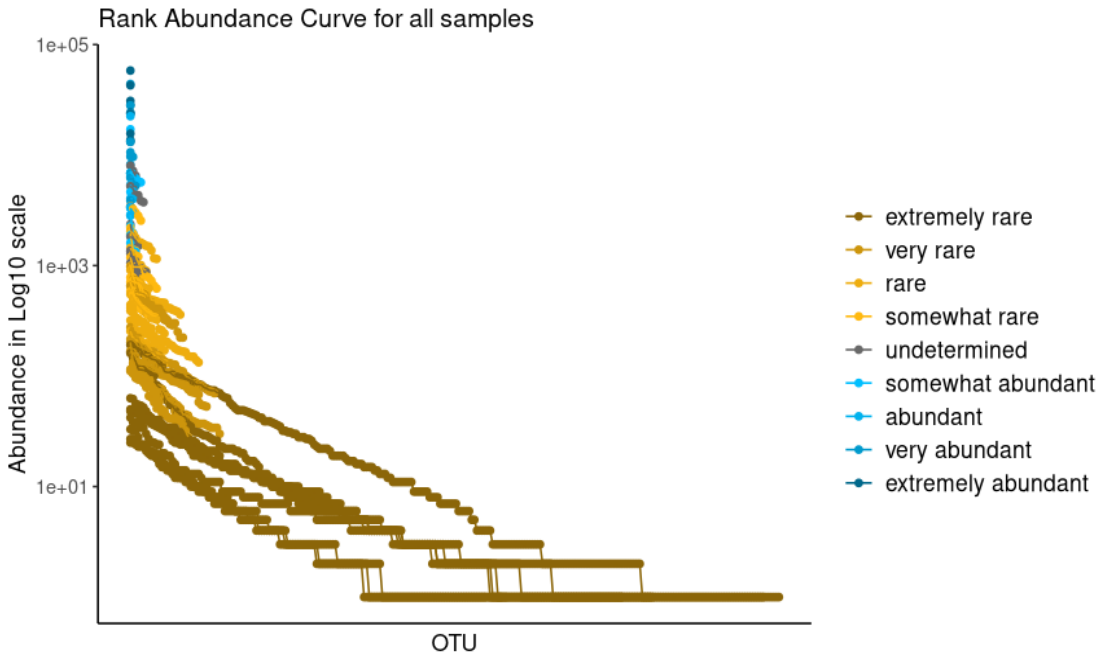


Figure 7.7. OTU abundance score, on a Log10 scale, ordered from the most to the least abundant. OTUs are colored according to their abundance classification and grouped by lines according to their sample.

The example with nine different classifications does not show the Silhouette score results, because the nine classifications were selected based on the Calinski-Harabasz index, which measures quality differently. The clustering result (Figure 7.7) showed that some classifications were substantially overlapped, making it unclear if they provided any new information. For example, the classifications “somewhat rare”, “undetermined” and “somewhat abundant” overlapped. The only classification that became more well-defined was “extremely rare”.

7.4. Summary and conclusion

The *ulrb* R package is an open-source software implemented in R, allowing for (1) the transformation of wide abundance table formats into long/tidy format, (2) the automatic classification of taxa into abundance classifications based on their abundance score, and (3) evaluating the classification results with relevant statistics and visualization tools. Using *ulrb*, we show that it is possible to automatically delineate rare, undetermined, and abundant taxa within biological communities without a human-predefined threshold of rarity. Alternatively, *ulrb* provides tools and options to use other abundance classifications that might work better for different experimental designs. Furthermore, the coding style used in *ulrb* follows tidy data principles (Wickham, 2014), which makes *ulrb* easy to integrate into other analysis pipelines.

As far as we know, no other R package uses unsupervised machine learning to define the rare biosphere of microbial and non-microbial datasets. Note that the most similar R package we could find was *FuzzyQ* (Balbuena *et al.*, 2021), but *FuzzyQ* uses a different approach, which is based on two features instead of one and results in a different classification (rare and common).

In future work, we will attempt to incorporate the concept of types of rarity (Lynch and Neufeld, 2015; Jia, Dini-Andreote and Falcão Salles, 2018) into *ulrb*. The types of rarity indicate how the classifications of a given taxon vary across samples. For example, a taxon that varies between rare and abundant across a time window could be defined as conditionally rare. Such types of rarity can be easy to define in specific experimental designs (Pascoal *et al.*, 2025a), however, to add this capability to *ulrb* we need an algorithm that can define the types of rarity for any study, which is challenging.

7.5. Declarations

Author contribution

Conceptualization (FP, PB, RC and CM); data curation (FP); formal analysis (FP); funding acquisition (all authors); investigation (FP and PB); methodology (FP, PB and LT); software (FP, PB and LT); supervision (PB, RC, LT, CM); validation (FP); Visualization (FP); Writing - original draft (FP); Writing - review & editing (all authors).

Financing

FCT funded this study through two grants to CM (PTDC/CTA-AMB/4946/2020; 2022.02983.PTDC) and through a PhD grant to FP (2020.04453). This research has also been supported by national funds by FCT through the projects UIDB/04565/2020 and UIDP/04565/2020 of iBB and the project LA/P/0140/2020 of i4HB. The work of Paula Branco was undertaken, in part, thanks to funding from the NSERC through a Discovery Grant.

Conflict of interests

The authors declare no conflict of interest.

Data availability

Raw data for the NICE dataset is available in the European Nucleotide Archive (accession number: PRJEB21950) and the OTU table is available in the MGnify platform (accession number: MGYS00001922). The wide and long format species abundance tables used are available in the *ulrb* R package.

Code availability

The R code used for the illustrative examples is available in a GitHub repository (https://github.com/pascoalf/*ulrb*_software_paper). The R code for the *ulrb* R package is available on CRAN (stable) (https://cran.r-project.org/web/packages/*ulrb*/index.html) and GitHub (development) (https://github.com/pascoalf/*ulrb*). Tutorials are available at https://pascoalf.github.io/*ulrb*/index.html.

Part IV. Ecology of the microbial rare biosphere in the Arctic Ocean.

Chapter 8. Combining unsupervised and rule-based machine learning to obtain novel insights into the prokaryotic rare biosphere response to a changing Arctic Ocean.

8. Combining unsupervised and rule-based machine learning to obtain novel insights into the prokaryotic rare biosphere response to a changing Arctic Ocean

Francisco Pascoal[1,2], Paula Branco[3], Rodrigo Costa[4,5] and Catarina Magalhães[1,2]

[1]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, rua do Campo Alegre s/n, 4169– 007 Porto, Portugal; [2]Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal; [3]School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa, Ottawa, Ontario, Canada; [4]Department of Bioengineering, Institute for Bioengineering and Biosciences (iBB), Instituto Superior Técnico, University of Lisbon, Portugal; [5]Institute for Bioengineering and Biosciences (iBB) and i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, University of Lisbon, Portugal.

Abstract

Climate change has a magnified impact on the Arctic Ocean, causing important shifts in the ecosystem. Prokaryotes are major engines of ecosystem functioning, mediating nutrient cycling and organismal interactions, but remain overlooked in the Arctic Ocean. In this study, we combined traditional microbial ecology analysis with unsupervised and rule-based machine learning approaches to derive more powerful insights into the prokaryotic community dynamics in the Arctic Ocean. We used a multi-year dataset (2016-2020), consisting of seawater samples collected at fixed stations with standardized methodologies for V4V5 16S rRNA gene amplicon sequencing and multiple environmental variables (depth, water mass, temperature, ammonium, nitrate, nitrate, nitrite, chlorophyll *a*, phosphate, silicates and salinity). Our results show that surface waters and higher temperatures were correlated with less rare prokaryotic diversity than lower temperatures and deeper water layers. By using unsupervised and rule-based machine learning, we were able to identify associations between specific taxonomic groups, environmental variables and abundance classifications. For example, taxa from genus *Clade Ia* (*SAR11* clade) were abundant in association with the samples with the highest nitrite and silicates concentrations. Overall, different environmental associations were found between rare, undetermined and abundant taxa, with scarcity of nutrients (except nitrites) and low temperature being associated with the rare microbial biosphere, while higher temperatures and higher concentration of nutrients like phosphate or nitrites

being associated with abundant taxa. The association rule mining results show that the classification obtained by unsupervised machine learning reflects relevant ecological patterns. Collectively, our study illustrated the potential of combining multiple machine learning approaches on multi-year datasets to obtain novel insights into microbial ecology.

Keywords: rare biosphere, association rule mining, machine learning, microbial ecology, microbiome, climate change.

8.1. Introduction

Climate change is magnified in Arctic biomes, due to the Arctic amplification effect (Box *et al.*, 2019; Rantanen *et al.*, 2022). The Arctic Ocean is subjected to both Atlantification and Pacification, whereby the Arctic waters become more similar to ocean waters from temperate regions (Polyakov *et al.*, 2017, 2020; Tesi *et al.*, 2021). Svalbard fjords and the Fram Strait are a particularly relevant region to study the impact of climate change on the Arctic Ocean, because they are characterized by environmental conditions already more temperate than other regions at similar latitudes (Tesi *et al.*, 2021). The Kongsfjorden (a Svalbard fjord) has been suggested as a natural laboratory to address climate change impacts on marine ecosystems (Wiencke and Hop, 2016; Assmy *et al.*, 2023). In fact, monitoring stations have been established from the coastal area within the Kongsfjorden up to the open sea area of the Fram Strait, revealing important temporal trends on phytoplankton, zooplankton, and prokaryote community dynamics (Wiencke and Hop, 2016; Assmy *et al.*, 2023; Wietz *et al.*, 2024).

Prokaryotes are cornerstones of ecosystem functioning as they mediate biogeochemical cycles and establish foundational interactions with eukaryotic organisms. Most species within prokaryotic communities belong to the microbial rare biosphere (Sogin *et al.*, 2006), which has the potential to work as a seed bank, ensuring ecosystem resistance and resilience (Jousset *et al.*, 2017). Thus, the rare biosphere has been addressed in many Arctic Ocean surveys (Galand *et al.*, 2009; Hubert *et al.*, 2009; Alonso-Sáez, Díaz-Pérez and Morán, 2015; Fernández-Gómez *et al.*, 2019; Cramm *et al.*, 2021; Kampouris *et al.*, 2023), including in the Fram Strait region (Carter-Gates *et al*., 2020; Thomas, Mohan and Krishnan, 2021; Pascoal *et al.*, 2022a) (Chapter 5). Collectively, these studies have found evidence of seed bank taxa, with the potential to respond to changing conditions (Hubert *et al.*, 2009; Cramm *et al.*, 2021; Thomas, Mohan and Krishnan, 2021). Additionally, the rare biosphere has been found to be highly influenced by water mass and geography (Galand *et al.*, 2009; Pascoal *et al.*, 2022a) (Chapter 5). It has also been noted that taxonomic groups found in specific geographical regions were mostly rare, while abundant groups were more cosmopolitan (Galand *et al.*, 2009; Fadeev *et al.*, 2018).

Due to the importance of prokaryotes to understand the impact of climate change on the Arctic Ocean, several multi-year surveys of prokaryotic diversity have been made, based on amplicon sequencing of short regions of the 16S rRNA gene (Priest *et al.*, 2023; Hörstmann *et al.*, 2024; Kraemer *et al.*, 2024; Wietz *et al.*, 2024). More specifically, Wietz

et al. (2024) published a dataset that ranges from the inner Kongsfjorden to the East and West Fram Strait, spanning five years (2015-2019), which allowed the identification of signature populations and highlighted the impact of variables like sampling region, sea-ice cover, temperature and depth (Wietz *et al.*, 2024). The dataset by Hörstmann *et al*. (2024) spans 6 years (2014-2019), but sampling was carried out at six distinct geographical locations (South Norway/Sweden, North Norway, Iceland, Svalbard, East and West Greenland), which complicates the ability to distinguish local from temporal effects. Kraemer *et al*. (2024) published a timeseries of Canada Basin, spanning 9 years (2004-2012), and identified an overall decrease of bacterial diversity. Finally, Priest *et al*. (2023) published a four year dataset (2016-2020) from the Fram Strait, finding evidence of Atlantification at prokaryotic compositional and functional levels (Priest *et al.*, 2023). To obtain functional information, the former study also employed metagenomic data analysis, besides the amplicon sequencing of the 16S rRNA gene (Priest *et al.*, 2023). Even though all the above mentioned datasets focused mainly on the epipelagic layer, they identified depth related patterns of diversity (Hörstmann *et al.*, 2024; Kraemer *et al.*, 2024; Wietz *et al.*, 2024), consistent with the wide agreement that the composition of Arctic Ocean prokaryotes is very stratified in the water column (Wilson *et al.*, 2017; Fadeev *et al.*, 2018; Müller *et al.*, 2018; Quero *et al.*, 2020; Cardozo-Mino *et al.*, 2021; Priest *et al.*, 2023; Vipindas *et al.*, 2023).

Besides multi-year analyses, many other studies have focused on geographical or local aspects of the Fram Strait and Svalbard ecosystems. On the terrestrial side, studies have addressed the impact of glacial meltwater on the fjord system (Han *et al.*, 2024; Yang, Li and Yin, 2024) and nearby sediments (Tian *et al.*, 2023), highlighting the impact of lowering salinity in fjord microbial communities and functions. Consistent with a study carried out during melt season across a range of river, intertidal, subtidal and fjord sediments of Svalbard, which found evidence that freshwater discharge modified community composition (Handler *et al.*, 2023). These results are also consistent with a marine microbiome analysis in the Fram Strait, which found salinity to be the most important factor shaping microbial composition (Mioduchowska *et al.*, 2023). A study of permafrost in Svalbard found no important differences in prokaryotic diversity from winter and summer (Loganathachetti *et al.*, 2022). This was consistent with a study of coastal sediments, where taxonomic and functional composition were stable across seasons (Miksch *et al.*, 2024), although the functional analysis was limited to glycan metabolism (Miksch *et al.*, 2024). On the marine side, analysis of microbial communities from meso- and bathypelagic layers of Eastern Fram Strait explained prokaryotic diversity according

to water masses (Quero *et al.*, 2020), consistent with findings using CARD-FISH (catalyzed reporter deposition-fluorescence in situ hybridization) and semi-automated counting on samples from the same region, collected at epi- and bathypelagic layers (Cardozo-Mino *et al.*, 2021). Studies performed in the north of Svalbard, using a fixed vessel on drifting ice, have compared prokaryotic dynamics across the water column and seasons (de Sousa *et al.*, 2019; Pascoal *et al.*, 2022a; Thiele *et al.*, 2022). They found that seasonal changes in sea-ice thickness influenced the composition of microbial functional groups, such as higher abundance of ammonia oxidizing groups at the surface layers under lower light conditions, and migration of these functional groups to deeper layers when light availability increases (de Sousa *et al.*, 2019; Thiele *et al.*, 2022). Another study from the Fram Strait in the summer of 2010 found evidence of increased net productivity in ice leads (García-Martín, Serret and Leakey, 2014).

Besides long-term data of Arctic Ocean microbial communities, powerful statistical analysis is required to detect complex and non-linear interactions. Some of the above-mentioned studies have already approached machine learning strategies. For example, using random forest to identify relevant variables (Kraemer *et al.*, 2024) or unsupervised machine learning to identify the rare biosphere (Pascoal *et al.*, 2025a) (Chapter 6). In other polar environments, rule-based machine learning has proved useful to identify complex and non-linear interactions between metagenomic derived genes in the McMurdo Dry Valleys of Antarctica (Pascoal *et al.*, 2022b), hinting at possible wider applications for microbial ecology data analysis.

The objective of this study is to improve the insights obtained from the rare biosphere and overall prokaryotic community in the context of a changing Arctic Ocean, using standardized data and multiple statistical and machine learning strategies. To survey the prokaryotic community, we used a V4V5 16S rRNA gene amplicon sequencing dataset with samples collected from Kongsfjorden to the Fram Strait, at fixed stations, during summer, from 2016 to 2020 (Pascoal *et al.*, 2025a) (Chapter 6). This dataset was built using standardized sampling methods (Pascoal *et al.*, 2023) (Chapter 3) and bioinformatics (Pascoal *et al.*, 2024) (Chapter 4), allowing direct comparisons between samples. Additionally, the dataset includes metadata for relevant environmental variables, including sampling depth (m), temperature (ºC), PO_4^- (µM), NO_2^- (µM), NO_3^- (µM), Si (µM), NH_4^+ (µM), chlorophyll *a* (Chl *a*) (mg/m^3), salinity (PSU) (Wold, Assmy and Duarte, 2024), and water mass classifications, based on Cottier *et al.*, (2005). We used unsupervised machine learning, with *ulrb* (Pascoal *et al.*, 2025a; 2025b) (Chapters 6 and 7), to divide all taxa into the abundance classifications: “rare”, “abundant”, or

“undetermined” (meaning neither rare, nor abundant). Then, we applied traditional microbial ecology analyzes to describe the general patterns of prokaryotic diversity across the different abundance classifications. Finally, we used rule-based machine learning to automatically extract complex interactions between taxonomy, environmental variables, and abundance classifications. We propose that the combination of unsupervised and rule-based machine learning applied to multi-year prokaryotic community data provides novel insights into rare, undetermined, and abundant taxa.

8.2. Methods

8.2.1. Sampling campaign and available metadata

The sampling campaigns to the Arctic Ocean and seawater sampling have been described in previous studies (Pascoal *et al.*, 2024; Pascoal *et al.*, 2025a) (Chapter 4 and 6). In summary, our team collects seawater samples at fixed stations every year, since 2016, in the framework of MOSJ (Renner, Dodd and Fransson, 2018). Samples were generally collected at three different depths, representing the surface, DCM (Deep Chlorophyll Maximum) and deeper depths, sometimes deeper than 1000 m. For this study, we divided all samples into pelagic layers: epipelagic (0-200 m), mesopelagic (200-1000 m), bathypelagic (> 1000 m). Every year, samples were collected between July and August. Furthermore, due to onsite sampling constraints (like weather conditions, human error or lack of funding), not all stations were available for all years. In Figure 8.1 we provide a map with all the stations where samples were collected.

All our samples include the assessment of environmental parameters, which were collected and processed in previous MOSJ studies by the Norwegian Polar Institute (Wold, Assmy and Duarte, 2024). Briefly, we selected the following environmental variables: sampling depth (m), temperature (ºC), PO_4^- (µM), NO_2^- (µM), NO_3^- (µM), Si (µM), NH_4^+ (µM), chlorophyll *a* (Chl *a*) (mg/m^3) and salinity (PSU). Water masses were classified based on temperature and salinity profiles (Cottier *et al.*, 2005): Atlantic Water (AW), Transformed Atlantic Water (TAW), Surface Water (SW), Winter Cooled Water (WCW), Intermediate Water (IW) and Local Water (LW). The metadata for the samples used in this study are available in Supplementary Data 8.1, adapted from previous studies (Wold, Assmy and Duarte, 2024; Pascoal *et al.*, 2025a) (Chapter 6).

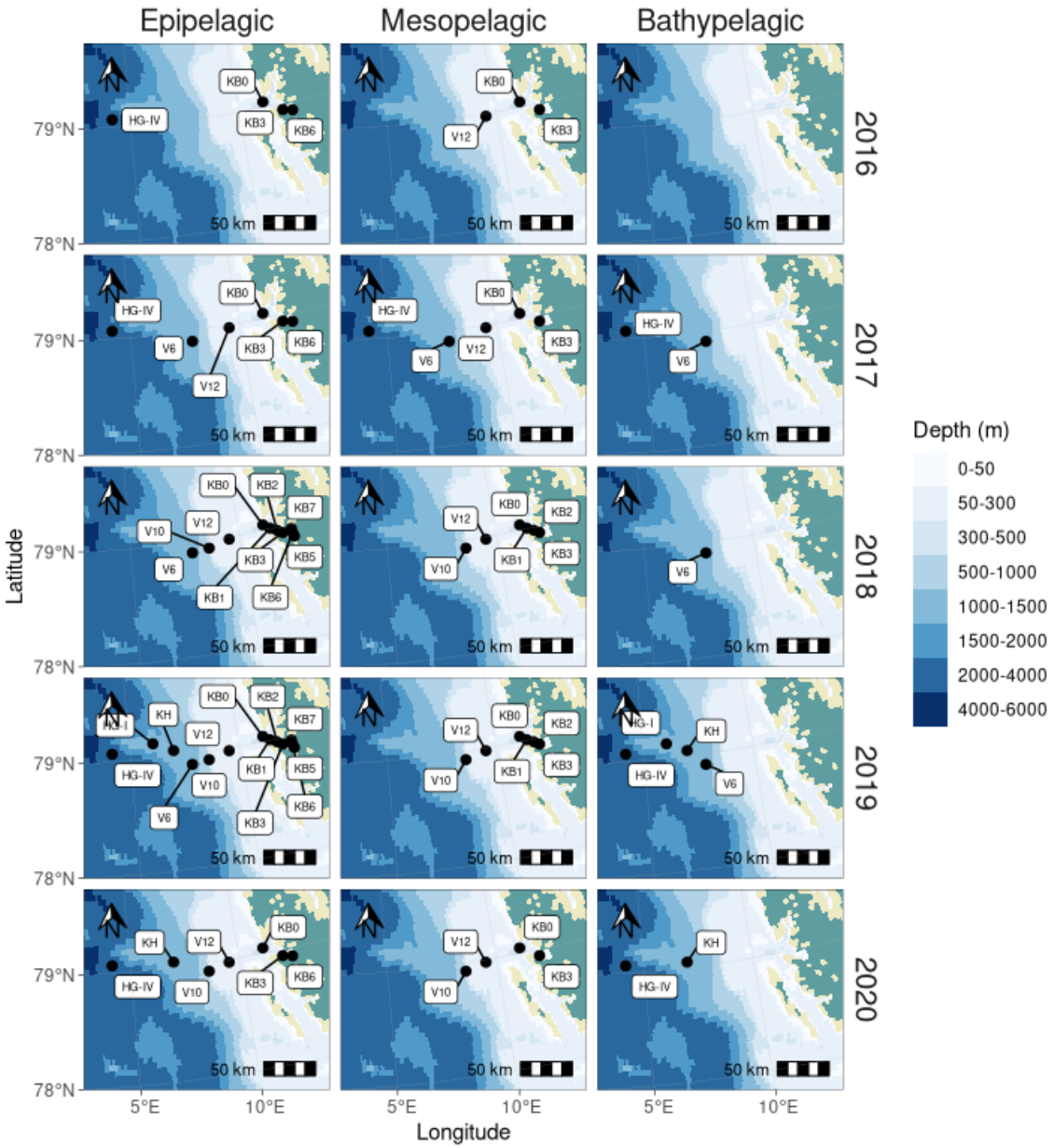


Figure 8.1. Map of the position of the sampling stations between the Fram Strait and Svalbard archipelago. Grids divide the sampling stations used by pelagic layers (columns) and year (rows). The depth of the seafloor is represented in shades of blue. The Svalbard archipelago is represented in light yellow, with the iceberg sheet in green.

8.2.2. Seawater sampling, DNA extraction and amplicon sequencing

Seawater sampling, DNA extraction and amplicon sequencing for the samples used in this study have been described previously (Pascoal *et al.*, 2024, 2025a) (Chapter 4 and 6), following standardized protocols (Kopf *et al.*, 2015; Pascoal *et al.*, 2023) (Chapter 3).

For amplicon sequencing, the V4V5 region of the 16S rRNA gene was selected, with primers optimized for a wide coverage of marine prokaryotes (Caporaso *et al.*, 2011;

Apprill *et al.*, 2015; Caporaso et al., 2012; Parada, Needham and Fuhrman, 2016): 515YF (5′-GTGYCAGCMGCCGCGGTAA-3′) and 926R (5′ - CCGYCAATTYMTTTRAGTTT- 3′). A previous study has confirmed the usefulness of V4V5 for standardized multi-year analysis of Arctic Ocean prokaryotic communities (Fadeev *et al.*, 2021). Sequencing was done by Illumina, with MiSeq platforms (2x300bp), following protocol used in de Sousa *et al*., (2019).

8.2.3. Bioinformatic processing of raw reads and taxonomic assignments

Raw reads were processed into ASVs, because they are comparable across datasets (Callahan, McMurdie and Holmes, 2017), making them ideal for our multi-year dataset. The processing into ASVs was done using the *DADA2* R package (Callahan *et al.*, 2016) and all steps were described previously (Pascoal *et al.*, 2024; Pascoal *et al.*, 2025a) (Chapters 4 and 6). Briefly, default parameters were used whenever possible and reads were trimmed to optimize quality (249nt forward, 214nt reverse). Taxonomic assignments were made using the Naive-Bayesian algorithm (Wang *et al.*, 2007), implemented in *DADA2* (Callahan *et al.*, 2016), with the Silva v138 database (Quast *et al.*, 2012; Yilmaz *et al.*, 2014). Rarefaction curves have been published previously (Pascoal *et al.*, 2025a) (Chapter 6) and revealed sufficient sequencing depth.

Recently, major taxonomic groups have faced nomenclature updates, some of the most relevant changes include the phylum *Proteobacteria*, which is now *Pseudomonadota* (Oren and Garrity, 2021), *Crenarcheota*, which is now *Thermoproteota* (Oren and Garrity, 2021), and *Thermoplasmatota*, which his now *Methanobacteriota* (Göker and Oren, 2023). However, not all taxonomic nomenclature modifications have been updated in the taxonomy databases. We used the taxonomy nomenclature as presented in the databases used, to avoid human error in the manual update of the nomenclature, which is a non-trivial task.

8.2.4. ASV abundance table pre-processing

The ASV abundance table was previously filtered to ensure high-quality (Pascoal *et al.*, 2025a) (Chapter 6). More specifically, ASVs assigned to chloroplasts, mitochondria, eukaryotes and unassigned at phylum level, if any, were removed automatically.

In this study, we grouped ASVs down to genus level, creating taxonomic groups with as much information as possible. This step was added to avoid overestimation of diversity, because ASVs can provide information down to strain level even in environmental

datasets (Pascoal *et al.*, 2024) (Chapter 4). Furthermore, we did not provide taxonomic information at species and subspecies level, since that would require exact matching (Callahan *et al.*, 2016) which is unreliable for short reads. Thus, we grouped ASVs at the highest possible level of taxonomic resolution enabled by the dataset, which is the genus level. If ASVs could not be classified at the genus level, the precedent taxonomic rank (*e.g.*, family, order) was used instead for taxonomic assignments.

We then inspected the distribution of relative abundance of the taxonomic groups, verified if they were present in a single sample and if they were singletons (Supplementary Figure 10.39). Based on this analysis, we decided to remove taxonomic groups with less than 0.01% relative abundance per sample, because they represent a small proportion of the taxonomic groups (18.5%) and are represented by very few reads. Regarding the taxonomic groups that were present in a single sample (1 out of 117 samples), we kept those that had more than 0.01% relative abundance, because we have no reason to believe that they are sequencing artifacts. Finally, we removed samples with less than 10 000 reads and we did not apply rarefaction to the samples, as to avoid removing valid biological information (McMurdie and Holmes, 2014) and always considered relative abundance values. This was a conservative approach that ensures all biological data is valid, at the cost of partially reducing the rare biosphere extent.

We provide a summary of the number of reads before and after bioinformatics raw read processing and abundance table pre-processing, as well as the total taxonomic groups obtained, grouped by year, in Supplementary Table 11.24.

8.2.5. Definition of the rare biosphere with unsupervised machine learning

We used *ulrb* (Pascoal *et al.*, 2025a; 2025b) (Chapters 6 and 7) to delineate the rare biosphere of all samples analyzed in this study. This approach uses unsupervised machine learning to classify taxonomic groups based on their relative abundance within a community/sample. The main advantage of *ulrb* is that it does not require the utilization of arbitrary thresholds and is comparable across studies (Pascoal *et al.*, 2025a) (Chapter 6). Furthermore, the *ulrb* method allows the automatic division of all taxonomic groups into abundance classifications: “rare”, “undetermined” and “abundant”. With the undetermined taxa being in an intermediate level of abundance between rare and abundant, *i.e.,* they are undetermined, because they are neither rare, nor abundant.

8.2.6. Rule-based machine learning with association rule mining

A microbial community dataset can include different kinds of information for each taxon, including abundance, taxonomy, and sample metadata. Each kind of information that is recorded for each taxon is a variable (or feature). For Association Rule Mining (ARM) (Agrawal, Imielinski and Swami, 1993), we assume that the variables registered in a dataset are items. A group of items is called an itemset, and a transaction stores itemsets. An association rule establishes and describes the association between itemsets in a transaction. Generally, in ARM, a database of transactions is first created, and then association rules are extracted (Kaur and Madan, 2015). The associations between itemsets describe relationships in the form of conditional statements. For example, if the genus *Clade Ia* (an item), from *Proteobacteria*, is present in a sample, and that sample also has very high nitrate concentration (another item) and silicates concentration (another item), then there is a 93% probability that the genus *Clade Ia* taxa are abundant (another item) in samples with those characteristics. Such probabilities are calculated based on how many transactions, within the database, included those itemsets ("genus *Clade Ia* + high nitrate concentration + high silicates concentration", and "abundant"). The ARM approach falls within the family of rule-based machine learning, data mining and knowledge based discovery in databases (Agrawal, Imielinski and Swami, 1993; Agrawal and Srikant, 1994).

Let *I* be a database with m itemsets, $I = I_1, I_2, \ldots, I_m$, let *T* be a transaction that includes itemsets from *I*, so that $T \subseteq I$, and let *D* be a database of transactions. Thus, for *n* transactions, a database of transactions can be described by Equation 8.1.

$$D: \{T_1, T_2, \ldots, T_n\}$$

Equation 8.1. Database of transactions.

An Association Rule (AR) is defined in the form of an implication, $X \Rightarrow Y$, where $X, Y \supset I$ and $X \cap Y = \emptyset$. The left side of the implication (*X*) is sometimes referred to as the Left Hand Side (LHS) or the antecedent of the rule, while the right side of the implication (*Y*) is sometimes referred to as the Right Hand Side (RHS) or the consequent of the rule. However, note that in the code implementation it is common to also find references to the LHS and RHS, which is why we are defining them here. We use the term "implication" for convention in describing the association rules, but this implication is not necessarily the same as the logical implications (more on this in the next section).

ARM is the process of mining association rules from a transactions database. ARM is divided into two main steps: 1) identification of large itemsets; and 2) construction of association rules. Most studies on ARM implementation focus on the optimization of the first step, which is not trivial and computationally demanding, while the second step is usually considered trivial. To do both steps, we used the apriori algorithm (Agrawal, Imielinski and Swami, 1993; Agrawal and Srikant, 1994), implemented by the R package *arules* (Hahsler *et al.*, 2011, 2021). The apriori algorithm produces rules with a single consequent, but with any number of antecedents.

For the first step of ARM, a large itemset is defined based on a pre-specified value of Support and the second step defines rules based on pre-specified requirements, most often Confidence. Let Count(X) be the number of occurrences of X. Then, the Support of an itemset, *X*, is defined as the proportion of occurrences of the itemset X in the database, with *n* transactions (Equation 8.2):

$$Support\ (X)\ =\ \frac{Count(X)}{n}$$

Equation 8.2. Support of an itemset.

Similarly, we can calculate the Support of the consequent and of the transaction (Equation 8.3):

$$Support(X\ \Rightarrow\ Y)\ =\ \frac{Count(X \cap Y)}{n}$$

Equation 8.3. Support of an association rule.

Note that the Support of an itemset is the probability of finding that itemset, since $P(X)\ =\ \frac{Count(X)}{n}\ =\ Support(X)$ (Equation 8.2). The Confidence is defined as the conditional probability of finding the consequent, given the antecedent (Equation 8.4):

$$Confidence\ (X\ =>\ Y)\ =\ P(Y|X)\ =\ \frac{P(X \cap Y)}{P(X)}\ =\ \frac{Support(X\ \Rightarrow Y)}{Support(X)}$$

Equation 8.4. Confidence of an association rule.

For the identification of large itemsets, we set a low Support value of 1%, because we do not have computing power limitations relative to this step and, with a low Support value, the chance of identifying interesting and rare rules increases. Since computing power was not a limitation, we also did not constrain the second step to a minimum Confidence value, even though we prioritized higher Confidence rules in the subsequent data analysis.

Complexity is defined as the size of the rule, which is the number of items in the antecedent and consequent. Thus, we set a minimum Complexity of 2, which translates to rules with, at least, one antecedent and one consequent.

Before applying the apriori algorithm for mining association rules, we made a subset of the original dataset with relevant variables for the construction of the transaction dataset. We removed data features that were inherently redundant and non-informative from an environmental perspective. For example, a rule with a sample ID will never be able to provide novel information, since the ID, by itself, is non-informative. Following the same reasoning, it was necessary to remove the Station ID, which simply indicates where the sample was collected. Thus, if such variables were kept, then they would result in rules that simply describe the sample or the station, which is repetitive information. Since the metadata table can be consulted to verify the content of a given sample, redundant metadata creates noise. Noise reduction is important to ensure that the ARM algorithm finds complex interactions of variables that do not repeat known metadata information.

Finally, we discretized the quantitative variables, because the apriori algorithm does not accept quantitative items. For the discretization, we used different approaches, depending on the kind of variable. Specifically, for the abundance scores of the taxonomic groups, they were replaced by the abundance classification (“rare”, “undetermined” and “abundant”). For the nutrient concentrations, we divided the values by quartiles and then labeled them: “very low”, “low”, “high” and “very high”. The reference values are present in Table 8.1. For the temperature, a different reference was used to account for negative values (Table 8.2).

Table 8.1. Reference values for the discretization of quantitative variables, relative to nutrient concentrations.

Variable	very low	low	high	very high
Chl *a* (mg/m^3)	]0,0.13]	]0.13,0.45]	]0.45,0.82]	]0.82,2.28]
NH_4^+ (µM)	]0,0.43]	]0.43,0.81]	]0.81,1.69]	]1.69,2.76]
NO_2^- (µM)	]0,0.06]	]0.06,0.1]	]0.1,0.18]	]0.18,0.56]
NO_3^- (µM)	]0,0.2]	]0.2,1.5]	]1.5,7.63]	]7.63,16.2]

PO_4^- (µM)	]0,0.16]	]0.16,0.25]	]0.25,0.7]	]0.7,1.04]
Salinity (PSU)	]0,33.8]	]33.8,34.81]	]34.81,34.95]	]34.95,35.12]
Si (µM)	]0,1.53]	]1.53,2.6]	]2.6,4.48]	]4.48,13.8]

Table 8.2. Reference values for the discretization of temperature (ºC) quantitative values.

Variable	very low	low	high	very high
Temperature (ºC)	]-0.81,2.32]	]2.32,4.66]	]4.66,6.32]	]6.31,8.54]

8.2.7. Statistical evaluation of association rules

The statistical strength of an association rule is usually measured by Support and Confidence (Equations 8.3 and 8.4), but those two metrics are usually not enough to identify the best rules or remove redundancy. Redundancy of rules refers to the fact that many rules with the same consequent might have antecedents that differ on only a few items. Thus, those redundant rules are different from each other, but provide little new information, resulting in noise.

For interpreting the meaning of a rule, the Confidence metric is the easier metric to understand, because it provides a conditional probability for the rule. However, it fails to provide a sense of the implication strength (Brin *et al.*, 1997). To solve this problem, the Conviction metric was introduced (Brin *et al.*, 1997), which emphasizes the direction of the implication of the rule, thus describing its implication strength. This is achieved by explicitly accounting for the probability of the antecedent appearing without the consequent (Equation 8.5):

$$Conviction(X \Rightarrow Y) = \frac{1 - Support(Y)}{1 - Confidence(X \Rightarrow Y)} = \frac{P(X)P(\neg Y)}{P(X \cap \neg Y))}$$

Equation 8.5. Conviction of an association rule.

It follows that Conviction will vary between 0 (no implication) and infinity (rule always holds, which is the case for Confidence = 100%), while Conviction = 1 indicates that the antecedent and the consequent are independent. However, Conviction is used to assert the magnitude of the implication, while there are other metrics to establish and evaluate the dependence between antecedent and consequent (more on that below).

As noted in the previous section, the term implication is used by convention, because there is no logical implication in an association rule. However, if Confidence = 100% (and, consequently, Conviction = infinity), then the association rule is equivalent to a

logical statement of implication. That is, in *D* it is always true that when X is present, then Y is also present.

Another aspect worth inspecting is the dependency between the antecedent and the consequent, because they might be associated because of random interactions in the database. One approach to identify which rules have antecedents statistically dependent on the consequents is the application of the Chi-squared test (Brin *et al.*, 1997). The Chi-squared test is used to test the independence of X and Y, which are two binary variables (2 by 2 contingency table). Thus, by establishing an alpha value (0.05 in here), it is possible to test the null hypothesis that X and Y are independent ($p > 0.05$); or reject it, if X and Y are dependent ($p < 0.05$). One problem with this approach is that each transaction will correspond to a statistical test. Thus, the chance of Type I error increases, which is why we applied the Bonferroni correction, before accepting/rejecting the null hypothesis.

The approach based on the Chi-squared test will divide rules into dependent or independent, but it will not provide information on the magnitude of the dependence between the antecedent and the consequent. To do so, the Lift metric was introduced (Brin *et al.*, 1997).

The Lift compares the frequency of X and Y occurring together relative to what would be expected if X and Y were independent (Equation 8.6):

$$Lift\ (X \Rightarrow Y)\ = Lift(Y \Rightarrow X)\ = \frac{Confidence(X\ =>\ Y)}{Support(Y)}\ = \frac{P(Y|X)}{P(Y)}$$

Equation 8.6. Lift of an association rule.

Thus, if Lift = 1, then X and Y are independent, the larger the value, the more dependent X and Y are. As noted in equation 8.6, Lift is symmetric, meaning that if the antecedent and consequent are switched, the Lift value will be the same.

We also used metrics that allow the comparison of redundant rules, by calculating how much information is gained by adding new antecedents to a rule. One such metric is Mutual Information (Tan, Kumar and Srivastava, 2004), which calculates the information that the antecedent provides about the consequent. Thus, if there are two similar rules, the rule that provides more information on the same consequent will be of more interest. Mutual Information varies between 0 and 1, the larger the value, the more information the antecedent gives regarding the consequent. Another alternative is the Improvement metric (Bayardo, Agrawal and Gunopulos, 1999), which calculates how much

Confidence is gained by the addition of a new item in a rule. Thus, it allows a direct comparison between similar rules. Improvement is calculated as the difference between the Confidence of a given rule and the minimum Confidence of any subrule. A subrule is a rule made from items present in another rule. For example, $A \Rightarrow Y$ is a subrule of the rule $A + B \Rightarrow Y$.

In summary, we selected metrics that can be used to assert the following characteristics: 1) the frequency of the association rule (Count, Support, Confidence); 2) strength of implication of the rule (Confidence, Conviction); 3) dependency between antecedent and consequent (Lift, Chi-squared test); and 4) redundancy (Mutual Information, Improvement, Complexity).

8.2.8. Statistical analysis and data visualization

All statistical analysis and data visualization was performed in R (R Core Team, 2023). Data wrangling was done with the help of *dplyr* (Wickham *et al.*, 2023), *stringr* (Wickham, 2023) and *tidyr* (Wickham, Vaughan and Girlich, 2023) R packages. Alpha and beta diversity calculations were made using the *vegan* R package (Oksanen *et al.*, 2025). Most plots were made using *ggplot2* R package (Wickham, 2016), except for: 1) the map in Figure 8.1, which was made with *ggspatial* (Dunnington, 2023) and *ggOceanMaps* (Vihtakari, 2024) R packages; 2) the ordination plots, which were plotted with base R functions and *vegan* R package (Oksanen *et al.*, 2025); and 3) network analysis software, Gephy (Bastian, Heymann and Jacomy, 2009), was used to visualize the connection between large numbers of association rules.

For statistical modeling we used the linear least squares and Poisson models, and for the correlations we used Pearson (linear) and Kendall (ranked) (Kendall, 1938), all were done in R using the *stats* R package (R Core Team, 2023). Briefly, we used the linear least squares modeling for a baseline description, but due to the difficulty with fitting the data, additional tests were added for additional support. The Poisson model was added, because it has the potential to fit better in a dataset of counts (consider number of taxa as counts), and then the Pearson correlation was used as confirmatory statistics of the linear least squares model, and the Kendall correlation was used to verify if there was any non-monotonic pattern missing from the Pearson correlation. Therefore, we used four statistical methods in tandem to compensate for the difficulty in fitting the data. However, note that the models were made to aid description of major patterns, and not with the objective of finding the best possible model to fit and explain the data.

To compare multiple means, within a grouping variable, the two-way ANOVA test was used, using the *rstatix* R package (Kassambara, 2022). For this test, the assumption of normality was tested with the Shapiro test and the assumption of homogeneity of variance was assessed with the Levene test, using the *rstatix* R package (Kassambara, 2019). For *post-hoc* tests of two-way ANOVA, pairwise t-test was used (Kassambara, 2019).

Statistics associated with association rule mining (described in previous section), were calculated using the *arules* R package (Hahsler *et al.*, 2011, 2021).

All statistical tests considered an alpha of 0.05, and whenever we identified a risk of Type I errors, the Bonferroni correction was applied.

8.3. Results

8.3.1. Overview of prokaryotic communities from 2016 to 2020, across pelagic layers

Taxonomic richness (number of taxa) was the main alpha diversity measure employed in this study, which resulted from the assignment of ASVs into taxonomic groups (down to genus level, see section 8.2.4). An initial overview of the dataset shows that taxonomic richness increased with depth, with the bathypelagic layer presenting the highest numbers of taxa, and variability among samples was higher for the epipelagic and mesopelagic layers (Figure 8.2).

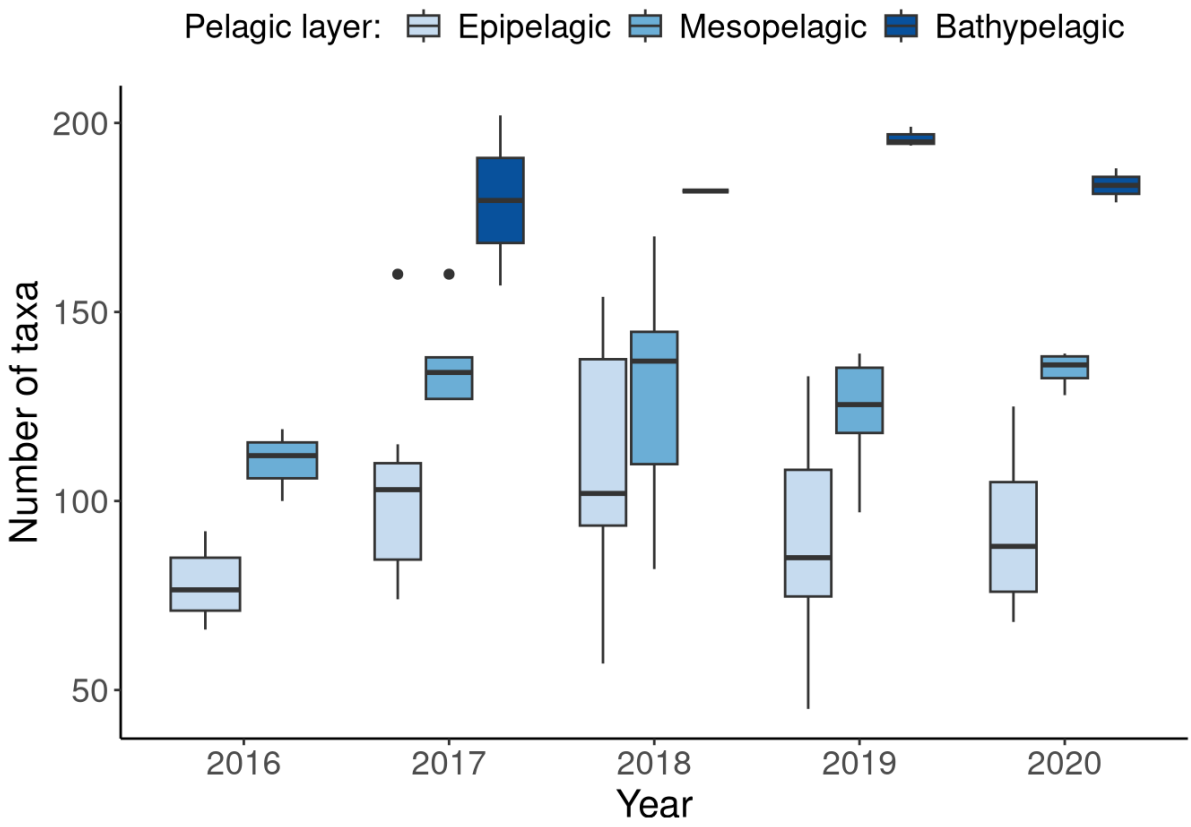


Figure 8.2. Boxplots of the number of taxa present per sample (alpha diversity estimate) as a function of sampling year. The number of taxa corresponds to the different taxonomic groups present in a given sample, grouped down to genus level. Samples are grouped by pelagic layers in boxplots (blue shading), with darker blues attributed to deeper samples. Boxplots indicate median, first and third quartile, minimum and maximum values over all samples (n = 117) and outliers (black dots, outside interquartile range).

Within the same pelagic layer, the number of taxa remained constant between the years from 2016 to 2020 (Figure 8.2), without significant differences in the mean number of taxa when accounting for the interaction between year and pelagic layer ($p > 0.05$, two-

way ANOVA test, Supplementary Table 11.25). *Post-hoc* tests revealed that there were significant differences between the number of taxa in 2016 compared to 2017 and 2018 ($p < 0.05$, pairwise t-test, Supplementary Table 11.26), but the significance was low and disappeared after applying the Bonferroni correction for Type I errors ($p > 0.05$, pairwise t-test, Supplementary Table 11.26). On the contrary, the *post-hoc* test revealed significant differences between all pelagic layers, even after the Bonferroni correction ($p < 0.05$, pairwise t-test, Supplementary Table 11.26). The normality assumption was verified ($p > 0.05$, Shapiro test, Supplementary Table 11.27). Note that for the two-way ANOVA test, the interaction between years and pelagic layers did not consider the bathypelagic layer, because of the low number of samples (n < 4 in each year). However, as shown by the pairwise t-test, this pelagic layer has a significantly higher number of taxa when the samples from all years are grouped together. When considering the year variable by itself, both linear least squares and Poisson models were not significant ($p > 0.05$, Supplementary Table 11.28 and 11.29), and this was further supported by both Pearson and Kendell correlations, which were close to zero and not significant ($p > 0.05$, Supplementary Table 11.30). Therefore, multiple statistical tests independently confirm the absence of any significant change in the number of taxa between 2016 and 2020, within each pelagic layer, except for the differences found for the year 2016. Instead, major differences in the number of taxa were attributed to the pelagic layers.

In terms of taxonomic composition, the most abundant phyla were *Bacteroidota* (41.8 %), *Proteobacteria* (41.4%), *Verrucomicrobiota* (4.9%), *Crenarchaeota* (2.7%) and *Thermoplasmatota* (1.5%) (calculated from the whole dataset, Supplementary Table 11.31). See Methods section for notes on recent nomenclature updates on *Proteobacteria (Pseudomonadota)*, *Crenarchaeota (Thermoproteota)* and *Thermoplasmatota (Methanobacteriota*) (Oren and Garrity, 2021; Göker and Oren, 2023). *Bacteroidota* relative abundance was generally constant through the years, except for a slight increase between 2019 and 2020 in the epipelagic layer, and a slight decrease in the bathypelagic layer (Figure 8.3). *Proteobacteria* decreased a little from 2016 to 2020 in the epipelagic and mesopelagic layer, but this decrease was within the standard deviation (Figure 8.3). *Verrucomicrobia* also decreased from 2016 to 2020 in all pelagic layers, but this decrease was more pronounced in the epipelagic layer (Figure 8.3). *Crenarchaeota* and *Thermoplasmatota*, both Archaea, increased their abundance from the epipelagic to the bathypelagic layer, but were constant through time in each pelagic layer, except for *Crenarchaeota* in the epipelagic layer (Figure 8.3). Interestingly, *Crenarchaeota* and *Thermoplasmatota* had generally high error bars, which increased in the bathypelagic

layer. Finally, the remaining phyla (grouped in “Other”) decreased slightly from 2016 to 2020 in the epipelagic and mesopelagic layers, but increased in the bathypelagic layer, even though by very little (Figure 8.3). Generally, archaeal phyla were more abundant in the mesopelagic and bathypelagic layers, while *Proteobacteria* was highest in the epipelagic layer (Figure 8.3).

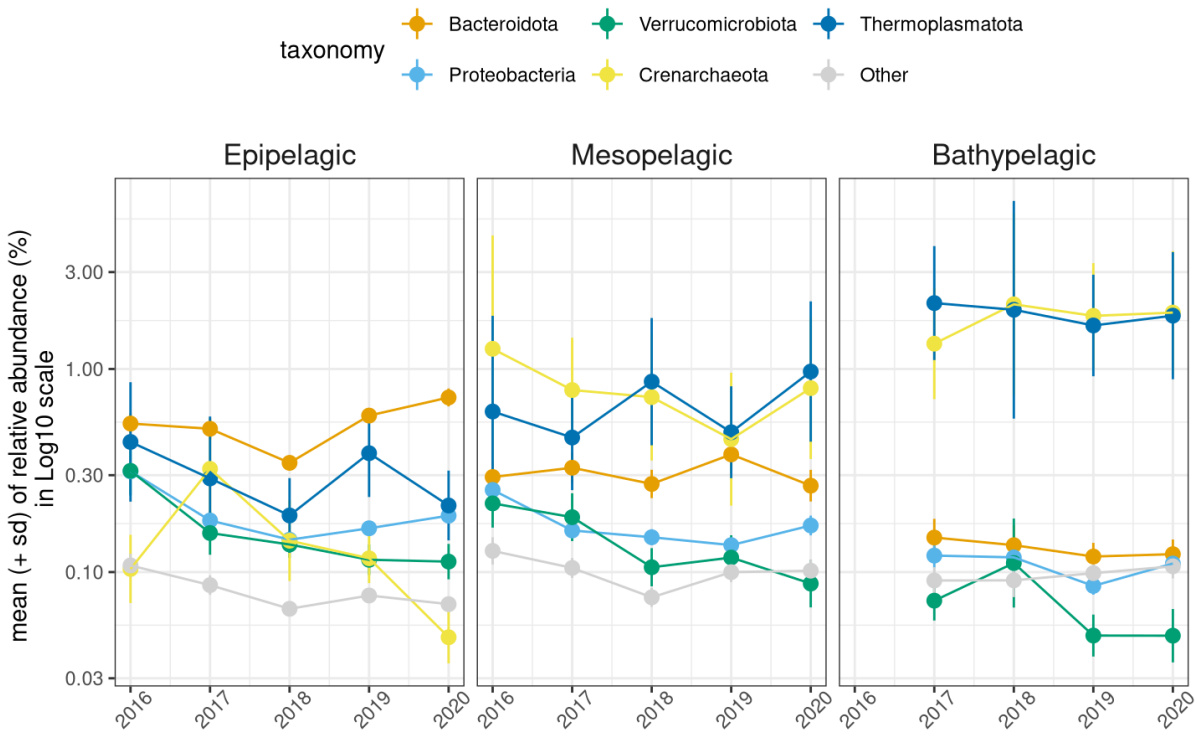


Figure 8.3. Mean ± standard deviation (sd) of relative abundance (%) in Log10 scale of phyla across year and pelagic layers. The five most abundant phyla across the entire dataset were illustrated with qualitative coloring and the remaining phyla were grouped into the “Other” category, in gray.

The main class level taxonomic groups were *Bacteroidia* (41.8%), *Alpha-* (23.4%) and *Gammaproteobacteria* (18.0%), followed by *Verrucomicrobiae* (4.8%), and *Nitrososphaeria* (2.7%) (calculated from whole dataset, Supplementary Table 11.31). An analysis of the major classes across multiple years and pelagic layers revealed similar patterns to the underlying phylum level taxonomic groups (Supplementary Figure 10.40). However, the only archaeal group to clearly surpass the abundance of all other groups in the mesopelagic and bathypelagic layer was only the *Nitrososphaeria* class, despite very high variance (Supplementary Figure 10.40). The taxa attributed to “Unknown” category was incorporated into the plot for completion’s sake, but it provides little biological information, because it groups unrelated taxa (Supplementary Figure 10.40). *Alpha-* and *Betaproteobacteria* shared the same patterns, and finally the

Verrucomicrobiae decreased abundance through the years in the Epipelagic and mesopelagic layer, but not so in the bathypelagic layer, where 2018 was abnormally high (Supplementary Figure 10.40).

Beta diversity indicated that the community composition was different between pelagic layers, with the most distinct prokaryotic communities being present in the bathypelagic layers (Figure 8.4). The community composition of the mesopelagic layer was within the observed range of variation of the epipelagic layer (Figure 8.4). Thus, the epi- and mesopelagic layers presented small differences with each other but were clearly separated from the bathypelagic layer. The PERMANOVA tests showed that the community composition was significantly different between pelagic layers and years ($p < 0.05$, Supplementary Table 11.32). However, note that the years variable lost its significance once the pelagic layers were accounted for ($p > 0.05$, Supplementary Table 11.32).

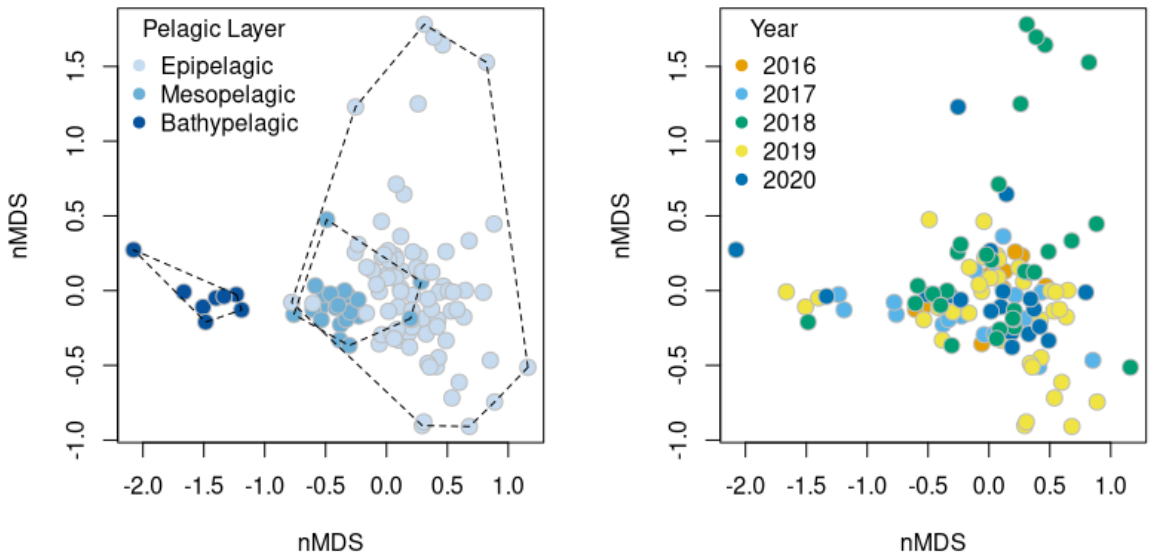


Figure 8.4. Beta diversity analysis of taxonomic groups, with nonmetric multidimensional scaling (nMDS) represented in an ordination plot. The left side illustrates samples according to their pelagic layer (shades of blue). The right side represents the sampling year, represented by qualitative coloring. Taxonomic groups represents ASVs grouped together down to genus-level, instead of the ASV-level, see Methods section.

To account for the impact of abundance in diversity estimates, we clustered taxa based on their abundance, using unsupervised machine learning, *ulrb* (Pascoal *et al.*, 2025a; 2025b) (Chapters 6 and 7) and classified them into rare, undetermined and abundant. From 2016 to 2020, across all pelagic layers, the rare biosphere presented the highest number of taxa, followed by the undetermined and abundant taxa (Figure 8.5).

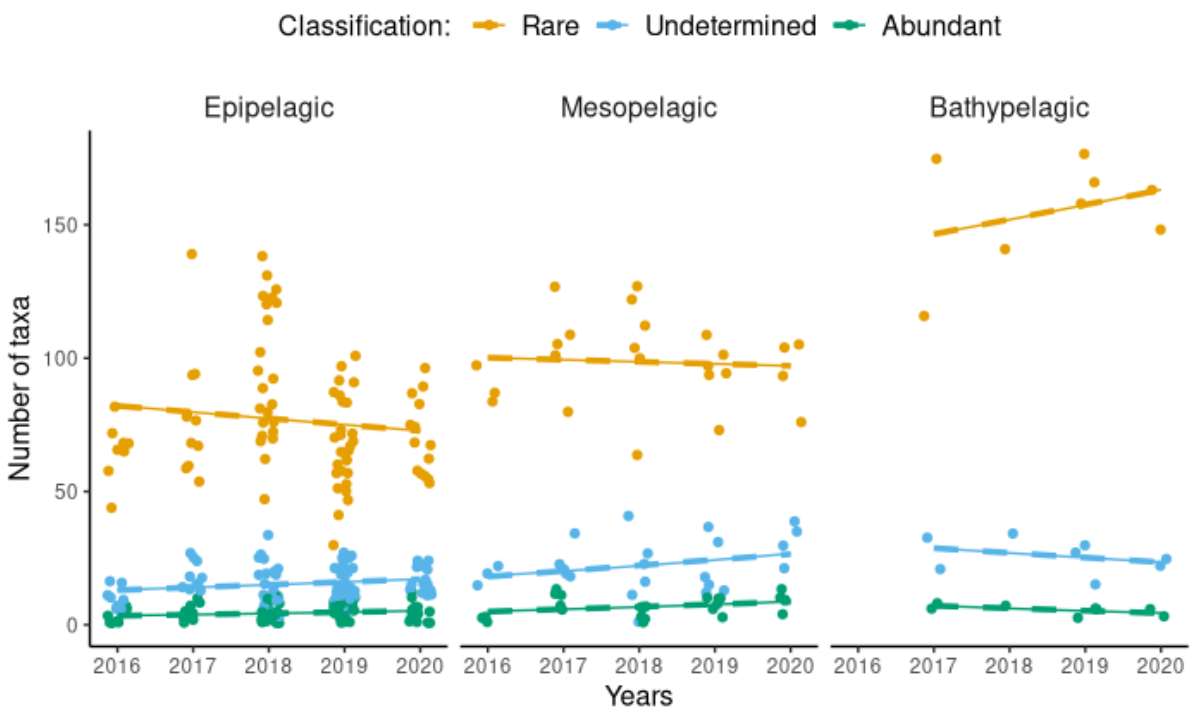


Figure 8.5. Number of rare, undetermined and abundant taxa, across years and pelagic layers. The classifications are divided by qualitative colors. Dashed regression lines indicate the linear least squares model, and the continuous lines indicate the Poisson model output.

When considering the abundance classifications (rare, undetermined and abundant) and the pelagic layers (Figure 8.5), the following patterns were identified: a small decrease in the number of rare taxa in the epi- and mesopelagic layers from 2016 to 2020; concomitant with a small increase in the number of abundant and undetermined taxa, for the same pelagic layers and time period (Figure 8.5). At the bathypelagic layer, however, the pattern was opposite, *i.e.*, a small increase in the number of rare taxa and a small decrease in the number of abundant and undetermined taxa (Figure 8.5). Notably, the linear least squares models generally failed to fit the data (Supplementary Table 11.28) and were never significant, but the Poisson one obtained significant models for the abundant and undetermined taxa in the epi- and mesopelagic layers, and for the rare biosphere in the epipelagic layer only (Supplementary Table 11.29). Both Pearson and Kendall correlations were not significant, and close to zero, except for the correlation between abundant taxa richness and years in the bathypelagic layer, which was below -0.5, but still not significant (Supplementary Table 11.30). Generally, Kendall correlations were weaker than Pearson correlations (Supplementary Table 11.30).

8.3.2. Diversity as a function of distinct environmental variables

Of all the environmental variables tested (temperature [ºC], silicates [µM], NO_2^- [µM], NO_3^- [µM], NH_4^+ [µM], and Chl *a* [mg/m^3]), temperature had the most important effect (Figure 8.6). According to the linear least squares model, by every 1ºC increase in temperature, approximately 10 taxa were lost ($p < 0.05$, Supplementary Table 11.33). However, when the abundance classification was accounted for, it became clear that most diversity loss with increasing temperatures was specific to rare taxa (Figure 8.6). Accordingly, by every 1ºC increase in temperature, approximately 9 rare taxa were lost, and 1 undetermined taxon was lost ($p < 0.05$, Supplementary Table 11.30), while there was no significant lost or gain for the abundant taxa ($p > 0.05$, Supplementary Table 11.33). The Poisson model was consistent with the linear least squares model (Figure 8.6, Supplementary Table 11.34). Both Pearson and Kendall correlations indicated negative correlations between the number of taxa and temperature for all abundance classifications, which were significant for all taxa ($p < 0.05$), except the abundant ($p > 0.05$, Supplementary Table 11.35).

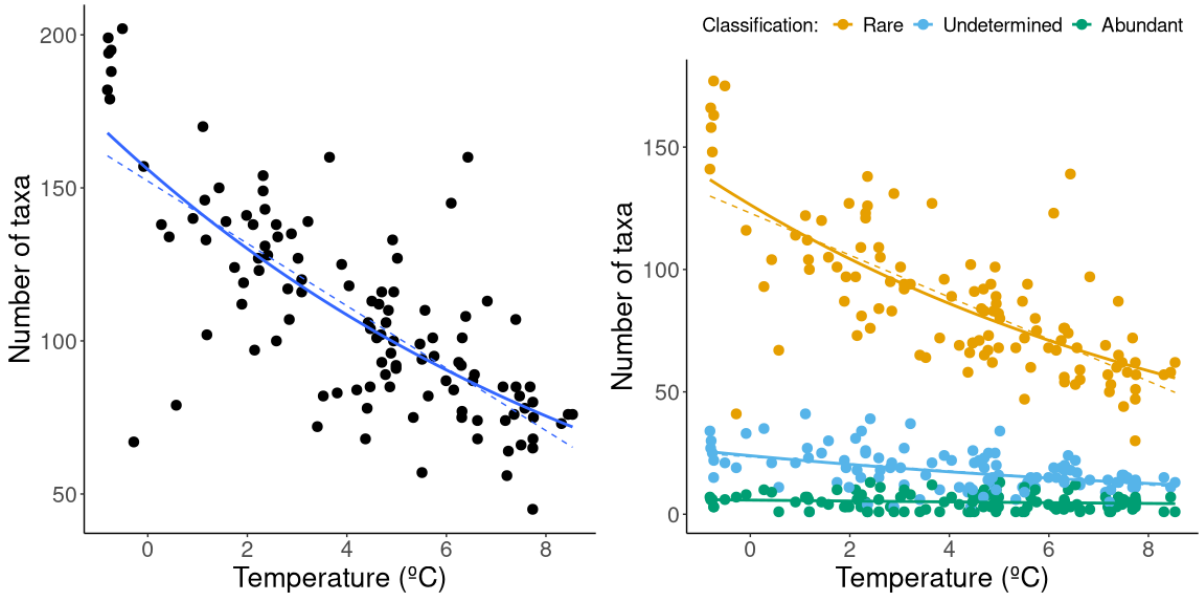


Figure 8.6. Number of taxa present in each sample (alpha diversity estimate) as a function of temperature (ºC). The left panel shows alpha diversity for the overall dataset. The right panel shows the number of taxa per sample, in each classification, with qualitative coloring. All samples with temperature (ºC) available were used (n = 114). Dashed regression lines represent the linear least squares model, and the continuous lines represent the Poisson model.

To further understand the impact of temperature on the number of taxa, we divided the previous analysis by the pelagic layers. This analysis showed that the highest number of taxa were found at deeper layers, which are also the coldest ones (Supplementary Figure 10.41, left side), followed by the meso- and epipelagic layers in terms of both

temperature and number of taxa. It was also verified that within the epi- and mesopelagic layers there were no trends in the number of taxa as a function of depth, which indicates that the temperature, indeed, reduced the number of taxa within those pelagic layers (Supplementary Figure 10.41). Additionally, the statistical models and correlations on the number of taxa as a function of depth, within pelagic layers, were generally non-significant and indicated no pattern (Supplementary Table 11.33-11.35).

Regarding the number of taxa as a function of temperature, the linear least squares model showed that approximately 7 taxa were lost per 1 ºC increase ($p < 0.05$, Supplementary Table 11.33), and a much higher rate of decline in the bathypelagic layer, but not significant ($p > 0.05$, Supplementary Table 11.33). The Poisson model, the Pearson and Kendall correlations between the number of taxa and temperature, within each pelagic layer, were consistent with the linear least squares model (Supplementary Tables 11.33-11.35). An analysis of temperature as a function of year and pelagic layers, grouped by water mass, showed that temperature increased the most within the AW of the epipelagic layer, 2016 to 2020 (Supplementary Figure 10.42). While SW, TAW, and WCW remained constant in terms of temperature (Supplementary Figure 10.42). Then, water masses were divided by both the number of taxa and the temperature, with WCW and TAW presenting lower temperatures and higher number of taxa, while SW and AW presented similar ranges of temperature and number of taxa (Supplementary Figure 10.43). The linear least squares model identified decreasing diversity with increasing temperatures within all water masses, except for the WCW, where an increase was identified, but was not significant ($p > 0.05$, Supplementary Table 11.33). The Poisson model detected a decrease in diversity as a function of temperature for all water masses, including the WCW, even though the Poisson model also failed to be significant for the WCW ($p > 0.05$, Supplementary Table 11.34). The Pearson and Kendall correlations obtained results that were consistent with the linear least squares model, but were not significant for the TAW and WCW water masses ($p > 0.05$, Supplementary Table 11.35). The difficulty in obtaining models with a significant fit to the WCW data is probably due to the low number of samples of that subset of the dataset. Geographically, the relationship between the number of taxa and temperature was equivalent in the Kongsfjorden and Fram Strait regions (Supplementary Figure 10.43), with all models being significant and indicating a loss of taxa as a function of increasing temperature within similar values ($p < 0.05$, Supplementary Tables 11.33-11.35).

The number of taxa increased with silicate concentration (Figure 8.7), with the highest increase for the rare biosphere (Figure 8.7), while the undetermined and abundant taxa

presented a very small increase (Figure 8.7). According to the linear least squares model, the number of overall taxa increased by 9 for each unit increase of silicate concentration ($p < 0.05$, Supplementary Table 11.36), while it increased 7 taxa per unit of silicate concentration for the rare biosphere ($p < 0.05$, Supplementary Table 11.36), 1 for undetermined taxa ($p < 0.05$, Supplementary Table 11.36) and none for the abundant taxa ($p < 0.05$, Supplementary Table 11.36). The Pearson model was consistent with the linear least squares (Supplementary Table 11.37). The Pearson and Kendall correlations showed a positive correlation between the number of taxa (overall) and silicate concentration ($p < 0.05$), but significantly negative correlations for the rare and undetermined taxa ($p < 0.05$, Supplementary Table 11.38).

The effect of silicate concentration on the number of taxa was probably a consequence of depth, as silicates concentration increased with depth (Supplementary Figure 10.45). All models and correlations did not identify any relationship between abundant taxa and silicates concentration (Figure 8.7 and Supplementary Tables 11.36-11.38).

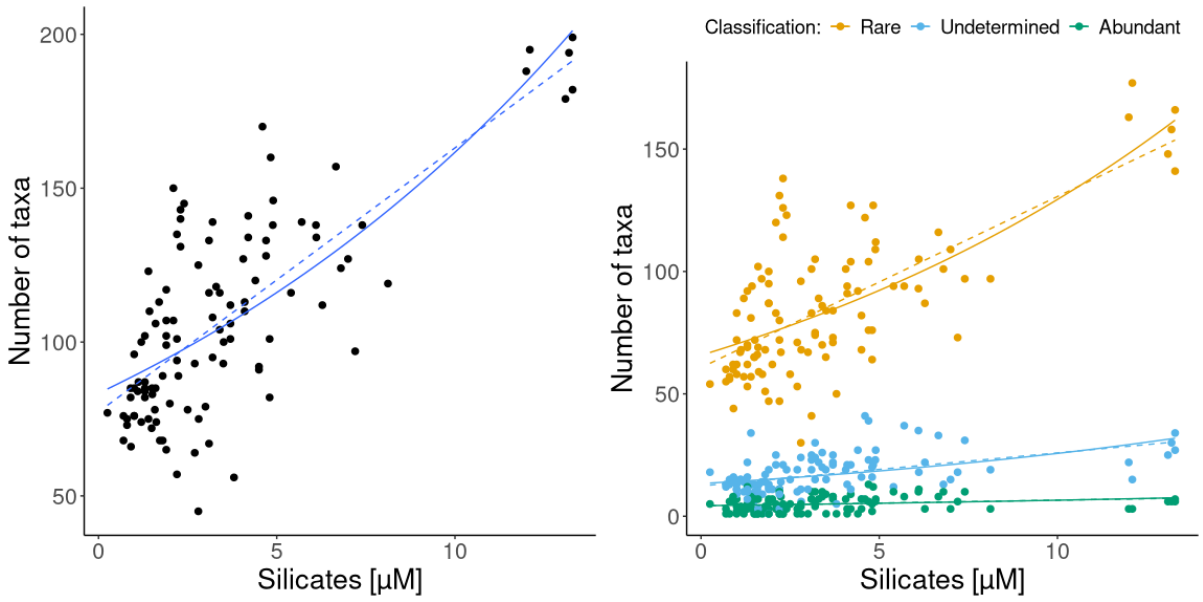


Figure 8.7. Number of taxa present in each sample (alpha diversity estimate) as a function of silicate concentration (µM). The left panel shows alpha diversity for the overall dataset. The right panel shows the number of taxa per sample, in each classification, with qualitative coloring. Dashed regression lines indicate the linear least squares model output, and the continuous lines indicate the Poisson model output. All samples with silicate concentration (µM) available were used (n = 108).

The number of taxa was not changed by the other environmental variables: Salinity (PSU), NO_2^- (µM), NO_3^- (µM), NH_4^+ (µM) and Chl *a* (mg/m³) (Supplementary Figures 10.46-10.50). Even though there were no clear patterns for all these variables, some of

the samples with more than 15 µM of NO_3^- and less than 0.5 µM NH_4^+ presented very high values of rare taxa and correspond to bathypelagic samples (Supplementary Figures 10.49 and 10.50).

8.3.3. Uncovering complex combinations between prokaryotes and multiple environmental variables

We applied association rule mining to obtain association rules and then evaluated them using metrics that assess the following four aspects: (i) frequency (Support, Confidence, and Count); (ii) dependency between antecedent and consequent (Chi-squared test, and Lift); (iii) strength of implication (Conviction, and Confidence); and (iv) redundancy (Mutual Information, Improvement, and Complexity), see Methods section. After applying ARM to the MOSJ 2016-2020 dataset (n = 116), we obtained a total of 18 550 association rules. Most of the rules had around 1% Support (mean = 1.1%, min = 0.1%, max = 21.2%, n = 18 550 rules), between 82% and 93% Confidence (mean = 88.3%, min = 80%, max = 100%, n = 18 620 rules), and the Lift was generally below 5 (mean = 1.47, min = 1.01, max = 20.28, n = 18 550 rules). Generally, the rules with the highest Lift had the lowest Support and high Confidence (Supplementary Figure 10.51).

Looking at the top 100 rules (ordered by descending Lift, Supplementary Table 11.39), several were redundant with each other, meaning that they provided similar information. For example:

{taxon = genus *Clade Ia* (*Proteobacteria*, *SAR11 clade*), NO_2^- = very high, Si = very high} => {Classification = Abundant} (Confidence = 93%)

Association rule 1 (AR1)

{taxon = genus *Clade Ia* (*Proteobacteria*, *SAR11 clade*), Pelagic Layer = Mesopelagic, NO_2^- = very high, Si = very high} => {Classification = Abundant} (Confidence = 93%)

Association rule 2 (AR2)

The examples of AR1 and AR2 show that different rules might have very similar information, making them redundant with each other. In Supplementary Table 11.39 there are more examples of variations of AR1 and AR2. In this example, the Lift was higher for AR1 than AR2, but the difference was negligible ($AR1_{Lift}$ = 20.28, $AR2_{Lift}$ = 20.18). There were 8 additional rules that were very similar to AR1 and AR2, with similar Lift values, in the top 100 (Supplementary Table 11.39). To obtain more information, it is

not possible to restrict to the top 100 rules, which represent < 1% of all rules obtained (18 550 rules), but it is impossible to manually inspect so many rules.

We obtained 593 dependent rules ($p < 0.05$, Chi-square test and Bonferroni correction for Type I error), which were grouped by their Confidence and Conviction (Figure 8.8). For similar values of Confidence, the rules with highest Conviction were explored more thoroughly, because they have stronger statistical implication between antecedent and consequent, which was further supported by the Lift values (Figure 8.8). By extracting dependent rules with Lift > 15, ordered by descending Conviction, all rules were relative to abundant and undetermined taxa (Supplementary Table 11.40), corresponding to the two clusters with highest Lift and Conviction scores (Figure 8.8).

We repeated the analysis of Confidence as a function of Conviction for all rules (dependent and independent) and verified that the three clusters were still present (Supplementary Figure 10.52), meaning that the relationship between Conviction and Confidence metrics was not specific to statistically dependent rules. The other metrics tested, Mutual Information and Improvement, were also divided based on the clusters of Conviction and Confidence values (Supplementary Figure 10.53), providing little new information on the association rules obtained. However, by comparing Mutual Information and Improvement, the Improvement metric provided more information within each of the three major groups of values, because within those groups the size of the points varied more with Improvement than with Mutual Information (Supplementary Figure 10.53).

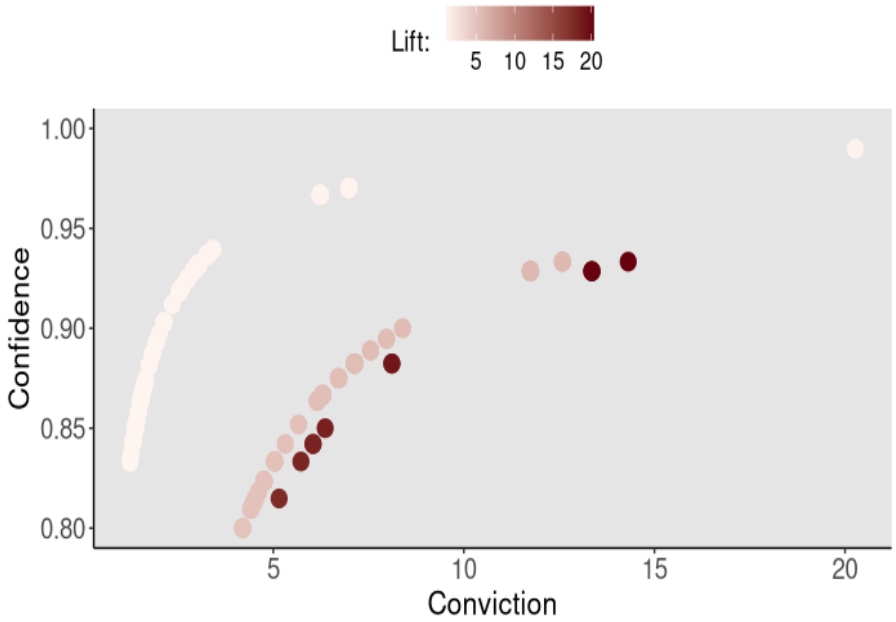


Figure 8.8. Confidence as a function of Conviction of rules with antecedents statistically dependent on their consequences. Rules were colored by Lift (red shading). Rules with Conviction equal to infinity were removed. n = 593 rules. Gray background was used to allow visibility of low Lift values.

After selecting statistically dependent rules ($p < 0.05$, Chi-squared test) with high Lift (Lift > 15), the number of rules was reduced to 11, but 10 of those rules were redundant (Supplementary Table 11.40), thus providing very little useful information. Without the Lift filtering, we obtain 595 rules, but this number of rules is too much information to be human-interpretable and keeps redundant information (Supplementary Data 8.2). To solve the redundancy problem and select the most interesting rules, we developed a method to automatically remove redundant rules based on the optimization of some target metrics. We considered several possibilities for target metrics, like the Mutual Information and Improvement indices, which compare redundant rules. Based on Supplementary Figure 10.53, the Improvement metric was better at distinguishing between rules within the same major groups of values (divided by Lift), so we selected Improvement for redundancy removal. By using the Improvement metric on the subset of dependent rules with the highest Lift value, we obtained two non-redundant and informative rules:

{taxon=family *Flavobacteriaceae* (*Bacteroidota*), Pelagic layer = Epipelagic, NH_4^+ = Unknown, Si = low} => {Classification = Abundant} (Confidence = 83%)

Association rule 3 (AR3)

{taxon=genus *Clade Ia* (*Proteobacteria*, *SAR11 clade*), Pelagic layer = Mesopelagic, Si = very high} => {Classification = Abundant} (Confidence = 88%)

Association rule 4 (AR4)

In fact, rules AR3 and AR4 provide different information from each other, and both are detailed and relevant. Specifically, AR3 indicates that if a given taxon belongs to the *Flacobacteriaceae* family, is present in the epipelagic layer and is under low silicate concentration, then there is an 83% chance it will also be abundant. Note that the samples responsible for AR3 probably had missing information regarding the ammonia concentration. However, due to the ARM algorithm and due to the structure of the rules, it can handle missing values without the need of removing valid samples. The interpretation of AR4 can be done similarly: if a taxon belongs to the *Clade Ia* genus of the *Proteobacteria* phylum, is present at the mesopelagic layer and is under very high silicate concentration, then there is an 88% probability that it is abundant (AR4) (more on biological interpretation of rules in the next sections).

Then, we applied the same redundancy removal method for all dependent rules, independently of the Lift value, and obtained 31 rules (Supplementary Table 11.41). This subset of non-redundant rules still included some level of redundancy in rules that did not have taxonomic information, *i.e.*, rules that only presented environmental information. Some insights of this subset of rules include: 1) low nutrient values were associated with the rare biosphere; 2) most of the relevant rules with taxonomic information were relative to abundant and undetermined taxa; 3) high granularity in taxonomic information allowed the identification of distinct rules for three subgroups of the *SAR11* order, whereby *Clade Ia* was abundant in the mesopelagic Layer, in samples with very high silicates concentrations, while the *Clades III* and *IV* were rare and were not associated to specific environmental conditions. Regarding this last example, there was little detail on *Clades III* and *IV*, which was probably a consequence of the Improvement metric not necessarily favoring complex rules. It may also reflect the fact that they were rare, so there was less environmental information to satisfy a complex rule.

By filtering out independent rules, we might be losing relevant information, so we tested the same redundancy removal methodology to the overall dataset (dependent + independent rules), to verify if new relevant information could be obtained. For all rules, the three main clusters of Confidence as a function of Conviction, colored by Lift, were kept (Supplementary Figure 10.52), hinting at a very structural pattern in the data.

To show the most relevant rules (dependent and independent), we selected rules with high Conviction (> 5), Confidence (> 90%) and Lift (> 1), then we used Improvement to select the best 100 non-redundant rules (Supplementary Table 11.42). In this subset, all 100 rules were relative to the rare biosphere, indicating that most statistically independent rules belong to the rare biosphere. Several rules provided very little information, because the Improvement metric favored shorter rules.

Another approach to select non-redundant rules could be to select for Complexity, which is measured by the size of the antecedent of the rule. We repeated the previous steps (Conviction > 5, Confidence > 90%, Lift > 1), but used Complexity instead of Improvement for non-redundant rules selection (Supplementary Table 11.43). Generally, the non-redundant rules selected based on Complexity (Supplementary Table 11.43) combined taxonomic information with environmental information to a larger extent than the rules selected by Improvement (Supplementary Table 11.42). The more complex rules included some rules with undetermined taxa, but most were still relative to the rare biosphere (Supplementary Table 11.43). Even though more complex rules are more informative, their contribution to the statistical strength of the rule might be negligible, which is why Improvement and Complexity selected different rules. However, both can be used to extract subsets of interesting rules.

Finally, one of the challenges in using ARM results is that there are a very large number of rules, which are hard to read in tabular form. To view many rules, focusing on the statistically dependent rules ($p < 0.05$, Chi-squared test and Lift > 1, n = 593), we used network visualization software (Gephy) to identify major patterns of the rules obtained and their interactions (Figure 8.9). Briefly, nodes represent rules (colored by Lift), which are connected by edges (colored by Lift) to multiple antecedents and one consequent (shown as labels). The same label can include any number of edges, connecting to rules that used that specific antecedent/consequent. Therefore, the rule A + B => C would be represented as a single node (one point), connected by edges to A, B, and C, which are shown as text labels. Then, A, B, and C can still be connected to other rules (other nodes), using additional edges. This plotting approach illustrates the different variables impacting rare, undetermined and abundant taxa (Figure 8.9). In fact, the Lift is consistent with the abundance classification: rare taxa have the lowest Lift, followed by the undetermined and abundant taxa, which constitute separate clusters of rules. This relationship between the abundance classification (rare, undetermined or abundant) and Lift was statistically confirmed (Supplementary Figure 10.54). This relationship indicates that rare taxa presented weaker statistical dependence with environmental variables.

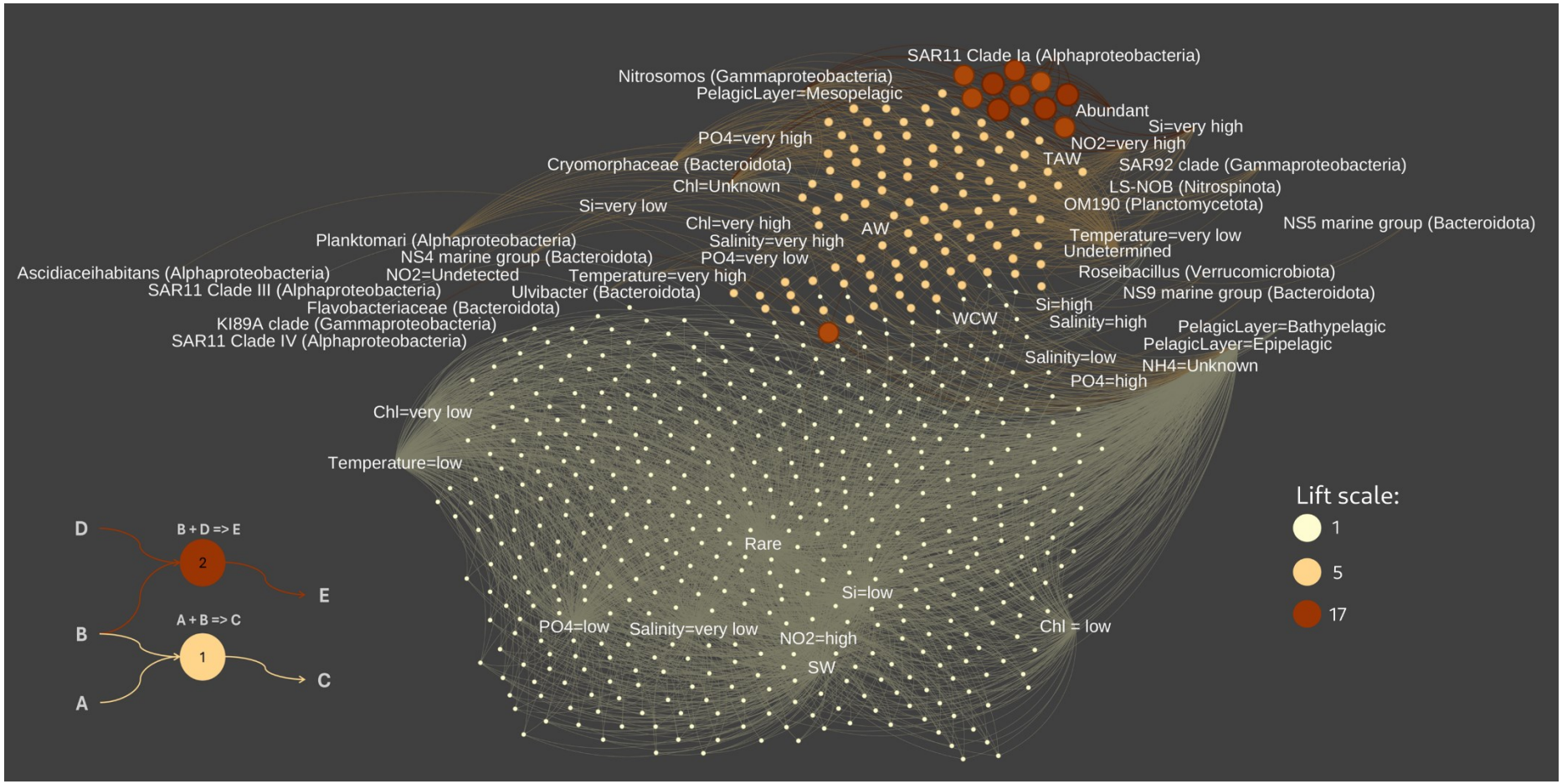


Figure 8.9. Visual representation of association rules (n = 593 rules). Each point represents an association rule, colored by Lift value (red for Lift > 17; dark yellow for Lift > 5; and light yellow for Lift > 1). Nodes represent rules, connecting their respective antecedents and consequents by edges. Antecedents and consequents are shown as text. The size of the rules (nodes) is proportional to the Lift of the rules. For the reference values of environmental variables, see Tables 8.1 and 8.2.

Note, however, that this analysis was made on a subset of statistically dependent rules, meaning that the difference in Lift across the rare, undetermined and abundant hints at the magnitude of the dependence.

The clusters of dependent rules obtained show that there are fundamentally different sets of variables implicated in the rare, undetermined and abundant taxa (Figure 8.9 and Supplementary Figure 10.54). Recall that the quantitative values for environmental variables were discretized for the ARM algorithm (see Methods, section 8.2) and the reference values were described in Tables 8.1 and 8.2.

For the rare biosphere, there was one large cluster characterized by scarcity of several nutrients (low PO_4^- concentration, low silicates concentration, very low salinity, very low to low Chl *a* concentration), by low temperature (ºC) and by the SW water mass (Figure 9). There were also a few rules with associations to the epipelagic layer. At the frontier between rules with rare or undetermined consequents, some antecedents included the WCW water mass, the bathypelagic layer and high levels of silicates concentration (Figure 8.9). The undetermined cluster was of an intermediate size, positioned within the rare and abundant taxa. In contrast to the rare biosphere cluster, the undetermined cluster highlighted many taxonomic groups, including *Bacteroidota* (*e.g.*, *Cryomorphaceae*), *Planctomycetota* (*e.g.*, *OM190*), *Nitrospinota* (*e.g.*, *LS-NOB*), *Gammaproteobacteria* (*e.g.*, *SAR92*), *Alphaproteobacteria* (*e.g.*, *SAR11 Clades III* and *IV*), and *Verrucomicrobia* (*e.g.*, *Roseibacillus*). Within the cluster of rules with an undetermined consequent, two main groups of environmental variables stood out. One of those was marked by the AW water mass, high levels of Chl *a*, salinity, and phosphate concentrations, as well as low silicates concentrations and the mesopelagic layer; on the other side, the TAW water mass stood out, with variables such as low temperature, and high concentrations of silicates and salinity. For the abundant taxa, there were very few rules, if compared with the rules with undetermined or rare taxa (Figure 8.9), which reflects the fact that there are few abundant taxa. However, they were the rules with the strongest statistical support in terms of the metrics previously described (Conviction, Confidence, Lift and Improvement) (Figure 8.8 and Supplementary Figure 10.53). This group of rules was constituted mainly of antecedent with taxonomic groups (genus *Clade Ia* from *SAR11*, class *OM190*, genus *LS-NOB*, genus *Nitrosomonas*), some included the mesopelagic layer and very high concentrations of NO_2^- and silicates (Figure 8.9). Note that the genus *Clade Ia*, from *SAR11*, was exclusive to abundant taxa rules, indicating that it was one of the most abundant taxonomic groups in our dataset (but it was also found as rare and undetermined, see next sections). It is also worth noting that the genus

Nitrosomonas was abundant when associated with the mesopelagic layer and high concentrations of NO_2^- (Figure 8.9).

Overall, there are some variables that worked as connection points between rare, undetermined and abundant taxa, like the epipelagic Layer (associated with specific groups of rare and undetermined taxa), PO_4^- concentration, which divided rare and undetermined taxa, silicates, salinity and temperature (Figure 8.9). Low temperatures were associated with both groups of rules with rare taxa, and very low temperatures were at the intersection between the rare biosphere and both undetermined and abundant taxa (Figure 8.9).

8.3.4. Fine-grained analysis of taxonomy with association rule mining

From the previously used subset of statistically dependent, non-redundant and high-quality rules (Conviction > 5, Confidence > 90%, Lift > 1), we made a subset of rules that include taxonomic information and removed redundancy based on complexity, to maximize the amount of information available. Likewise, we obtained a subset of 7 rules (Supplementary Table 11.44), which is a number of rules that is human interpretable. One of those rules was previously described (AR1). The remaining 6 rules provide relevant information and are described below.

{taxon = genus *SAR92 clade* (*Gammaproteobacteria*), NO_2^- = very high, NH_4^+ = Unknown} => {Classification = Undetermined} (Confidence = 93.33%)

Association rule 5 (AR5)

{taxon = order *Cryomorphaceae* (*Bacteroidota*), Pelagic Layer = Mesopelagic, PO_4^- = very high, NO_2^- = very high, Chl *a* = Unknown} => {Classification=Undetermined} (Confidence = 92.86%)

Association rule 6 (AR6)

{taxon = class *OM190* (*Planctomycetota*), PO_4^- = very high, Si = very high, Chl *a* = Unknown, Temperature = very low} => {Classification=Undetermined} (Confidence = 92.86%)

Association rule 7 (AR7)

{taxon = order *Clade III* (*Alphaproteobacteria*, *SAR11 clade*)} => {Classification = Rare} (Confidence = 100%)

Association rule 8 (AR8)

{taxon = order *KI89A clade* (*Gammaproteobacteria*, *Pseudomonadales*)} => {Classification=Rare} (Confidence = 100%)

Association rule 9 (AR9)

{taxon = *clade IV* (*Alphaproteobacteria*, *SAR11 clade*)} => {Classification=Rare} (Confidence = 98.98%)

Association rule 10 (AR10)

From this subset of rules, the ones related to undetermined and abundant taxa presented environmental information (AR1, 5-7), while the ones associated with the rare biosphere did not show environmental information (AR8-10).

AR5 indicates that when there were high values of NO_2^-, the *SAR92 clade* would tend to have an intermediate level of abundance. According to AR6, at the mesopelagic layer, under very high concentrations of NO_2^- and PO_4^-, the order *Cryomorphaceae* would tend to have an intermediate level of abundance. And, according to AR7, under very high concentrations of PO_4^- and silicates, and very low temperatures, the class *OM190* (from *Planctomycetota*) would tend to be at intermediate levels of abundance. Regarding rules AR8-10, they indicate that ASVs with taxonomic information down to order *Clade III* (AR8) and order *KI89A clade* (AR9) were always rare, and that *Clade IV* genus of *Alphaproteobacteria* was almost always rare (98.98%).

There are more rules available, if the quality metrics are relaxed. For more examples, see Supplementary Tables 11.35 and Supplementary Data 8.13. For additional dependent and non-redundant rules with taxonomic information, see Supplementary Table 11.41. Some of the unmentioned rules include potentially relevant information. For example, within family *Flavobacteriaceae*, genera *NS4*, *NS5*, and *NS9* marine groups, and *Ulvitobacter* were undetermined in association with different variables, with genus *NS5* being associated with the highest values of temperature and silicates (Confidence = 88.24%), while *NS4* was only associated with the epipelagic layer (Confidence = 82.35%), *NS9* was associated with low salinity values (Confidence = 87.5%), and *Ulvibacter* was associated with the highest temperatures (Confidence = 82.35%) (Supplementary Table 11.41). Additionally, some taxa relevant for nitrogen cycling were also identified in association rules (genus *LS-NOB* and *Nitrosomonas*) (Supplementary Table 11.41). Genus *LS-NOB* was undetermined in association with very high silicates, but very low temperatures (Confidence = 87.5%), and genus *Nitrosomonas* was undetermined in association with the mesopelagic layer and very high silicates

concentrations (Confidence = 86.67%) (Supplementary Table 11.41). Within the family *Rhodobacteraceae*, there were different rules for the genera *Ascidiaceihabitans* and *Planktomarina*, with *Ascidiaceihabitans* being undetermined in association with the epipelagic layer and high values of silicate concentration (Confidence = 81.25%), and *Planktomarina* being undetermined in association with very low values of phosphate concentration (Confidence = 88.89%) (Supplementary Table 11.41). From phylum *Verrucomicrobiota*, the genus *Roseibacillus* was undetermined in association with very phosphate and nitrite concentrations (Confidence = 86.67%) (Supplementary Table 11.41).

To describe association rules, it is not possible to do so for hundreds or thousands of them. However, it is possible to search for rules that include specific taxonomic groups of interest, or specific environmental variables, for targeted analysis. To provide an example we analyzed the rules that include *Cyanobacteria*. Accordingly, we can select all rules from the original dataset with information on *Cyanobacteria* in the antecedent, then we select the rules with highest Lift and remove redundant rules based on the Complexity metric. For this example, we use Complexity to remove redundancy, because we are interested in finding rules with a lot of information. By using this approach, we obtained one statistically strong rule:

{taxon = genus *Synechococcus CC9902* (*Cyanobacteria*), Pelagic layer = Epipelagic, Salinity = very low, Water mass = SW} => {Classification=Rare} (Confidence = 100%)

Association rule 11 (AR11)

AR11 shows that if the genus *Synhechococcus* was present at the epipelagic layer, under very low salinity conditions and at the SW water mass, then it was a member of the rare biosphere.

8.3.5. Monitoring targeted taxonomic groups by combining unsupervised and rule-based machine learning

By using unsupervised machine learning, with *ulrb*, we were able to attribute abundance classifications (rare, undetermined, and abundant) to all taxa, within each sample. Thus, it becomes possible to compare how a taxonomic group oscillates between different abundance classifications, across specific environmental variables. Furthermore, the rules obtained by ARM can be used as guides for more targeted analysis. Thus, we used the ARM results to monitor taxa, across the variables highlighted by selected ARs (AR1, AR2, AR3; AR5; AR6; AR7; and AR11). Only statistically strong rules holding both

taxonomic and environmental information were selected, so as to avoid an excessive number of comparisons.

For rule AR1 (as well as other redundant rules, like AR2), the group *SAR11 Clade Ia*, from *Proteobacteria*, were associated with samples with the highest levels of silicate and nitrite concentrations. By plotting the abundance classification obtained by *SAR 11 Clade Ia* as a function of silicate and nitrite concentrations (Supplementary Figure 10.15), it was possible to see that this group was very often rare, undetermined and abundant in the epipelagic layer, become more abundant than undetermined in the mesopelagic layer, with only a few occurrences as rare, and was mostly rare in the bathypelagic layer. By also accounting for AR2, the impact of nitrite and silicate concentrations becomes clearer, since AR2 accounted for the mesopelagic layer, where *SAR11 Clade Ia* is mostly abundant, under some of the highest concentrations of nitrite and silicate (Supplementary Figure 10.55). For rule AR3, we compared the abundance classifications obtained by *Flavobacteriaceae* (*Bacteroidota*) as a function of silicate concentration (Supplementary Figure 10.55). Again, the impact of silicate concentration followed the highly stratified water column, whereby *Flavobacteriaceae* was mostly abundant and undetermined in the epipelagic layer, under lower concentrations of silicate, but become mostly intermediate in the mesopelagic layer, and rare in the bathypelagic layer. For AR5, we compared the abundance classifications of the genus *SAR92 clade* (*Gammaproteobacteria*) across nitrite concentrations (Supplementary Figure 10.55). *SAR92* clade oscillated mostly between rare and undetermined, with a few exceptions in the mesopelagic layer where it was occasionally abundant (Supplementary Figure 10.55). However, *SAR92* clade was a group that was mostly undetermined, *i.e.*, neither rare nor abundant. From AR6, we compared *Cryomorphaceae* (*Bacteroidota*) against nitrite and phosphate concentrations (Figure 8.10a). This analysis clearly showed that *Cryomorphaceae* were mostly undetermined in the mesopelagic layer, with samples of high nitrate and phosphate concentrations, which reflect AR6 (Figure 8.10a). Furthermore, it showed that *Cryomorphaceae* was able to oscillate from rare to abundant in the epipelagic layer but diminished to mostly undetermined in the mesopelagic layer and almost disappeared in the bathypelagic layer, as always rare (Figure 8.10a). From AR7, the class *OM190* (*Planctomycetota*) was selected and compared against temperature and phosphate concentration (Figure 8.10b). *OM190* was always rare in the epipelagic layer, but in the mesopelagic layer it was found across the entire range of abundance classifications (from rare to abundant), and was undetermined in the bathypelagic layer, where conditions reflected the lowest

temperature values and highest phosphate concentrations (Figure 8.10b), which reflects AR7. Finally, AR11 associated the rare genus *Synechococcus CC9902* (*Cyanobacteria*) with low salinity samples in SW and in epipelagic layers (Figure 8.10c). In fact, *Synechococcus CC9902* was always rare in the SW water mass, while it varied from rare to abundant in other water masses of the epipelagic layer and was mostly rare below the epipelagic layer (Figure 8.10c).

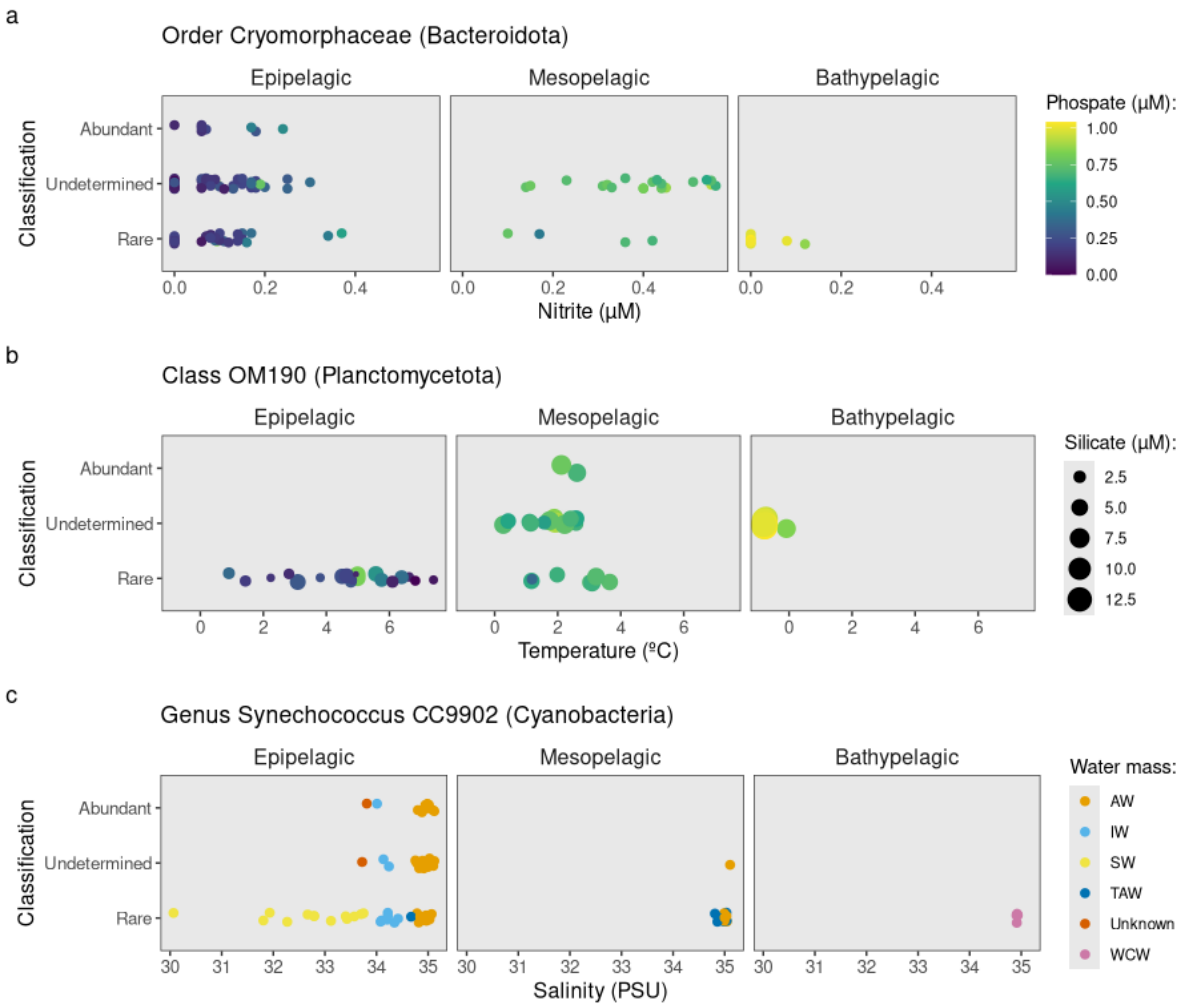


Figure 8.10. Abundance classification of target taxonomic groups. **a.** Abundance classifications of *Cryomorphaceae* (*Bacteroidota*) as a function of nitrite concentration, in x axis, and colored by phosphate concentration (n = 108 samples); **b**. Abundance classifications of *OM190* (*Planctomycetota*) as a function of temperature (ºC), in x axis, silicate concentration, represented by point size, and phosphate concentration, in "viridis" color palette (n = 52); **c**. Abundance classifications of *Synechococcus CC9902* as a function of salinity (PSU), with points colored by water masses (n = 82). All figures were divided into grids for the epipelagic, mesopelagic and bathypelagic layers.

8.4. Discussion

8.4.1. Prokaryotic composition follows the highly stratified water column

Climate change is rapidly changing the entire Arctic, with increasing temperatures, lower salinity near ice melting and Atlantification being of particular importance for the Fram Strait and Svalbard regions (Skogseth *et al.*, 2020; Tesi *et al.*, 2021). A growing number of studies in the area have been addressing several aspects of this changing ecosystem, including multi-year datasets (Priest *et al.*, 2023; Hörstmann *et al.*, 2024; Kraemer *et al.*, 2024; Wietz *et al.*, 2024).

Our multi-year analysis found no clear pattern of prokaryotic diversity across years. Instead, the main factor structuring prokaryotic communities during the five year period was temperature, silicates and the water column, *i.e.*, pelagic layers and respective water masses. The impact of the water column agrees with previous literature (Wilson *et al.*, 2017; Fadeev *et al.*, 2018; Müller *et al.*, 2018; Quero *et al.*, 2020; Cardozo-Mino *et al.*, 2021; Pascoal *et al.*, 2022a; Priest *et al.*, 2023; Vipindas *et al.*, 2023) (Chapter 5), but the impact of temperature and silicates was overlooked by previous studies.

The water mass effect was consistent with the clear pattern between temperature and diversity, because water masses are defined based on temperature and salinity profiles (Cottier *et al.*, 2005), and deeper pelagic layers are generally cooler. Furthermore, it is worth noting that deep layer prokaryotic communities were very distinct from the surface and characterized by a high diversity of rare taxa, possibly representing a natural seed bank. Studies that compared pole-to-pole diversity, *i.e.*, Arctic vs Antarctic prokaryotic diversity, have consistently reported that the prokaryotic communities at deeper sites tend to be similar in taxonomic composition and functioning across poles (Ghiglione *et al.*, 2012; Zhang *et al.*, 2020). From our statistical analysis, it was clear that temperature negatively correlated with diversity of the rare biosphere, indicating a loss of rare taxa, but a maintenance of abundant taxa. This pattern might indicate that rare taxa are more responsive to changing conditions, while the abundant are more resistant. Combining the effect of temperature with the water column profiles, it seems that there are isolated hotspots of diversity in the Arctic, which are probably well adapted to extremely cold conditions (near 0 ºC). Additionally, by comparing the impact of both depth and temperature, we confirmed that depth, by itself, did not explain changes in the number of taxa within pelagic layers, contrary to temperature, which supports the impact of

temperature on prokaryotic diversity. However, we also found that silicates correlated strongly with diversity, which might indicate a possible co-effect of temperature and silicates, and not merely a response to increasing depth.

Previous studies have looked at the effect of temperature on the prokaryotes of the Arctic (Delpech *et al.*, 2021; Vipindas *et al.*, 2023; Hörstmann *et al.*, 2024; Wietz *et al.*, 2024), which is particularly important because of the Arctic amplification processes (Box *et al.*, 2019; Rantanen *et al.*, 2022). In general, those studies agree with our results, in that temperature and water mass were the most important factors in shaping prokaryotic diversity (Delpech *et al.*, 2021; Vipindas *et al.*, 2023; Hörstmann *et al.*, 2024; Wietz *et al.*, 2024) and we show that deeper sites represent more unique communities, because of rare taxa. However, contrary to our analyses, those studies did not identify the impact of silicates, and did not distinguish between the impact of specific water masses, pelagic layers and temperature. In fact, our analyses revealed counter-intuitive patterns, whereby low temperatures correlated with higher diversity. This was unexpected, because temperature is usually positively correlated with diversity, which is why there is a general trend of decreasing diversity with increasing latitude, shown in studies of the Arctic region (Ghiglione *et al.*, 2012; Hooper *et al.*, 2024) and global surveys (Salazar *et al.*, 2019; Sunagawa *et al.*, 2020; Martin *et al.*, 2021). Our results, therefore, do not support the notion that temperature increases diversity by generally increasing metabolic rates. Instead, our results support the existence of local communities adapted for the specific conditions of their water masses, which is also supported by the previous finding that the community composition of the rare biosphere is distinct at different geographic regions of the Arctic Ocean results reported by previous literature (Galand *et al.*, 2009; Fadeev *et al.*, 2018; Pascoal *et al.*, 2022a) (Chapter 5). In the Barents Sea, a study has found no correlation between temperature and decreasing diversity, but it was limited to surface samples (Aalto *et al.*, 2022). In fact, our study shows that the epipelagic layer may not suffice to fully understand the impact of different environmental variables on prokaryotic diversity, including temperature and silicates.

Salinity is often mentioned as another important factor for prokaryotic community composition, often together with temperature (Fadeev *et al.*, 2018; Gorrasi *et al.*, 2019; Aalto *et al.*, 2022), or by itself (Yergeau *et al.*, 2017; Handler *et al.*, 2023; Mioduchowska *et al.*, 2023). The effect of salinity might be even stronger in areas where sea-ice melting is directly involved (Fernández-Gómez *et al.*, 2019). Even though we did not find a strong signal for salinity specifically, the impact of both temperature and salinity on prokaryotic communities seems to be present in the Arctic Ocean. This is consistent with the highly

stratified nature of the Arctic Ocean water column. Water masses may be important to shape local communities by working as a dispersal barrier and by providing unique conditions (Pascoal *et al.*, 2022a; Priest *et al.*, 2023; Vipindas *et al.*, 2023) (Chapter 5), which is probably why the most isolated water masses (like WCW) accumulate such unique diversity of rare taxa. Another variable that has been reported to influence prokaryotic communities together with temperature is silicates concentration (Vipindas *et al.*, 2023). Our results are consistent with this, showing that silicates concentrations increased with depth, together with temperature decreases, culminating in highest diversity at the lowest temperatures and highest silicates concentrations. This high diversity was derived specifically from the rare biosphere, because the number of abundant and undetermined taxa remained constant as a function of silicates concentration. Previous studies have justified the increased diversity of microorganisms in deeper layers based on the availability of carbon from sinking particles in the mesopelagic layer, and chemoautotrophic and mixotrophic metabolisms at deeper layers (Swan *et al.*, 2011; Vipindas *et al.*, 2023). In this context, higher concentration of silicates, likely derived from lysed and decanting phytoplankton, might result from marine snow aggregation processes which ultimately would support the establishment of particle-attached microbial communities in deeper waters. However, in other environments, like subglacial ecosystems, silicates have been found to support microbial communities in the absence of carbon sources of energy (Telling *et al.*, 2015; Toubes-Rodrigo *et al.*, 2021).

Other nutrients analyzed in our dataset did not have a clear impact on diversity or community composition, but a previous study on the Beaufort gyre observed that Chl *a* and ammonia concentrations impacted more the prokaryotic communities in surface waters than other inorganic nutrients (Vipindas *et al.*, 2023).

8.4.2. Rare, undetermined, and abundant taxa follow distinct environmental patterns

As mentioned previously, the rare biosphere was shaped locally, which agrees with previous literature (Galand *et al.*, 2009; Pascoal *et al.*, 2022a) (Chapter 5). By means of rule-based machine learning, we were able to confirm that the rare biosphere responds to a different set of environmental variables than undetermined or abundant taxa. More specifically, the rare biosphere was associated with nutrient scarcity, consistent with higher rare taxa diversity at the bathypelagic layer, together with high salinity and silicate concentration. The ARM results, in general, were more informative than the classical

statistical modeling results (linear least squares and Poisson models, Pearson and Kendall correlations). However, most rules for the rare biosphere provided more taxonomic than environmental information (*e.g.*, AR8), while the rules for abundant and undetermined taxa revealed more complex environmental interactions and had stronger statistical support (*e.g.*, AR1). This difference in statistical strength was expected, because the rare biosphere, by definition, will have low abundance and the abundant taxa will have high abundance. The stronger environmental associations with abundant taxa might also reflect a response to selective pressures, while the rare biosphere might be more randomly distributed. However, this contradicts the notion that the rare biosphere is locally distributed in particular water masses and/or geographical regions (Galand *et al.*, 2009; Pascoal *et al.*, 2022a) (Chapter 5). A possible explanation is that the rare biosphere accumulates in certain water masses and geographical regions by stochastic processes, where it remains rare, but viable, and does not respond strongly to local variables. Such a scenario would be consistent with the "stochastic vs deterministic processes" framework used to explain the distribution of the rare biosphere (Jia, Dini-Andreote and Falcão Salles, 2018; Pascoal, Costa and Magalhães, 2021) (Chapter 2).

Importantly, the rule-based machine learning approach was able to attribute clearly distinct environmental factors and taxonomy to the rare, undetermined and abundant taxa, which were classified by unsupervised machine learning, *ulrb*, (Pascoal *et al.*, 2025a; 2025b) (Chapters 6 and 7). This methodology to define the rare biosphere (together with undetermined and abundant taxa) was proposed to substitute arbitrary thresholds, such as 0.1% relative abundance per sample, because it can provide biologically and ecologically meaningful classifications (Pascoal *et al.*, 2025a) (Chapter 6). Thus, our rule-based machine learning results provide further evidence that unsupervised machine learning can group taxa into meaningful and useful abundance classifications. As far as we know, this is the first report of the utilization of rule-based machine learning approaches to disentangle complex interactions between abundance classifications, taxonomic groups and environmental variables.

Some advantages of ARM include the ability to keep all samples available in the dataset, which is not possible when applying classical statistical models and/or correlations, because of the existence of missing values. Thus, since ARM can deal with missing values without compromising the statistical strength of the rules obtained, it helps to maximize the sample size used in experimental works.

8.4.3. Targeted taxonomic analysis

Generally, there is an agreement between studies, whereby at phylum and class level, *Proteobacteria* (most notably *Gamma*-, *Alpha*- and *Betaproteobacteria*) and *Bacteroidetes* are some of the most abundant groups in the Arctic Ocean (Comeau *et al.*, 2011; Ghiglione *et al.*, 2012; Yergeau *et al.*, 2017; de Sousa *et al.*, 2019; Fernández-Gómez *et al.*, 2019; Cardozo-Mino *et al.*, 2021; Delpech *et al.*, 2021; Pascoal *et al.*, 2022a; Thiele *et al.*, 2022). This was consistent with our general overview, but some studies found differences in the dynamics of these groups under different conditions (like proximity to ice meltwater, regions, seasonality) (Fadeev *et al.*, 2018; Cardozo-Mino *et al.*, 2021). For example, it has been argued that groups like *Gammaproteobacteria* and *Flavobacteria* grow in the summer, under post-bloom and ice-free conditions (Fadeev *et al.*, 2018; Cardozo-Mino *et al.*, 2021), which might be an indicator of increased heterotrophy. However, our data did not tap into seasonality, instead looking at a single season (summer) across different years. In that regard, major differences at phylum level were detected across the water column but were fairly constant between distinct years. Except for decreasing abundance levels of *Alphaproteobacteria* relative to other groups in the epipelagic layer, and the growth of “Other” taxonomic groups (outside the top 5 most abundant) in the bathypelagic layer. The distinct taxonomic composition of the bathypelagic layer was consistent with alpha and beta diversity, supporting the existence of a distinct prokaryotic community at deeper layers. Our ARM results provided more detail to *Gammaproteobacteria* and *Bacteroidetes*. Specifically, *SAR92 Clade* (*Gammaproteobacteria*) was associated with the samples with highest nitrite concentration as being neither rare nor abundant (undetermined). *SAR92* group is known as an abundant oligotrophic group in the ocean (Stingl *et al.*, 2007) and has been identified in previous reports of prokaryotic diversity in the Arctic Ocean (Lovejoy, Galand and Kirchman, 2011), which have associated it with early summer conditions (de Sousa *et al.*, 2019). However, our follow up analysis of *SAR92 clade* revealed that it was always undetermined or rare in the epipelagic layer and only became abundant in the mesopelagic layer.

On the *Bacteroidetes* side, the *Flavobacteriia* class is often mentioned as an abundant group in Arctic Ocean surveys (Comeau *et al.*, 2011; Bowman *et al.*, 2012; Ghiglione *et al.*, 2012; Ortega-Retuerta *et al.*, 2013; Yergeau *et al.*, 2017; Fadeev *et al.*, 2018; de Sousa *et al.*, 2019; Fernández-Gómez *et al.*, 2019; Thomas, Sinha and Krishnan, 2020; Vipindas *et al.*, 2023). Members of the *Flavobacteriales* order have been linked to degradation of polysaccharides (Teeling *et al.*, 2016), consistent with reports linking

Flavobacteriaceae taxa with blooms of *Phaeocystis pouchetii* both in the Arctic (de Sousa *et al.*, 2019) and Antarctic waters (Delmont *et al.*, 2014). Within this order, ARM was able to obtain informative rules on two families, *Flavobacteriaceae* and *Cryomorphaceae*. More specifically, ARM identified rules associating abundant *Flavobacteriaceae* family with epipelagic samples and with the lowest concentrations of silicates, which is consistent with more general reports on the *Flavobacteriia* class (de Sousa *et al.*, 2019). A follow up analysis of this group further revealed that it was also mostly abundant and undetermined in the mesopelagic layer, only becoming mostly rare in the bathypelagic layer. ARM associated the family *Cryomorphaceae* with being neither rare nor abundant in the mesopelagic layer samples, which also have the highest PO_4^- and NO_2^- concentrations. A follow up analysis on this group further revealed that it was most versatile in the epipelagic layer, because it was present as rare, undetermined and abundant quite often; but was mostly undetermined and rare in the mesopelagic layer; and rare in the bathypelagic layer, making *Cryomorphaceae* a good example of water column stratification impact on a widely distributed group. Thus, ARM was able to distinguish between closely related groups (two families within the *Flavobacteriales* order), distinguishing a widely reported group (*Flavobacteriaceae*) from a previously overlooked group (*Cryomorphaceae*).

Another relevant group that was highlighted by the ARM results and many Arctic Ocean reports is the class *SAR11* (also known as *Pelagibacterales*), from *Alphaproteobacteria* (Comeau *et al.*, 2011; Lovejoy, Galand and Kirchman, 2011; Bowman *et al.*, 2012; Fadeev *et al.*, 2018; de Sousa *et al.*, 2019; Kraemer *et al.*, 2020; Cardozo-Mino *et al.*, 2021; Delpech *et al.*, 2021; Han *et al.*, 2021; Wietz *et al.*, 2024). *SAR11* is recognized as one of the most abundant heterotrophs in marine environments and is known for having a streamlined genome (Morris *et al.*, 2002; Tripp, 2013; Giovannoni, 2017). An in-depth genomic analysis of *SAR11* in the Arctic Ocean has identified at least six different clades (Kraemer *et al.*, 2020), but we will focus on *Clade Ia*, which was selected by several rules (*e.g.*, AR1). A previous study reported an increase of *SAR11 Clade Ia* from June to August, justified by the oligotrophic preferences of this genus (Delpech *et al.*, 2021). The same report found positive correlations between *SAR11 Clade Ia* and silicates concentration, temperature, and particulate organic matter (Delpech *et al.*, 2021). Consistent with other studies that related *SAR11 Clade Ia* with the stratified water column, mainly because of temperature (Selje, Simon and Brinkhoff, 2004; Giebel *et al.*, 2011; Sperling *et al.*, 2012; Tada *et al.*, 2013; Giovannoni, 2017; Kraemer *et al.*, 2020; Cardozo-Mino *et al.*, 2021). We did not test particulate organic matter, but our rules

identified associations between abundant *SAR11 Clade Ia* taxa, temperature and silicates concentration. Specifically, *SAR11 Clade Ia* was widespread as both rare, undetermined and abundant in the epipelagic layer, becoming more abundant in the mesopelagic layer, but mostly rare in the bathypelagic layer. The differences in abundance classifications across the water column clearly reflect differences in both temperature and silicate concentrations. Thus, our results support the notion that *SAR11 Clade Ia* takes advantage of the stratified water column in the Arctic Ocean (Cardozo-Mino *et al.*, 2021; Delpech *et al.*, 2021). Even though *SAR11* are aerobic heterotrophs adapted to live in the oligotrophic ocean conditions, which explains why they are widespread in surface layers, they were mostly abundant in the mesopelagic layer, which might be explained by the finding that *SAR11* can also thrive in anoxic regions and reduce nitrate to nitrite in the absence of oxygen (Tsementzi *et al.*, 2016), which is consistent with our associations between *SAR11* being abundant and present in the mesopelagic layer, together with high concentrations of nitrites.

In this study we also identified members of the genus *Synechococcus*, which was rare in surface waters, but varied from rare to abundant in other water masses of the epipelagic layer. They were very scarce below the epipelagic layer, which was expected from a cyanobacterial group. *Cyanobacteria* has been considered generally scarce in the Arctic Ocean by earlier reports (Lovejoy, Galand and Kirchman, 2011; Pedrós-Alió, Potvin and Lovejoy, 2015). However, many studies have reported *Cyanobacteria* in the Arctic Ocean (Waleron *et al.*, 2007; Paulsen *et al.*, 2016; Delpech *et al.*, 2021; von Jackowski *et al.*, 2022), including the results presented in this chapter. The presence of *Cyanobacteria* in the Arctic Ocean has been considered allochthonous, coming mainly from ice melting, riverine sources and Atlantic waters (Waleron *et al.*, 2007; Paulsen *et al.*, 2016; Gorrasi *et al.*, 2019; Coelho *et al.*, 2022; Mioduchowska *et al.*, 2023). Some of those reports have also specifically identified the genus *Synechococcus* (Waleron *et al.*, 2007; Paulsen *et al.*, 2016). Since SW is very characteristic of the Arctic Ocean, with lower salinity values, it makes sense that *Synechococcus* was always rare in those waters. Under climate warming conditions, *Cyanobacteria* are expected to become more abundant in the Arctic Ocean, because their biogeographical distribution mainly follows temperature (Mioduchowska *et al.*, 2023; Hooper *et al.*, 2024), and they might outcompete local algae in primary production (Edwards *et al.*, 2020). Some studies have raised questions about the validity of cyanobacterial findings in the Arctic Ocean, because of 16S rRNA gene limitations, finding that some of their cyanobacterial assigned sequences were derived from chloroplasts (Fernández-Gómez *et al.*, 2019; Quero *et al.*,

2020). However, other studies have identified cyanobacterial groups in the Arctic Ocean using either metagenomics or other marker genes (Paulsen *et al.*, 2016; Duncan *et al.*, 2022), supporting the overall evidence that *Cyanobacteria* are present, often as rare, in the Arctic Ocean. Regarding our data, we manually verified that *Synechococcus* sequences were from *Cyanobacteria*, using BLASTN against the NCBI reference (Zhang *et al.*, 2000).

8.4.4. Assessing types of rarity by monitoring targeted taxa

The concept of “types of rarity” refers to the transitions that taxa undergo between abundance classifications (Lynch and Neufeld, 2015; Jia, Dini-Andreote and Falcão Salles, 2018). For example, a taxon that grows from rare to abundant is considered a conditionally rare taxon (Shade *et al.*, 2014). However, describing the types of rarity is a non-trivial problem. A major challenge in the description of types of rarity is the numerous possible outcomes. For example, if we consider three abundance classifications (rare, undetermined and abundant), there are nine possible paths between two points. Even if many combinations are manageable between two points, in datasets with hundreds of samples and multiple environmental variables at play, the number of possibilities blows up quickly. And, finally, the order in which samples are compared also matters, further increasing the number of possible outcomes. Thus, we consider that comparing and quantifying the different types of rarity is still open to research, specifically for large datasets. For smaller datasets, with simple experimental designs, however, we have previously shown that it is possible to easily quantify types of rarity by grouping them into commonly used categories (permanently rare/abundant, conditionally rare, and transiently rare) (Pascoal *et al.*, 2022a) (Chapter 5). Our former analysis of types of rarity in the Arctic Ocean, in the north of Svalbard region, showed that most rare taxa remained rare or disappeared across different water masses (Pascoal *et al.*, 2022a) (Chapter 5). Other studies of the rare biosphere in the Arctic Ocean have also highlighted that most rare taxa remain so, or disappear (Galand *et al.*, 2009; Kirchman, Cottrell and Lovejoy, 2010). Collectively, those reports were consistent with our finding of hotspots of rare biosphere diversity in specific water masses (particularly, the deeper ones) and net loss of rare taxa under changing conditions.

Even though we did not quantify the different types of rarity, because of the abovementioned unresolved challenges, we monitored the abundance classifications of specific taxa, under relevant environmental variables. This approach was possible thanks to *ulrb* ability to make abundance classifications for each taxon within specific

samples (Pascoal *et al.*, 2025a; 2025b) (Chapters 6 and 7). Furthermore, by combining *ulrb* with ARM, it was possible to select what taxonomic groups and environmental variables to look at more closely, aiding in the overall interpretation of the rare, undetermined and abundant taxa dynamics. By monitoring specific taxa, even though we are not ascribing explicit types of rarity, it becomes clear what are the dynamics at play. For example, *SAR11 Clade Ia*, was present as rare, undetermined and abundant in the epipelagic layer, but was mostly undetermined in the mesopelagic layer and rare in the bathypelagic layer. Showing, for example, that *SAR11 Clade Ia* behaves as a conditionally rare taxon in the epipelagic layer, where it can thrive under different conditions, but becomes mostly abundant at mesopelagic pelagic layers. Another interesting example was class *OM190*, which was permanently rare at the epipelagic layer, which is the typical type of rarity expected in the Arctic Ocean (Galand *et al.*, 2009; Kirchman, Cottrell and Lovejoy, 2010; Pascoal *et al.*, 2022a) (Chapter 5), but became conditionally rare in the mesopelagic layer, and then stayed as undetermined in the bathypelagic layer.

8.5. Conclusion and future work

This study used a multi-year dataset of V4V5 16S rRNA gene amplicon sequencing of the Arctic Ocean (from Kongsfjorden to Fram Strait) to access prokaryotic diversity, with a particular focus on the dynamics of different abundance classifications. Our main results show that there is a clear negative relationship between prokaryotic diversity and temperature, with an accentuated loss of rare taxa under increasing temperatures. The prokaryotic community stratification along the water column was expected, but we found that temperature and silicates were correlated with diversity, which are parameters overlooked in previous studies. We also found that the impact of such parameters might have been underestimated based on the limited depth range of most studies in the Arctic Ocean.

The impact of temperature was mostly derived from differences in pelagic waters and water masses, which was consistent with the literature.

Besides the baseline, traditional ecological analysis, we show that unsupervised and rule-based machine learning can be combined to achieve more powerful analysis of prokaryotic dynamics, by sorting organisms into informative abundance classifications (rare, undetermined and abundant), and by automatically identifying interesting associations between taxonomic groups, complex combinations of environmental variables and abundance classifications. Furthermore, we showed for the first time that the rule-based machine learning approach can connect the abundance classifications obtained by unsupervised machine learning with distinct biological and environmental patterns. Thus, we provide evidence that it is possible to define the rare biosphere in a biologically meaningful way, without the use of arbitrary thresholds. We also show that rule-based machine learning is an interesting approach to identify potentially interesting taxonomic groups for deeper analysis, because it automatically identifies associations that are derived exclusively from the available data. Indeed, the approach presented here was able to distinguish complex environmental associations between closely related taxa, including taxa that are usually overlooked in microbial ecology surveys, like *NS9* and *NS5* genera of the family *Flavobacteriaceae*.

Future work is still necessary both on the ecological and on the methodological side. On the ecological side, it will be important to distinguish between seasonal, interannual and geographical impacts on prokaryotic diversity. Work is also needed to disentangle the stochastic from deterministic processes at play in the Arctic Ocean. More specifically, some variables are usually correlated with prokaryotic communities composition,

together with the highly stratified water column, but it is hard to know to what extent such variables are selecting for particular taxa, or if they are just confounding variables relative to the water masses, which could simply impose dispersal limitation to randomly distributed taxa. We propose that it is probably a combination of both processes, with deterministic processes explaining the conditionally rare taxa (like *SAR11 Clade Ia* in the epipelagic layer), and stochastic processes explaining the permanently rare taxa (like *Synechococcus* in surface waters). This framework is consistent with previous hypotheses on the impact of stochastic and deterministic processes on the rare biosphere (Jia, Dini-Andreote and Falcão Salles, 2018). Namely, the deterministic processes of homogeneous selection to explain permanently rare taxa, selective pressure to explain taxa that follow seasonal patterns, and dispersal to explain stochastic processes.

On the methodological side, rule-based machine learning proved useful to identify potentially interesting taxa. Thus, future work should attempt to develop tools with broader applicability for microbial ecology studies. Furthermore, rule-based machine learning still provides a huge number of rules, which are hard to interpret by the human user. Thus, future work will be necessary to improve noise reduction and the ability to select interesting and non-redundant rules. Finally, the rules obtained should be interpreted together with the underlying data for a better understanding of the microbial ecology itself. We illustrated this by monitoring how the abundance classification of targeted taxa changed according to the variables selected by the rules obtained.

Future work should also add full-length 16S rRNA gene sequencing, to enable higher taxonomic resolution and accuracy, and metagenomes should also be considered, to account for the functional component.

8.6. Declarations

Author contributions

Conceptualization (all authors); data curation (FP); formal analysis (FP); funding acquisition (all authors); investigation (FP); methodology (FP); project administration (CM, PB and RC); resources (CM); software (FP); supervision (CM, PB and RC); validation (all authors); visualization (FP); writing - original draft (FP); writing – review and editing (all authors).

Acknowledgments

We acknowledge the precious help from Pedro Duarte and Philip Assmy, and their institution, Norwegian Polar Institute, because of their help in the sampling campaigns and ongoing collaboration to obtain environmental data from the Arctic Ocean.

Funding

FCT funded this work through the projects 2022.02983.PTDC and PTDC/CTA-AMB/4946/2020 of CM, UIDB/04423/2020 and UIDP/04423/2020 of CIIMAR, and UIDB/04565/2020 and UIDP/04565/2020 of iBB. FCT also funded a PhD grant to FP (2020.04453). The work of Paula Branco was undertaken, in part, thanks to funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant.

Data and code availability

The source data and code for all analysis is available in a GitHub repository (https://github.com/pascoalf/PhD_thesis_chapter_8). Source FASTAQ files are available at the European Nucleotide Archive: MOSJ2016 (PRJEB24517); MOSJ2017 (PRJEB72025); MOSJ2018 (PRJEB72030); MOSJ2019 (PRJEB60815); MOSJ2020 (PRJEB72034). For sample metadata see section 8.7. Supplementary Data and Zenodo (10.5281/zenodo.14755201).

8.7. Supplementary Data

Supplementary Data of Chapter 8 is publicly available in Zenodo under the doi: 10.5281/zenodo.14755201

Supplementary Data 8.1. Sample metadata, adapted from Pascoal *et al*. (2025a). Nutrient measurements from Wold *et al*., (2024).

Supplementary Data 8.2. Dependent rules, for any Lift value, ordered by Conviction. Additional metrics are available (Support, Confidence, Coverage, Count and Chi-squared test).

General Discussion and Conclusion

9. General Discussion and Conclusion

9.1. General remarks

The microbial rare biosphere is a topic of increasing interest in the microbial ecology field, because of the recognition of its role in the ecosystem (Jousset *et al.*, 2017), as a source of genetic diversity, and even the potential for novel biotechnological applications (Pascoal, Magalhães and Costa, 2020). However, since its inception (Sogin *et al.*, 2006), the study of the microbial rare biosphere faces methodological difficulties, including controversies in the application of its definition, methodology selection and analysis.

This thesis aimed at solving methodological issues in the study of the microbial rare biosphere, with a particular focus on Arctic marine environments, but also addressing other environmental contexts, like host-associations. We reviewed all aspects of the microbial rare biosphere literature (Part I), compared how multiple methodologies describe microbial communities in their environmental context (Part II), proposed a new approach to delineate rare from non-rare microbial taxa (Part III), and used standardized prokaryotic data, together with unsupervised and rule-based machine learning approaches, to obtain novel insights into the ecology of the microbial rare biosphere in the Arctic Ocean (Part IV). Importantly, we were careful never to consider the microbial rare biosphere as an isolated and independent component of the community. Instead, we always analyzed the entire microbial communities and specific abundance classifications, and compared the rare against abundant taxa, where applicable. In accordance, our results and insights collectively support the notion that the microbial rare biosphere is an important piece of the puzzle in understanding overall microbial community dynamics, but it is also a component that should not be analyzed in isolation.

In this concluding chapter, we discuss the major findings and contributions of this thesis, but we also acknowledge its limitations, and open questions for future research.

9.2. Major findings and contributions

The major findings of this thesis were relative to the ecology of the microbial rare biosphere in the Arctic Ocean, inter-comparison of methodological aspects of microbial ecology, including aspects specific to the microbial rare biosphere, and application of machine learning strategies. There were also scientific contributions in the creation of an R package and publication of datasets. Below we showcase the major findings and contributions from each chapter and how they interconnect.

9.2.1. Insights from reviewing literature

The literature review from Part I extended beyond the microbial rare biosphere, to have an outlook on historical findings and concepts related to the study of rare species in general, an approach that proved useful to understand how different concepts came to be and how they impacted the emergence of the microbial rare biosphere field. In this instance, we found that Darwin was one of the first researchers to acknowledge the widespread existence and importance of rare species in the context of evolution (Darwin, 1859). However, studies in the XX century also identified the pattern of a high number of rare taxa, followed by a few abundant ones (Corbet, 1942; Fisher, Corbet and Williams, 1943; Preston, 1948), a pattern that was also identified for microorganisms thanks to the emergence of high-throughput sequencing (Sogin *et al.*, 2006). The perspective laid out by this thesis showcases the method-dependency of the microbial rare biosphere but also showed that both microbial and non-microbial communities share a similar rank abundance curve shape (Chapter 2). In fact, such common rank abundance curves partially explain why *ulrb* was applicable to non-microbial datasets (Chapter 6).

A historical perspective of recent methodological advances in the microbial rare biosphere in the XXI century is also useful. For instance, the earlier emergence of studies using amplicon sequencing of the 16S rRNA gene, compared to the 18S rRNA gene, explains why most studies of the microbial rare biosphere focus on prokaryotes. Similarly, there were much fewer studies on rare fungi, which require ITS (Internal Transcribed Spacer) amplicon sequencing, and even less for viruses (Chapter 2). The increasing availability of shotgun metagenomic sequencing, however, makes any taxonomic group accessible, as well as opening opportunities to study the functional side of the microbial rare biosphere. Accordingly, we argued in favor of the need to introduce metagenomes in the field of the microbial rare biosphere (Chapter 2) and further showed that shotgun metagenomic sequencing can, indeed, be used to survey the rare biosphere (Chapter 5 and 6). Regarding other recent methodological advances, we also supported the argument being made (Callahan, McMurdie and Holmes, 2017) for the utilization of ASVs instead of OTUs in microbial ecology (Chapter 2). Even though we recognized some of the disadvantages of using ASVs (Chapters 4, 5 and 6), they proved key for the modular nature of the MOSJ 2016-2020 dataset (Chapters 6 and 8). Currently, ASV utilization is becoming standard for prokaryotic community analysis.

We also summarized other recent advances that show potential to improve the analysis of the microbial rare biosphere, from host associations, to accessing of metagenome

assembled genomes, transcriptomics for identification of active functions, or even the assessment of the metabolism level.

Beyond methodological aspects, another noteworthy contribution from Part I of this thesis was the conceptual framework of the rare biosphere, whereby we tried to organize different theories and concepts in an interconnected way. Such an approach was based on community assembly theory applied to the rare biosphere (Jia, Dini-Andreote and Falcão Salles, 2018). Since then, the concepts organized in the Chapter's 2 framework are still being used in studies of the microbial rare biosphere (Jia, Dini-Andreote and Salles, 2022; Xu *et al*., 2022; Araujo *et al*., 2023; Ramond *et al*., 2023; Bontemps, Moënne-Loccoz and Hugoni, 2024).

9.2.2. Standardization enables the creation of modular, continuously growing multi-year datasets

The creation of the multi-year dataset of the Arctic Ocean (MOSJ 2016-2020) was not trivial (Chapter 6). To be able to continuously add new layers of data yearly, the dataset needed to be modular, whereby each "module" represents a batch of samples collected at the same time and which can be connected to another module. For this purpose, we needed to ensure that ASVs identified in one module represent the same biological unit in another module. This was possible by accounting for small variations in sampling methodology for each different module, ascertaining the impact of such variations, and then applying the same bioinformatics processing for all samples. We summarize how different parts of the thesis contributed to the modular nature of the MOSJ 2016-2020 in Figure 9.1.

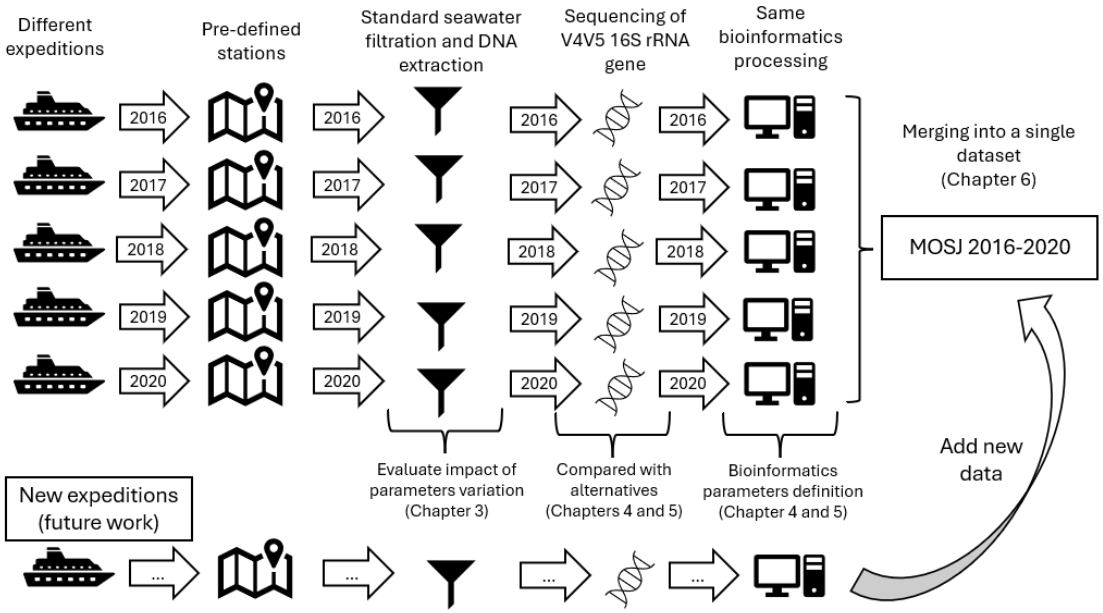


Figure 9.1. Schematic representation of MOSJ 2016-2020 modularity. Example of MOSJ 2016-2020 and how different chapters of this thesis contributed for its generation.

In the chapters of Part II, we systematically compared multiple methodologies, representative of the steps necessary for the analysis of the microbial rare biosphere, from seawater sampling to bioinformatics and data processing. Such comparisons were useful to establish strategies to combine multiple smaller datasets into a single standardized and bigger dataset (Figure 9.1). Not only the comparisons themselves, but also the optimization of R code for standardized raw data processing and high-quality filtering (Chapters 4 to 6). Such an effort allowed for the direct comparison of ASVs obtained across distinct environmental settings, which was used by machine learning to obtain informative patterns from the data (Chapter 8).

Collectively, these findings and approaches can help other researchers in the design of their own multi-year datasets and contribute to data interoperability.

9.2.3. Towards a solution to the microbial rare biosphere definition problems

Initially, this thesis considered the traditional approaches to define the microbial rare biosphere, *i.e.,* the relative abundance thresholds. However, we find such an approach to be ineffective and problematic, complicating inter-comparability between methodological variations (Chapter 5). We also verified that MultiCoLA (Gobet, Quince and Ramette, 2010; Jia, Dini-Andreote and Falcão Salles, 2018) was not a satisfactory

alternative, because it was reliant on human interpretation of a plot to decide a threshold, and because it was complicated to implement (Chapter 5). One important contribution of this thesis was the clarification of the problems regarding threshold-based approaches (Chapters 5 and 6), as well as the limitations of other alternatives, such as MultiCoLA (Chapter 5), and we surveyed the literature, highlighting that threshold-based approaches are overwhelmingly used (Chapter 2, 5 and 6). Therefore, this thesis argues that studies should provide a clear description of the method to define the rare biosphere and advises against using threshold-based methods or MultiCoLA.

In Part III, we presented a new methodology to define the rare biosphere, *ulrb*, including the necessary R code to run it, and relevant statistics to evaluate its results (Chapters 6 and 7). We showed that *ulrb* is applicable to abundance tables derived from different bioinformatics approaches, sequencing strategies, number of samples, number of taxa, sequencing depth, and counts data of animals and plants (Chapter 6). And we also showed that the classifications obtained by *ulrb* are directly tied to their ecology, by showing examples of taxa that thrive within healthy coral tissue (abundant), but decline in necrotic coral tissue, seawater and sediment (rare and undetermined). This last example clearly shows that *ulrb* provides a classification that can be used directly for ecological interpretation, *i.e.,* correctly identified the consequences of an eubiosis vs dysbiosis scenario (Chapter 6). Furthermore, it was consistent with the common finding that microbial symbionts persist as members of the microbial rare biosphere outside the host (reviewed in Chapter 2, but see also Tedersoo *et al*. (2021)).

It is important to note that in Part III we also highlighted the R code aspects of *ulrb*, whereby we developed an algorithm that can classify thousands of taxa, across hundreds of samples in a few seconds. Therefore, in Chapter 7 we extensively described the *ulrb* R package, highlighting all its functionalities, and explaining the R code strategy. For example, we followed “tidy” principles of data analysis (Wickham, 2014), and avoided the utilization of “for loops” entirely, using nested functions instead. Such advances were important to enable a quick and easy application of our definition of the rare biosphere to the big dataset used in Chapter 8. For instance, if we had used MultiCoLA instead, this step would have taken significantly more time. Furthermore, our learning materials and coding approach promoted a tidy style of data wrangling, which can be useful for other colleagues.

9.2.4. Identification of environmental variables shaping the rare biosphere in the Arctic Ocean

To provide tangible ecological insights into the microbial rare biosphere, we looked specifically at the Arctic Ocean (Chapter 8), in a region that is particularly influenced by Atlantification, the Fram Strait and Kongsfjorden (Promińska, Cisek and Walczowski, 2017; Tverberg *et al.*, 2019). Additionally, Chapter 5 addressed ecological questions regarding the winter to summer transition North of Svalbard, although with a much more limited dataset and analysis.

Collectively, all our datasets from the Arctic Ocean reflected the well-known vertical stratification of the water column, whereby water masses provide physical and chemical barriers to prokaryotic dispersion, which results in unique communities within specific water masses. This was more evident in the winter cooled water, using the MOSJ 2016-2020 dataset, whereby prokaryotic community remained constant over the years, but changed across depth. Of the environmental variables studied, temperature was the most important one (Chapter 8), because it indicated a possible climate change scenario in which warming temperatures reduce the Arctic Ocean rare prokaryotic diversity. We also verified that within epi- and mesopelagic layers, depth itself did not change the number of rare taxa, but temperature did so. Interestingly, silicate concentration was also shown to be correlated with prokaryotic diversity and associated with several groups (*e.g.*, *Clade Ia* genus of *SAR11*), by the rule-based machine learning strategy. However, we found surprisingly scarce literature on the relationship between prokaryotic diversity and silicates, with most of the literature focusing on diatoms (Tréguer *et al.*, 2018). Indeed, prokaryotes are virtually negligible in the global net silica cycle (Tréguer *et al.*, 2021).

9.2.5. Combining unsupervised and rule-based machine learning for in-depth analysis of microbial communities

In Chapter 8, we combined unsupervised and rule-based machine learning. In fact, both machine learning approaches were performing symmetric classifications, *i.e.,* they were both explaining the abundance classifications of taxa. More specifically, unsupervised machine learning obtained an abundance classification using the abundance score of taxa; then, rule-based machine learning explained the abundance classification using the environmental data instead of the abundance score (Figure 9.2).

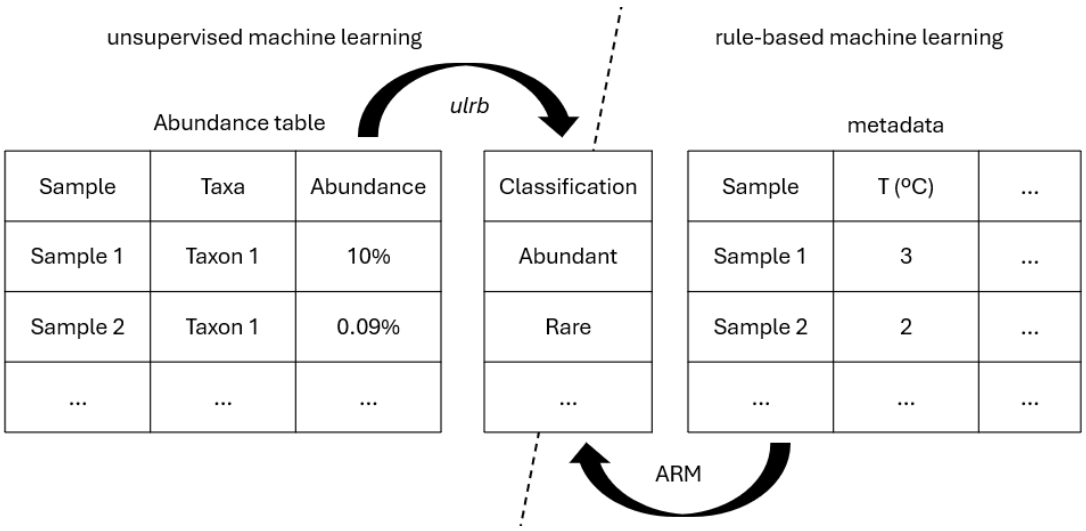


Figure 9.2. Schematic representation of symmetry between unsupervised and rule-based machine learning approaches taken. On the left side, unsupervised machine learning was used to classify taxa as either rare, undetermined, or abundant, using *ulrb*. On the right side, rule-based machine learning was used to identify associations between complex sets of variables and the classifications obtained by *ulrb*, using ARM.

This strategy (Figure 9.2) ultimately obtained new insights into the microbial rare biosphere, by confirming that taxa classified in different abundance levels (rare, undetermined, and abundant) were associated with different sets of environmental variables. It is important to consider the high granularity of this approach, whereby taxa of interest were automatically selected independently of our previous expectations or literature bias. For instance, we identified interesting patterns for both taxa commonly reported in the literature (*e.g.*, *SAR11*), and for taxa underreported in the literature (*e.g.*, *Cryomorphaceae*). Note that such taxa are commonly found in many microbial surveys, but there are so many different taxa, divided on multiple levels (usually, from phyla down to genus), that traditional microbial ecology analyses limit taxonomic analysis to the most well-known groups.

9.3. Experimental limitations

Despite this thesis' contribution for advances in the study of the microbial rare biosphere, and the study of general microbial ecology data, some limitations of this thesis are noteworthy and discussing them is a way of improving future research and reflecting on future perspectives.

9.3.1. High dimensionality and imbalanced datasets

In data science, high dimensionality datasets refer to datasets that have many variables, or cofactors, relative to the number of samples. Imbalanced datasets refer to the

existence of some variables that are represented by many samples, while some target variables are represented by very few samples.

The EMOSE dataset (Chapter 3) was equilibrated, because it attempted to obtain, at least, three replicates of each variable. Even though some replicates were lost either during sampling processing or during bioinformatics quality filtering, differences in sampling effort were minimal. However, this dataset included many variables, but few samples represented each variable, which resulted in a high dimensionality data. Consequently, Chapter 3 posed a challenge from a statistical perspective. The solution was to apply consistent statistics and focus interpretation on the statistical tests with stronger support. Indeed, the major conclusions of Chapter 3 were relative to variables well represented (*e.g.*, filtered volume).

The N-ICE dataset suffered from low sample size, being composed of 18 original samples, nine of which were further multiplied by two bioinformatics approaches, resulting in a total of 27 samples (Chapter 5). However, this sample size was enough for the original purpose of Chapter 5, which focused on the comparison of methodological variations in the interpretation of the prokaryotic rare biosphere and focused on a few environmental variables. The small size of the N-ICE was an advantage for Chapters 6 and 7 in testing and developing *ulrb*, because a small dataset is easy to handle during software development, and it is easy to explain and share in learning materials.

The MOSJ 2016-2020 dataset did not suffer from low sample size, however, it did include many environmental variables, to have a multi-year comprehensive monitoring of microbiological communities. An effort which extends beyond this thesis. Consequently, it is high dimensional and imbalanced, because some variables are easier to sample than others (consider the effort of collecting samples at 5 m *vs* 1000 m of depth). Such limitations of the dataset were not a problem for Chapter 6, where this dataset was used for testing *ulrb*, but they were a factor that we needed to account for in Chapter 8. This was because some variables were not well represented, resulting in weaker statistical support.

An important insight from MOSJ 2016-2020 is that the effort to build big datasets is not linear, because as the sampling grows, some variables get very well represented, but others become underrepresented. However, such obstacles derive from the nature of the data itself, beyond funding, human resources and timing issues. For instance, it is understandable that the number of samples from the Arctic Ocean during the summer season grows quicker than data obtained during the winter; or that the number of

samples collected at the epipelagic layer grows much faster than that of samples collected in the bathypelagic layer. It is also worth noting that imbalanced and high-dimensional data might be unavoidable in microbial ecology, because the number of cofactors that can be considered at once is very high. For example, to discard any confounding effect on the effect of temperature on diversity using the MOSJ 2016-2020 dataset (Chapter 8), taking a classical statistical approach, we must consider the impact of depth, pelagic layer, water mass, and geography, but by such division of factors the true sampling effort becomes quite small and the statistical models too complex.

Collectively, as stated in the General Introduction section, we attempted to use centrality metrics and visualization practices consistent with sampling size. However, particularly in unbalanced datasets, such principles were hard to follow straightforwardly, and such considerations were adapted and evaluated in a case-by-case scenario. Notwithstanding, our guiding principle was to show data clearly and interpret it accounting for its limitations.

Finally, several data science approaches can deal with high dimensional and imbalanced data better than the classical statistics approach, which is part of the reason why incorporating rule-based machine learning in this thesis proved to be so useful.

9.3.2. Phylogenetic signal, community assembly theory, and classification of the types of rarity

One of the missing points in the thesis was the connection between community assembly theory and the rare biosphere in the Arctic Ocean, as expected from the ecological framework proposed in Chapter 2. Community assembly theory is useful to distinguish stochastic and deterministic effects and to further subdivide them in specific ecological processes (Jia, Dini-Andreote and Falcão Salles, 2018). However, the statistical approach used to calculate such processes requires the calculation of the phylogenetic signal. Indeed, obtaining phylogenetic trees could be advantageous not only to calculate the phylogenetic signal and then estimate community assembly processes, but also to calculate phylogenetic diversity, and thus provide a more comprehensive ecological interpretation of the microbial rare biosphere.

During the thesis experimental work, we attempted to produce phylogenetic trees using the data from MOSJ 2016-2020. Such a task proved challenging, because we were not able to obtain phylogenetic trees with sufficient statistical support (data not shown), using either maximum likelihood or Bayesian methods. This was expected from using short

regions of the 16S rRNA gene. Therefore, we did not endeavor to estimate certain metrics that would ultimately be compromised by an ineffective phylogenetic analysis. We note that some studies have used such metrics with data from short regions of the 16S rRNA gene but providing little to no information on the statistical support of the corresponding phylogenetic trees, *e.g.*, (Ren, Luo and Zhang, 2022).

The application of such metrics of community assembly theory could help discriminate the types of rarity (Jia, Dini-Andreote and Falcão Salles, 2018). However, such an approach is deductive, *i.e.,* considers general ecological principles (like stochasticity) and then specific ecological processes (like drift) and associates such processes with a range of values from certain statistics. For example, the values of βNTI (β Nearest Taxon Index) indicate stochastic or deterministic mechanisms, and further calculations can distinguish specific stochastic and deterministic processes (Jia, Dini-Andreote and Falcão Salles, 2018).

Even though we initially considered this approach (Chapter 2) and were intending to use it, the challenges with the definition of different types of rarity (Chapter 5) hinted that a deductive approach might be naïve. Instead, we opted for an inductive approach to define types of rarity, whereby we monitor the abundance classifications obtained by specific taxa across a range of variables (Chapters 5 and 6). This approach is independent of the definition of rarity used, as we applied a similar paradigm to both threshold-based definitions, MultiCoLA and *ulrb*. However, our approach becomes extremely complicated for complex datasets and is non-trivial to apply for general purpose datasets. Instead, it must be used in a case-by-case scenario. Even if we attempted to use community assembly theory to deductively predict types of rarity, we consider that we did not have the data to accurately use it and we consider that an inductive approach is more useful, because it might identify types of rarity that do not follow the community assembly theory assumptions. However, community assembly theory is a framework that is very useful for ecological interpretation of data (see section 9.4).

9.3.3. Fast-paced sequencing advances

Based on Chapter 4 and the literature (Klemetsen, Willassen and Karlsen, 2019; Pootakham *et al.*, 2019, 2021; Brede *et al.*, 2020; Wang *et al.*, 2022; Yu *et al.*, 2022), we know that the best approach to survey prokaryotic communities would be using full-length 16S rRNA gene amplicon sequencing, because there are no longer significant limits on high throughput and the sequencing quality is better. Therefore, it is an

approach that both provides a good number of reads and high-quality information. However, for the purpose of building long term datasets that aim to uncover temporal patterns, which is the case with MOSJ 2016-2020 dataset, it is important to balance technological advance with standardization. This is the reason why the final dataset relied on V4V5 16S rRNA gene sequencing instead of the full gene, even though we knew that the full gene would provide more accurate information. On the same note, the GTDB taxonomy database also has several advantages compared with the Silva database, and together with full-length 16S rRNA gene sequencing provides the highest degree of information. We showed this in Chapter 4, where this combination (full-length 16S rRNA gene and GTDB) was able to classify approximately 75% of all reads down to species level. The database selection is also particularly important in the context of both the rare biosphere and the Arctic Ocean, because we require a general-purpose database that includes as much information as possible, since we cannot rely on habitat-specific databases. Therefore, it would be expected that we would use the GTDB database in Chapter 8, for the MOSJ 2016-2020 dataset. However, we opted for the Silva database, because it was better at distinguishing between chloroplasts and *Cyanobacteria* than GTDB (data not shown). An alternative that we did not consider at the time of writing Chapter 8 was the utilization of GTDB but using Silva to filter out potential chloroplasts and verify against the NCBI database. However, that approach would require several additional steps.

Metagenomic based methods were also greatly improved, being able to extract MAGs, which undoubtedly provides more information than metabarcoding methods. In general, even though we used metagenomes, we did not cover functional aspects of the rare biosphere and instead focused mostly on questions regarding diversity and taxonomy. However, this was a consequence of a bigger effort and focus on methodological issues of microbiome sampling, creation of standardized data and improvements of data analysis.

Finally, the data analysis advances of Parts III and IV are independent of the sequencing strategy employed, and database selection, ensuring their relevance for the evolving field of microbial ecology.

9.4. Open questions and future perspectives

9.4.1. Continued methodological improvements

It is challenging to achieve a fully standardized workflow, while keeping up with updates in technology. One such challenge is the number of possible variations. In this regard, the EMOSE dataset (Chapter 3) holds a significant potential to further elucidate finer methodological differences, which might have been missed in the general analyses performed in Chapter 3. Specifically, we assumed that methodological differences would have a linear impact on the whole community *vs* its subdivisions in abundance classifications. However, with the *ulrb* method, we might test if that is the case, or if the rare biosphere is disproportionately impacted by methodological variations. This is a credible hypothesis, considering that by subdividing the community by taxonomic groups, we obtained evidence that specific taxa could be differently impacted by methodological variations (Chapter 3). However, we were unable to understand why, and we were unable to provide a detailed analysis of that phenomenon, due to data analysis limitations. In future work we could use strategies like rule-based machine learning to identify taxa that are missed by specific methodological combinations, or vice-versa. We have reasons to believe that rule-based machine learning can solve this data analysis problem based on the results obtained in Chapter 8. Outside the scope of the EMOSE dataset, further methodological variations could be explored, such as DNA extraction methods, in future studies.

9.4.2. Data science potential to boost microbial ecology studies

Several of the limitations presented in the previous section (9.3) can be solved or, at least, minimized by data science approaches. Indeed, the meaning of “low abundance” was hard to establish methodologically, but we showed that unsupervised machine learning (with *ulrb*) was able to circumvent the need to define “low abundance” (Chapter 6). This was achieved by clustering taxa of similar abundance together and grouping them in abundance classifications.

In future versions of the *ulrb* R package, we aim at developing an algorithm to inductively establish the types of rarity in any experimental setting. This is quite challenging, because we must consider the possible abundance classifications (k + 1, *i.e.*, the abundance classifications plus the possible absence), for any number of taxa (*m*), across any number of samples (*n*). If the ordering of the combinations matters and replacement is allowed, then the combinations possible, per taxon, are given by permutations with

replacement, $(k+1)^n$. It follows that, for all taxa, the total combinations in the dataset should be given by,

$$\sum_{1}^{m}(k+1)^n$$

Equation 9.1. Number of possible combinations of abundance classifications.

The combinations would then be grouped into types of rarity and would be summarized in relevant metrics and connected with ecological principles, like community assembly theory (Jia, Dini-Andreote and Falcão Salles, 2018). This approach has potential to obtain deeper insights into microbial dynamics overall, because it can be connected to specific ecological mechanisms. Since the number of categorical combinations is so large, data science tools, including machine learning, will be key for its interpretation. For perspective, applying the above-mentioned formula to just 10 samples with 100 distinct taxa, using the default *ulrb* parameters (k = 3), would result in 1 048 576 possible combinations of abundance classifications for each taxon.

The last-mentioned example also showcases another important insight from this thesis, which is that even small microbial ecology datasets are very information dense, meaning that they contain a lot of information relative to the number of samples (this is another way of looking at high dimensionality data). Therefore, even with a small sample size, to optimize the cost-effectiveness of investment made in ecological monitoring, and to avoid wasting high quality data, data science tools are important. This notion of information density might become even more relevant for metagenomic based methods, whereby the functional information available is very complex.

As we showed in Chapter 8, rule-based machine learning is an approach that can process complex and noisy datasets to obtain informative patterns. Furthermore, it holds the possibility of merging data from different sources. We showed this by combining abundance table data, taxonomic data and sample metadata together in the rule-based machine learning approach (Chapter 8). Therefore, it is possible, in principle, to add further information, such as metagenomic data, into larger, multilayered datasets. This approach would hold the promise to simultaneously account for taxonomy, the environmental context, and the actual functions. In the long term, we could possibly expand this notion to concomitantly integrate multiple *omics* levels. For instance, combining in a single rule-based machine learning model the environmental context, the

prokaryotic composition, but also the gene, transcript, protein, and metabolite level. Such an endeavor would allow for a very comprehensive analysis of a given biological system.

Besides rule-based and unsupervised machine learning, other machine learning families of algorithms should be tested in the microbial ecology field, because they might solve new problems. And, within the approaches of machine learning already used, there might also be other algorithmic alternatives that can be considered in future work. For instance, it is worth exploring different machine learning strategies to deal with both high-dimensional and imbalanced datasets.

Generally, future work should both improve the analysis provided by the classical statistics framework to confirm specific hypotheses and actively search for alternative data analyses.

Finally, in the short to medium term, we can predict the implementation of rule-based machine learning in a dedicated R package, with the specific purpose of optimizing taxonomic data analyses. Such improvement is very relevant, because the interpretation of taxonomic data is hampered by current microbial ecology data analyses strategies. Specifically, the taxonomic data problem can be described by the large number of possible categories, which are then grouped in several taxonomic levels, across multiple samples, grouped by multiple environmental variables. Therefore, it is a similar problem to that of defining types of rarity, *i.e.,* the existence of an unpredictably large number of combinations. Even though this was not an explicit objective of the thesis, while applying rule-based machine learning to explain the abundance classifications, the results revealed the potential to solve the problem of taxonomy analysis. Essentially, rule-based machine learning can take an unpredictably large number of combinations and automatically identify informative, non-redundant patterns (following the methodology proposed in Chapter 8). By backtracking, *i.e.,* by searching for a taxonomic group of interest in the rule-based machine learning output, we were also able to obtain detailed information, which can then be explored with additional analyses, including traditional statistical testing, for hypothesis confirmation. Therefore, there is foreseeable application in the short-term, even though we will need to adapt our algorithm to specific purpose of explaining taxonomy and then build an R package for easy access to our peers.

9.4.3. The Arctic Ocean as a laboratory for accessing climate change impacts

Climate warming is one of the main concerns in the Arctic, which justifies the need to address how the microbial rare biosphere will respond to it. Our thesis contributed to this effort, but currently available datasets and data analysis are not enough to identify the specific impact of climate change in the Arctic Ocean. Such an endeavor will be possible in the future, by adding new data and applying more complex data analyses. For instance, it was quite difficult to distinguish time from local patterns in Chapter 8, making it difficult to establish a relationship with climate change. For that purpose, future work will need to add new years of data and apply detailed analysis of cofactors. Besides the effect of temperature, depth and water mass, it is also important to distinguish the geographical from the temporal effects. In that sense, geography should also be accounted for in future studies. As we exposed in the previous sections, this poses a data analyses challenge that should account for both classical statistics and machine learning. In the short- to mid-term, new data will be readily available, providing the possibility to update and improve the results obtained in Chapter 8, by increasing the five-year range to almost a full decade. Indeed, sequencing results are already available to expand MOSJ from 2016-2020 to 2016-2024, and future expeditions have been planned. In the long-term, as we obtain more data, time series analyses may be used to investigate the yearly patterns found and distinguish naturally occurring cycles, from specific reactions to climate warming.

9.4.4. The microbial rare biosphere in the microbial ecology landscape

Microbial ecology studies can benefit from the abundance classifications provided by *ulrb*, whereby they identify the rare biosphere microorganisms, but they also identify which microorganisms are abundant or undetermined. This approach can be extended in future studies to consider abundance classifications tailored for specific experimental designs and is a powerful tool to identify abundance patterns in microorganisms, regardless of methodology used. Even though we generally focused on prokaryotes in this thesis, our approach and contributions should be applicable to protists and fungi.

In future studies, we also propose not to look exclusively at the rare biosphere but always consider the whole microbial community context. That is, we argue that the rare

biosphere should be analyzed as one of many components of the entire community and should always be compared against abundant taxa.

It is worth noting that the rare biosphere is usually the component most affected by sequencing noise and potential artifacts, because it always represents the subset of the data with less representatives (for example, rare ASVs are represented by fewer reads than abundant ASVs). Future work should attempt to better understand to what extent it is possible to reduce such noise. Such steps should be applied at the bioinformatics level (raw read processing and sequencing quality assessment), *i.e.*, before *ulrb*. Once the original data obtains better quality, then the *ulrb* output will also reflect that higher quality. One approach to achieve such objective would be to directly calculate sequencing biases, using mock communities as references (McLaren, Willis and Callahan, 2019). Even though using mock communities for the rare biosphere is challenging, it is one of the future questions that might be answered by the EMOSE dataset, for example.

Additionally, *ulrb* holds the potential to explore how many abundance classifications a given microbial community has, by accessing the value of k obtained by the automatic settings. This sort of analysis could provide relevant information about microbial communities, whereby microbial communities with fewer classifications (*e.g.*, k = 2) could be considered more extreme than microbial communities with several different k values (*e.g.*, k = 4). For example, in future research we could test the hypothesis that extreme environments tend to have fewer abundance classifications compared to temperate environments. This hypothesis, therefore, would assume that under extreme conditions taxa are either very abundant (reflecting adaptation to the environment), or subsist as rare; while under temperate environments we would expect to find taxa that are very rare, but also several intermediate groups, as well as the most abundant ones. Finally, comparison of k values could be inspected in future studies to evaluate a possible relationship between the ideal k of a given community and its higher/lower resistance to environmental disturbance.

Finally, future work should explore the functional capabilities of the microbial rare biosphere by integrating metagenomic and metatranscriptomic data. Additionally, we may expand the notion of rare taxa to rare genes, whereby we hypothesize that common genes are abundant and are responsible for housekeeping functions, while genes responsible for specific functions (such as nitrogen cycling) are rare. Indeed, some of the advances made in this thesis will be helpful in the clarification of the functional role of the rare biosphere in the environment. For example, by combining *ulrb* and rule-based machine learning (similar to the approach used in Chapter 8) it might be possible to

connect complex combinations of functional genes and rare taxa in their environmental context.

9.5. Final remarks

This thesis demonstrated how advancements in microbial ecology methodologies can enhance the study of the rare biosphere in natural environments; and showed that accounting for the microbial rare biosphere in standard microbial ecology studies is important. In fact, we provided both an ecological framework to do so, as well as a concrete methodology to make the study of the microbial rare biosphere accessible to any laboratory studying microbial communities, in any ecological context. Finally, we showed that machine learning holds promise in the analysis of microbial communities, by providing novel insights into microbial rare biosphere ecology in the Arctic Ocean.

References

Aalto, N.J. *et al.* (2022) 'Diversity and selection of surface marine microbiomes in the Atlantic-influenced Arctic', *Frontiers in Microbiology*, 13(892634), pp. 1–18. Available at: https://doi.org/10.3389/fmicb.2022.892634.

Agrawal, R., Imielinski, T. and Swami, A. (1993) 'Mining association in large databases', *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*, pp. 207–216.

Agrawal, R. and Srikant, R. (1994) 'A fast algorithm for mining association rules', in *Proceedings of the 20th international conference on very large data bases*. Santiago de Chile, 12-15 September.

Ai, D. *et al.* (2013) 'Migration and niche partitioning simultaneously increase species richness and rarity', *Ecological Modelling*, 258(6), pp. 33–39. Available at: https://doi.org/10.1016/j.ecolmodel.2013.03.001.

Alberti, A. *et al.* (2017) 'Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition', *Scientific Data*, 4(1), p. 170093. Available at: https://doi.org/10.1038/sdata.2017.93.

Albertsen, M. *et al.* (2013) 'Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes', *Nature Biotechnology*, 31(6), pp. 533–538. Available at: https://doi.org/10.1038/nbt.2579.

Alivisatos, A.P. *et al.* (2015) 'A unified initiative to harness Earth's microbiomes', *Science*, 350(6260). Available at: https://doi.org/10.1126/science.aac8480.

Alonso-Sáez, L. *et al.* (2014) 'Winter bloom of a rare betaproteobacterium in the Arctic Ocean', *Frontiers in Microbiology*, 5(425), pp. 1–9. Available at: https://doi.org/10.3389/fmicb.2014.00425.

Alonso-Sáez, L., Díaz-Pérez, L. and Morán, X.A.G. (2015) 'The hidden seasonality of the rare biosphere in coastal marine bacterioplankton', *Environmental Microbiology*, 17(10), pp. 3766–3780. Available at: https://doi.org/10.1111/1462-2920.12801.

Alteio, L. V. *et al.* (2020) 'Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil', *mSystems*, 5(2), pp. 1–18. Available at: https://doi.org/10.1128/mSystems.00768-19.

Amaral-Zettler, L. *et al.* (2010) 'A global census of marine microbes', in *Life in the World's Oceans*. Oxford, UK: Wiley-Blackwell, pp. 221–245. Available at: https://doi.org/10.1002/9781444325508.ch12.

Amaral-Zettler, L.A. *et al.* (2009) 'A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes', *PLoS ONE*, 4(7), pp. 1–9. Available at: https://doi.org/10.1371/journal.pone.0006372.

De Anda, V. *et al.* (2018) 'Understanding the mechanisms behind the response to environmental perturbation in microbial mats: A metagenomic-network based approach', *Frontiers in Microbiology*, 9(11), pp. 1–24. Available at: https://doi.org/10.3389/fmicb.2018.02606.

Anderson, R.E., Sogin, M.L. and Baross, J.A. (2015) 'Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents', *FEMS Microbiology Ecology*, 91(1), pp. 1–11. Available at: https://doi.org/10.1093/femsec/fiu016.

Apprill, A. *et al.* (2015) 'Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton', *Aquatic Microbial Ecology*, 75(2), pp. 129–137. Available at: https://doi.org/10.3354/ame01753.

Araujo, A.S.F. *et al.* (2023) 'Changes in the bacterial rare biosphere after permanent application of composted tannery sludge in a tropical soil', *Chemosphere*, 313(11). Available at: https://doi.org/10.1016/j.chemosphere.2022.137487.

Arnan, X., Gaucherel, C. and Andersen, A.N. (2011) 'Dominance and species co-occurrence in highly diverse ant communities: A test of the interstitial hypothesis and discovery of a three-tiered competition cascade', *Oecologia*, 166(3), pp. 783–794. Available at: https://doi.org/10.1007/s00442-011-1919-y.

Assmy, P. *et al.* (2023) 'Seasonal plankton dynamics in Kongsfjorden during two years of contrasting environmental conditions', *Progress in Oceanography*, 213(12), p. 102996. Available at: https://doi.org/10.1016/j.pocean.2023.102996.

Auguie, B. (2017) 'gridExtra: Miscellaneous functions for "grid" graphics'. R package v2.3. Available at: https://cran.r-project.org/web/packages/gridExtra/index.html.

Auld, R.R. *et al.* (2013) 'Characterization of the microbial acid mine drainage microbial community using culturing and direct sequencing techniques', *Journal of Microbiological Methods*, 93(2). Available at: https://doi.org/10.1016/j.mimet.2013.01.023.

Bachy, C. *et al.* (2013) 'Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study', *The ISME Journal*, 7(2), pp. 244–255. Available at: https://doi.org/10.1038/ismej.2012.106.

Bahrndorff, S. *et al.* (2016) 'The microbiome of animals: Implications for conservation biology', *International Journal of Genomics*, 2016, pp. 1–7. Available at: https://doi.org/10.1155/2016/5304028.

Balbuena, J.A. *et al.* (2021) 'Fuzzy quantification of common and rare species in ecological communities (FuzzyQ)', *Methods in Ecology and Evolution*, 12(6), pp. 1070–1079. Available at: https://doi.org/10.1111/2041-210X.13588.

Baltar, F. *et al.* (2015) 'Response of rare, common and abundant bacterioplankton to anthropogenic perturbations in a Mediterranean coastal site', *FEMS Microbiology Ecology*, 91(6), pp. 1–12. Available at: https://doi.org/10.1093/femsec/fiv058.

Baltar, F. *et al.* (2016) 'Oceanic fronts: Transition zones for bacterioplankton community composition', *Environmental Microbiology Reports*, 8(1), pp. 132–138. Available at: https://doi.org/10.1111/1758-2229.12362.

Bastian, M., Heymann, S. and Jacomy, M. (2009) 'Gephi: An open source software for exploring and manipulating networks', in *International AAAI Conference on Weblogs and Social Media*, pp. 361–362.

Bauer, D.F. (1972) 'Constructing confidence sets using rank statistics', *Journal of the American Statistical Association*, 67(339), p. 687. Available at: https://doi.org/10.2307/2284469.

Bay, S.K. *et al.* (2020) 'Soil bacterial communities exhibit strong biogeographic patterns at fine taxonomic resolution', *mSystems*, 5(4), pp. 1–16. Available at: https://doi.org/10.1128/msystems.00540-20.

Bayardo, R.J., Agrawal, R. and Gunopulos, D. (1999) 'Constraint-based rule mining in large, dense databases', in *Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337)*, Sydney, NSW, Australia, pp. 188–197. Available at: https://doi.org/10.1109/ICDE.1999.754924.

Belheouane, M. *et al.* (2020) 'Assessing similarities and disparities in the skin microbiota between wild and laboratory populations of house mice', *ISME Journal*, 14(10). Available at: https://doi.org/10.1038/s41396-020-0690-7.

Bontemps, Z., Moënne-Loccoz, Y. and Hugoni, M. (2024) 'Stochastic and deterministic assembly processes of microbial communities in relation to natural attenuation of black stains in Lascaux Cave', *mSystems*, 9(2). Available at: https://doi.org/10.1128/msystems.01233-23.

Bosch, T.C.G. and McFall-Ngai, M. (2021) 'Animal development in the microbial world: Re-thinking the conceptual framework', in *Current Topics in Developmental Biology*. Academic Press Inc., pp. 399–427. Available at: https://doi.org/10.1016/bs.ctdb.2020.11.007.

Boström, K.H. *et al.* (2004) 'Optimization of DNA extraction for quantitative marine bacterioplankton community analysis', *Limnology and Oceanography: Methods*, 2(11), pp. 365–373. Available at: https://doi.org/https://doi.org/10.4319/lom.2004.2.365.

Boulotte, N.M. *et al.* (2016) 'Exploring the *Symbiodinium* rare biosphere provides evidence for symbiont switching in reef-building corals', *The ISME Journal*, 10(11), pp. 2693–2701. Available at: https://doi.org/10.1038/ismej.2016.54.

Bowen, J.L. *et al.* (2012) 'Salt marsh sediment diversity: a test of the variability of the rare biosphere among environmental replicates', *The ISME Journal*, 6(11), pp. 2014–2023. Available at: https://doi.org/10.1038/ismej.2012.47.

Bowman, J.S. *et al.* (2012) 'Microbial community structure of Arctic multiyear sea ice and surface seawater by 454 sequencing of the 16S RNA gene', *The ISME Journal*, 6(1), pp. 11–20. Available at: https://doi.org/10.1038/ismej.2011.76.

Bowman, J.S. *et al.* (2013) 'Selective occurrence of *Rhizobiales* in frost flowers on the surface of young sea ice near Barrow, Alaska and distribution in the polar marine rare biosphere', *Environmental Microbiology Reports*, 5(4), pp. 575-582. Available at: https://doi.org/10.1111/1758-2229.12047.

Box, J.E. *et al.* (2019) 'Key indicators of Arctic climate change: 1971–2017', *Environmental Research Letters*, 14(4), p. 045010. Available at: https://doi.org/10.1088/1748-9326/aafc1b.

Brede, M. *et al.* (2020) 'PacBio and Illumina MiSeq amplicon sequencing confirm full recovery of the bacterial community after subacute ruminal acidosis challenge in the

RUSITEC System', *Frontiers in Microbiology*, 11(1813), pp. 1-19. Available at: https://doi.org/10.3389/fmicb.2020.01813.

Brin, S. *et al.* (1997) 'Dynamic itemset counting and implication rules for market basket data', in *Proceedings of the 1997 ACM SIGMOD international conference on Management of data - SIGMOD '97*. New York, New York, USA: ACM Press, pp. 255–264. Available at: https://doi.org/10.1145/253260.253325.

Brown, C.T. *et al.* (2015) 'Unusual biology across a group comprising more than 15% of domain Bacteria', *Nature*, 523(7559), pp. 208–211. Available at: https://doi.org/10.1038/nature14486.

Calatayud, J. *et al.* (2019) 'Positive associations among rare species and their persistence in ecological assemblages'. *figshare*, Dataset. Available at: https://doi.org/10.6084/m9.figshare.9906092.v1.

Califano, G. *et al.* (2017) 'Molecular taxonomic profiling of bacterial communities in a gilthead seabream (*Sparus aurata*) hatchery', *Frontiers in Microbiology*, 8(204), pp. 1–16. Available at: https://doi.org/10.3389/fmicb.2017.00204.

Califano, G. *et al.* (2020) 'Cultivating the macroalgal holobiont: effects of integrated multi-trophic aquaculture on the microbiome of *Ulva rigida* (*Chlorophyta*)', *Frontiers in Marine Science*, 7(2), pp. 1–19. Available at: https://doi.org/10.3389/fmars.2020.00052.

Calinski, T. and Harabasz, J. (1974) 'A dendrite method for cluster analysis', *Communications in Statistics - Theory and Methods*, 3(1), pp. 1–27. Available at: https://doi.org/10.1080/03610927408827101.

Callahan, B.J. *et al.* (2016) 'DADA2: High-resolution sample inference from Illumina amplicon data', *Nature Methods*, 13(7), pp. 581–583. Available at: https://doi.org/10.1038/nmeth.3869.

Callahan, B.J. *et al.* (2019) 'High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution', *Nucleic Acids Research*, 47(18), pp. e103–e103. Available at: https://doi.org/10.1093/nar/gkz569.

Callahan, B.J., McMurdie, P.J. and Holmes, S.P. (2017) 'Exact sequence variants should replace operational taxonomic units in marker-gene data analysis', *The ISME Journal*, 11(12), pp. 2639–2643. Available at: https://doi.org/10.1038/ismej.2017.119.

Campbell, B.J. *et al.* (2011) 'Activity of abundant and rare bacteria in a coastal ocean', *Proceedings of the National Academy of Sciences*, 108(31), pp. 12776–12781. Available at: https://doi.org/10.1073/pnas.1101405108.

Campos, A.B. *et al.* (2021) 'CPR and DPANN have an overlooked role in corals' Microbial Community Structure', *Microbial Ecology,* 83(3), 252-255. Available at: https://doi.org/10.1007/s00248-021-01737-4.

Caporaso, J.G. *et al.* (2011) 'Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample', *Proceedings of the National Academy of Sciences*, 108(supplement_1), pp. 4516–4522. Available at: https://doi.org/10.1073/pnas.1000080107.

Caporaso, J.G. *et al.* (2012) 'Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms', *The ISME Journal*, 6(8), pp. 1621–1624. Available at: https://doi.org/10.1038/ismej.2012.8.

Cardozo-Mino, M.G. *et al.* (2021) 'Spatial distribution of Arctic bacterioplankton abundance is linked to distinct water masses and summertime phytoplankton bloom dynamics (Fram Strait, 79°N)', *Frontiers in Microbiology*, 12(658803), pp. 1–14. Available at: https://doi.org/10.3389/fmicb.2021.658803.

Caron, D. and Countway, P. (2009) 'Hypotheses on the role of the protistan rare biosphere in a changing world', *Aquatic Microbial Ecology*, 57(3), pp. 227–238. Available at: https://doi.org/10.3354/ame01352.

Carter-Gates, M. *et al.* (2020) 'Implications of increasing Atlantic influence for Arctic microbial community structure', *Scientific Reports*, 10(1), pp. 1–13. Available at: https://doi.org/10.1038/s41598-020-76293-x.

Catão C. P., E. *et al.* (2021) 'Temperate and tropical coastal waters share relatively similar microbial biofilm communities while free-living or particle-attached communities are distinct', *Molecular Ecology*, 30(12), pp. 2891–2904. Available at: https://doi.org/10.1111/mec.15929.

Cerro-Gálvez, E. *et al.* (2019) 'Microbial responses to anthropogenic dissolved organic carbon in the Arctic and Antarctic coastal seawaters', *Environmental Microbiology*, 21(4), pp. 1466–1481. Available at: https://doi.org/10.1111/1462-2920.14580.

Chase, J.M. (2010) 'Stochastic community assembly causes higher biodiversity in more productive environments', *Science*, 328(5984), pp. 1388–1391. Available at: https://doi.org/10.1126/science.1187820.

Chase, J.M. and Myers, J.A. (2011) 'Disentangling the importance of ecological niches from stochastic processes across scales', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1576), pp. 2351–2363. Available at: https://doi.org/10.1098/rstb.2011.0063.

Chaumeil, P.A. *et al.* (2020) 'GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database', *Bioinformatics*, 36(6), pp. 1925–1927. Available at: https://doi.org/10.1093/bioinformatics/btz848.

Chaumeil, P.A. *et al.* (2022) 'GTDB-Tk v2: memory friendly classification with the genome taxonomy database', *Bioinformatics (Oxford, England)*, 38(23), pp. 5315–5316. Available at: https://doi.org/10.1093/bioinformatics/btac672.

Chave, J. (2004) 'Neutral theory and community ecology', *Ecology Letters*, 7(3), pp. 241–253. Available at: https://doi.org/10.1111/j.1461-0248.2003.00566.x.

Chesson, P. (2000) 'Mechanisms of maintenance of species diversity', *Annual Review of Ecology and Systematics*, 31(1), pp. 343–366. Available at: https://doi.org/10.1146/annurev.ecolsys.31.1.343.

Chiriac, C.M. *et al.* (2018) 'Microbial composition and diversity patterns in deep hyperthermal aquifers from the western plain of Romania', *Microbial Ecology*, 75(1), pp. 38–51. Available at: https://doi.org/10.1007/s00248-017-1031-x.

Choi, J. and Park, J.S. (2020) 'Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform', *Scientific Reports*, 10(1), p. 6519. Available at: https://doi.org/10.1038/s41598-020-63561-z.

Coelho, L.F. *et al.* (2022) 'Structural shifts in sea ice prokaryotic communities across a salinity gradient in the subarctic', *Science of the Total Environment*, 827(154286), pp. 1-15. Available at: https://doi.org/10.1016/j.scitotenv.2022.154286.

Comeau, A.M. *et al.* (2011) 'Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum', *PLoS ONE*. Edited by J.A. Gilbert, 6(11), p. e27492. Available at: https://doi.org/10.1371/journal.pone.0027492.

Condit, R. *et al.* (2002) 'Beta-diversity in tropical forest trees', *Science*, 295(5555), pp. 666–669. Available at: https://doi.org/10.1126/science.1066854.

Condit, R. *et al.* (2019) 'Complete data from the Barro Colorado 50-ha plot: 423617 trees, 35 years'. *Dryad*, Dataset. Available at: https://doi.org/10.15146/5xcp-0d46.

Corbet, A.S. (1942) 'The distribution of butterflies in the Malay peninsula', *Proceedings of the Royal Entomological Society of London*, A(16), pp. 101–116.

Costa, L.V. da *et al.* (2022) 'MALDI-TOF MS database expansion for identification of Bacillus and related genera isolated from a pharmaceutical facility', *Journal of Microbiological Methods*, 203(10), p. 106625. Available at: https://doi.org/10.1016/j.mimet.2022.106625.

Cottier, F. *et al.* (2005) 'Water mass modification in an Arctic fjord through cross-shelf exchange: The seasonal hydrography of Kongsfjorden, Svalbard', *Journal of Geophysical Research: Oceans*, 110(C12), pp. 1–18. Available at: https://doi.org/10.1029/2004JC002757.

Coveley, S., Elshahed, M.S. and Youssef, N.H. (2015) 'Response of the rare biosphere to environmental stressors in a highly diverse ecosystem (Zodletone spring, OK, USA)', *PeerJ*, 3(e1182), pp. 1-19. Available at: https://doi.org/10.7717/peerj.1182.

Cramm, M.A. *et al.* (2021) 'Characterization of marine microbial communities around an Arctic seabed hydrocarbon seep at Scott Inlet, Baffin Bay', *Science of the Total Environment*, 762, p. 143961. Available at: https://doi.org/10.1016/j.scitotenv.2020.143961.

Crespo, B.G. *et al.* (2016) 'Probing the rare biosphere of the North-West Mediterranean Sea: An experiment with high sequencing effort', *PLOS ONE*, 11(7), p. e0159195. Available at: https://doi.org/10.1371/journal.pone.0159195.

Crevecoeur, S., Prairie, Y.T. and del Giorgio, P.A. (2022) 'Tracking the upstream history of aquatic microbes in a boreal lake yields new insights on microbial community assembly', *PNAS Nexus*, 1(4), pp. 1–12. Available at: https://doi.org/10.1093/pnasnexus/pgac171.

Cui, W. *et al.* (2023) 'Weak environmental adaptation of rare phylotypes sustaining soil multi-element cycles in response to decades-long fertilization', *Science of the Total Environment*, 871(2), p. 162063. Available at: https://doi.org/10.1016/j.scitotenv.2023.162063.

Dam, H.T. *et al.* (2020) 'Targeted Cell Sorting Combined With Single Cell Genomics Captures Low Abundant Microbial Dark Matter With Higher Sensitivity Than

Metagenomics', *Frontiers in Microbiology*, 11(7). Available at: https://doi.org/10.3389/fmicb.2020.01377.

Darwin, C. (1859) *The Origin of Species*. Amsterdam: Amsterdam University Press. Available at: https://doi.org/10.5117/9781904633785.

Davidov, K. *et al.* (2020) 'Identification of plastic-associated species in the Mediterranean Sea using DNA metabarcoding with Nanopore MinION', *Scientific Reports*, 10(1), p. 17533. Available at: https://doi.org/10.1038/s41598-020-74180-z.

Davies, D.L. and Bouldin, D.W. (1979) 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2). Available at: https://doi.org/10.1109/TPAMI.1979.4766909.

Dawson, W. *et al.* (2017) 'A small number of low-abundance bacteria dominate plant species-specific responses during rhizosphere colonization', *Frontiers in Microbiology*, 8(975), pp. 1–13. Available at: https://doi.org/10.3389/fmicb.2017.00975.

Debroas, D., Hugoni, M. and Domaizon, I. (2015) 'Evidence for an active rare biosphere within freshwater protists community', *Molecular Ecology*, 24(6), pp. 1236–1247. Available at: https://doi.org/10.1111/mec.13116.

Deiner, K. *et al.* (2018) 'Optimising the detection of marine taxonomic richness using environmental DNA metabarcoding: The effects of filter material, pore size and extraction method', *Metabarcoding and Metagenomics*, 2, pp. 1–15. Available at: https://doi.org/10.3897/mbmg.2.28963.

Delmont, T.O. *et al.* (2014) '*Phaeocystis antarctica* blooms strongly influence bacterial community structures in the Amundsen Sea polynya', *Frontiers in Microbiology*, 5(646), pp. 1-13. Available at: https://doi.org/10.3389/fmicb.2014.00646.

Delmont, T.O. *et al.* (2015) 'Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics', *Frontiers in Microbiology*, 6(358), pp. 1-15. Available at: https://doi.org/10.3389/fmicb.2015.00358.

Delpech, L.M. *et al.* (2021) 'Terrestrial inputs shape coastal bacterial and archaeal communities in a high Arctic fjord (Isfjorden, Svalbard)', *Frontiers in Microbiology*, 12(614634), pp. 1–19. Available at: https://doi.org/10.3389/fmicb.2021.614634.

Denonfoux, J. *et al.* (2013) 'Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration', *DNA Research*, 20(2), pp. 185–196. Available at: https://doi.org/10.1093/dnares/dst001.

Dethlefsen, L. *et al.* (2008) 'The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing', *PLoS Biology*, 6(11), p. e280. Available at: https://doi.org/10.1371/journal.pbio.0060280.

Dini-Andreote, F. *et al.* (2015) 'Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession', *Proceedings of the National Academy of Sciences*, 112(11), pp. E1326–E1332. Available at: https://doi.org/10.1073/pnas.1414261112.

Djurhuus, A. *et al.* (2017) 'Evaluation of filtration and DNA extraction methods for environmental DNA biodiversity assessments across multiple trophic levels', *Frontiers in Marine Science*, 4(314), pp. 1-11. Available at: https://doi.org/10.3389/fmars.2017.00314.

Doud, D.F.R. *et al.* (2020) 'Function-driven single-cell genomics uncovers cellulose-degrading bacteria from the rare biosphere', *The ISME Journal*, 14(3), pp. 659–675. Available at: https://doi.org/10.1038/s41396-019-0557-y.

Du, S. *et al.* (2020) 'Divergent co-occurrence patterns and assembly processes structure the abundant and rare bacterial communities in a salt marsh ecosystem', *Applied and Environmental Microbiology*, 86(13), pp. 1–12. Available at: https://doi.org/10.1128/AEM.00322-20.

Duarte, C.M. (2015) 'Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition', *Limnology and Oceanography Bulletin*, 24(1), pp. 11–14. Available at: https://doi.org/10.1002/lob.10008.

Duhaime, M.B. *et al.* (2012) 'Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: A rigorous assessment and optimization of the linker amplification method', *Environmental Microbiology*, 14(9), pp. 2526–2537. Available at: https://doi.org/10.1111/j.1462-2920.2012.02791.x.

Duncan, A. *et al.* (2022) 'Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans', *Microbiome*, 10(1), pp. 1–21. Available at: https://doi.org/10.1186/s40168-022-01254-7.

Dunn, O.J. (1964) 'Multiple comparisons using rank sums', *Technometrics*, 6(3), pp. 241–252. Available at: https://doi.org/10.1080/00401706.1964.10490181.

Dunnington, D. (2023) 'ggspatial: Spatial Data Framework for ggplot2', R package v1.1.9. Available at: https://doi.org/10.32614/CRAN.package.ggspatial.

Earl, J.P. *et al.* (2018) 'Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes', *Microbiome*, 6(1), p. 190. Available at: https://doi.org/10.1186/s40168-018-0569-2.

Edgar, R.C. (2018) 'Updating the 97% identity threshold for 16S ribosomal RNA OTUs', *Bioinformatics*, 34(14), pp. 2371–2375. Available at: https://doi.org/10.1093/bioinformatics/bty113.

Edwards, A. *et al.* (2020) 'Microbial genomics amidst the arctic crisis', *Microbial Genomics*, 6(5), pp. 1–20. Available at: https://doi.org/10.1099/mgen.0.000375.

Elshahed, M.S. *et al.* (2008) 'Novelty and uniqueness patterns of rare members of the soil biosphere', *Applied and Environmental Microbiology*, 74(17), pp. 5422–5428. Available at: https://doi.org/10.1128/AEM.00410-08.

Epstein, H.E. *et al.* (2021) 'Inconsistent patterns of microbial diversity and composition between highly similar sequencing protocols: a case study with reef-building corals', *Frontiers in Microbiology*, 12(740932), pp. 1-11. Available at: https://doi.org/10.3389/fmicb.2021.740932.

Eren AM *et al.* (2013) 'Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data', *Methods in Ecology and Evolution*, 4(12), pp. 1111–1119. Available at: https://doi.org/10.1111/2041-210x.12114.

Erwin, P.M. *et al.* (2014) 'Down under the tunic: bacterial biodiversity hotspots and widespread ammonia-oxidizing archaea in coral reef ascidians', *ISME Journal*, 8(3), pp. 575-588. Available at: https://doi.org/10.1038/ismej.2013.188.

Espinosa-Asuar, L. *et al.* (2022) 'Diversity of an uncommon elastic hypersaline microbial mat along a small-scale transect', *PeerJ*, 10(e13579), pp. 1-30. Available at: https://doi.org/10.7717/peerj.13579.

Esquivel-Hernández, D.A. *et al.* (2021) 'Microbial ecology in selenate-reducing biofilm communities: rare biosphere and their interactions with abundant phylotypes', *Biotechnology and Bioengineering*, 118(7), pp. 2460–2471. Available at: https://doi.org/10.1002/bit.27754.

F Escapa, I. *et al.* (2020) 'Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets', *Microbiome*, 8(1), p. 65. Available at: https://doi.org/10.1186/s40168-020-00841-w.

Fadeev, E. *et al.* (2018) 'Microbial communities in the East and West Fram Strait during sea ice melting season', *Frontiers in Marine Science*, 5(429), pp. 1–21. Available at: https://doi.org/10.3389/fmars.2018.00429.

Fadeev, E. *et al.* (2021) 'Comparison of two 16S rRNA primers (V3–V4 and V4–V5) for studies of Arctic microbial communities', *Frontiers in Microbiology*, 12(637526), pp. 1–11. Available at: https://doi.org/10.3389/fmicb.2021.637526.

Fargione, J., Brown, C.S. and Tilman, D. (2003) 'Community assembly and invasion: an experimental test of neutral versus niche processes', *Proceedings of the National Academy of Sciences of the United States of America*, 100(15), pp. 8916–8920. Available at: https://doi.org/10.1073/pnas.1033107100.

Fernández-Gómez, B. *et al.* (2019) 'Bacterial community structure in a sympagic habitat expanding with global warming: brackish ice brine at 85–90 °N', *The ISME Journal*, 13(2), pp. 316–333. Available at: https://doi.org/10.1038/s41396-018-0268-9.

Fernandez-Gonzalez, N., Huber, J.A. and Vallino, J.J. (2016) 'Microbial communities are well adapted to disturbances in energy input', *mSystems*, 1(5), pp. 1–15. Available at: https://doi.org/10.1128/mSystems.00117-16.

Fidler, G. *et al.* (2020) 'Tendentious effects of automated and manual metagenomic DNA purification protocols on broiler gut microbiome taxonomic profiling', *Scientific Reports*, 10(3419), pp. 1-16. Available at: https://doi.org/10.1038/s41598-020-60304-y.

Fisher, M.M. and Triplett, E.W. (1999) 'Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities', *Applied and Environmental Microbiology*, 65(10), pp. 4630–4636. Available at: https://doi.org/10.1074/jbc.M311300200.

Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943) 'The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population', *The Journal of Animal Ecology*, 12(1), pp. 42-58. Available at: https://doi.org/10.2307/1411.

Fuentes, S. *et al.* (2016) 'From rare to dominant: a fine-tuned soil bacterial bloom during petroleum hydrocarbon bioremediation', *Applied and Environmental Microbiology*, 82(3), pp. 888–896. Available at: https://doi.org/10.1128/AEM.02625-15.

Gagic, D. *et al.* (2015) 'Improving the genetic representation of rare taxa within complex microbial communities using DNA normalization methods', *Molecular Ecology Resources*, 15(3), pp. 464–476. Available at: https://doi.org/10.1111/1755-0998.12321.

Gaidos, E., Rusch, A. and Ilardo, M. (2011) 'Ribosomal tag pyrosequencing of DNA and RNA from benthic coral reef microbiota: community spatial structure, rare members and nitrogen-cycling guilds', *Environmental Microbiology*, 13(5), pp. 1138–1152. Available at: https://doi.org/10.1111/j.1462-2920.2010.02392.x.

Galand, P.E. *et al.* (2009) 'Ecology of the rare microbial biosphere of the Arctic Ocean', *Proceedings of the National Academy of Sciences*, 106(52), pp. 22427–22432. Available at: https://doi.org/10.1073/pnas.0908284106.

Galand, P.E. *et al.* (2010) 'Hydrography shapes bacterial biogeography of the deep Arctic Ocean', *The ISME Journal*, 4(4), pp. 564–576. Available at: https://doi.org/10.1038/ismej.2009.134.

Ganesan, S.M. *et al.* (2017) 'A tale of two risks: smoking, diabetes and the subgingival microbiome', *ISME Journal*, 11(9), pp. 2075–2089. Available at: https://doi.org/10.1038/ismej.2017.73.

García-Martín, E.E., Serret, P. and Leakey, R.J.G. (2014) 'Plankton community and bacterial metabolism in Arctic sea ice leads during summer 2010', *Deep-Sea Research Part I: Oceanographic Research Papers*, 92, pp. 152–161. Available at: https://doi.org/10.1016/j.dsr.2014.06.007.

Gaston, K. and Fuller, R. (2008) 'Commonness, population depletion and conservation biology', *Trends in Ecology & Evolution*, 23(1), pp. 14–19. Available at: https://doi.org/10.1016/j.tree.2007.11.001.

Gattoni, G. *et al.* (2023) 'Genomic study and lipidomic bioassay of *Leeuwenhoekiella parthenopeia*: a novel rare biosphere marine bacterium that inhibits tumor cell viability', *Frontiers in Microbiology*, 13(1090197), pp. 1–22. Available at: https://doi.org/10.3389/fmicb.2022.1090197.

Gevers, D. *et al.* (2012) 'The human microbiome project: a community resource for the healthy human microbiome', *PLoS Biology*, 10(8), p. e1001377. Available at: https://doi.org/10.1371/journal.pbio.1001377.

Ghiglione, J.-F. *et al.* (2012) 'Pole-to-pole biogeography of surface and deep marine bacterial communities.', *Proceedings of the National Academy of Sciences of the United*

States of America, 109(43), pp. 17633–8. Available at: https://doi.org/10.1073/pnas.1208160109.

Giebel, H.A. *et al.* (2011) 'Distribution of *Roseobacter* RCA and *SAR11* lineages in the North Sea and characteristics of an abundant RCA isolate', *ISME Journal*, 5(1), pp. 8–19. Available at: https://doi.org/10.1038/ismej.2010.87.

Gilbert, J.A. *et al.* (2010) 'The Earth Microbiome Project: meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010.', *Standards in Genomic Sciences*, 3(3), pp. 249–253. Available at: https://doi.org/10.4056/aigs.1443528.

Gilbert, J.A., Jansson, J.K. and Knight, R. (2014) 'The Earth Microbiome project: successes and aspirations', *BMC Biology*, 12(69), pp. 1-4. Available at: https://doi.org/10.1186/s12915-014-0069-1.

Gilbert, J.A., Jansson, J.K. and Knight, R. (2018) 'Earth Microbiome Project and global systems biology', *mSystems*, 3(3), pp. e00217-17. Available at: https://doi.org/10.1128/mSystems.00217-17.

Giovannoni, S.J. *et al.* (1990) 'Genetic diversity in Sargasso Sea bacterioplankton', *Nature*, 345, pp. 60–62. Available at: https://doi.org/10.1016/0021-9797(80)90501-9.

Giovannoni, S.J. (2017) 'SAR11 bacteria: the most abundant plankton in the oceans', *Annual Review of Marine Science*, 9(1), pp. 231–255. Available at: https://doi.org/10.1146/annurev-marine-010814-015934.

Gittel, A. *et al.* (2015) 'Ubiquitous presence and novel diversity of anaerobic alkane degraders in cold marine sediments', *Frontiers in Microbiology*, 6(1414), pp. 1–15. Available at: https://doi.org/10.3389/fmicb.2015.01414.

Giulia, A. *et al.* (2021) 'Extending association rule mining to microbiome pattern analysis: tools and guidelines to support real applications', *Frontiers in Bioinformatics*, 1(794547), pp. 1–14. Available at: https://doi.org/10.3389/fbinf.2021.794547.

Glassing, A. *et al.* (2016) 'Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples', *Gut Pathogens*, 8(1), pp. 1–12. Available at: https://doi.org/10.1186/s13099-016-0103-7.

Glassman, S.I. and Martiny, J.B.H. (2018) 'Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units', *mSphere*, 3(4), pp. e00148-18. Available at: https://doi.org/10.1128/mSphere.00148-18.

Glöckner, F.O. *et al.* (2017) '25 years of serving the community with ribosomal RNA gene reference databases and tools', *Journal of Biotechnology*, 261(2), pp. 169–176. Available at: https://doi.org/10.1016/j.jbiotec.2017.06.1198.

Gloor, G.B. *et al.* (2017) 'Microbiome datasets are compositional: and this is not optional', *Frontiers in Microbiology*, 8(2224), pp. 1–6. Available at: https://doi.org/10.3389/fmicb.2017.02224.

Gobet, A. *et al.* (2012) 'Diversity and dynamics of rare and of resident bacterial populations in coastal sands', *The ISME Journal*, 6(3), pp. 542–553. Available at: https://doi.org/10.1038/ismej.2011.132.

Gobet, A., Quince, C. and Ramette, A. (2010) 'Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets', *Nucleic Acids Research*, 38(15), pp. e155–e155. Available at: https://doi.org/10.1093/nar/gkq545.

Göker, M. and Oren, A. (2023) 'Valid publication of four additional phylum names', *International Journal of Systematic and Evolutionary Microbiology*, 73(9), pp. 1–3. Available at: https://doi.org/10.1099/ijsem.0.006024.

Gokul, J.K. *et al.* (2019) 'Illuminating the dynamic rare biosphere of the Greenland Ice Sheet's Dark Zone', *FEMS Microbiology Ecology*, 95(12), pp. 1-17. Available at: https://doi.org/10.1093/femsec/fiz177.

Gonzalez, J.M. *et al.* (2012) 'Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities', *PLoS ONE*, 7(1), p. e29973. Available at: https://doi.org/10.1371/journal.pone.0029973.

Gorrasi, S. *et al.* (2019) 'Structure and diversity of the bacterial community of an Arctic estuarine system (Kandalaksha Bay) subject to intense tidal currents', *Journal of Marine Systems*, 196(2), pp. 77–85. Available at: https://doi.org/10.1016/j.jmarsys.2019.04.004.

Granskog, M. *et al.* (2016) 'Arctic research on thin ice: consequences of Arctic sea ice loss', *Eos*, 97(3), pp. 22-26. Available at: https://doi.org/10.1029/2016EO044097.

Granskog, M.A. *et al.* (2018) 'Atmosphere-ice-ocean-ecosystem processes in a thinner Arctic sea ice regime: the Norwegian Young Sea ICE (N-ICE2015) Expedition', *Journal*

of Geophysical Research: Oceans, 123(3), pp. 1586–1594. Available at: https://doi.org/10.1002/2017JC013328.

Grattepanche, J.-D. *et al.* (2016) 'Unexpected biodiversity of ciliates in marine samples from below the photic zone', *Molecular Ecology*, 25(16), pp. 3987–4000. Available at: https://doi.org/10.1111/mec.13745.

Gruber-Dorninger, C. *et al.* (2015) 'Functionally relevant diversity of closely related Nitrospira in activated sludge', *ISME Journal*, 9(3), pp. 643–655. Available at: https://doi.org/10.1038/ismej.2014.156.

Gschwend, F. *et al.* (2021) 'Long-term stability of soil bacterial and fungal community structures revealed in their abundant and rare fractions', *Molecular Ecology*, 30(17), pp. 4305–4320. Available at: https://doi.org/10.1111/mec.16036.

Gurevich, A. *et al.* (2013) 'QUAST: quality assessment tool for genome assemblies', *Bioinformatics*, 29(8), pp. 1072–1075. Available at: https://doi.org/10.1093/bioinformatics/btt086.

Hahsler, M. *et al.* (2011) 'The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets', *Journal of Machine Learning Research*, 12(2011), pp. 2021–2025. Available at: https://www.jmlr.org/papers/v12/hahsler11a.html

Hahsler, M. *et al.* (2021) 'arules: Mining Association Rules and Frequent Itemsets', R package version 1.7-9. Available at: https://doi.org/10.32614/CRAN.package.arules.

Hakimzadeh, A. *et al.* (2024) 'A pile of pipelines: an overview of the bioinformatics software for metabarcoding data analyses', *Molecular Ecology Resources*, 24(e13847), p. e13847. Available at: https://doi.org/10.1111/1755-0998.13847.

Hamasaki, K. *et al.* (2016) 'Active populations of rare microbes in oceanic environments as revealed by bromodeoxyuridine incorporation and 454 tag sequencing', *Gene*, 576(2), pp. 650–656. Available at: https://doi.org/10.1016/j.gene.2015.10.016.

Hamp, T.J., Jones, W.J. and Fodor, A.A. (2009) 'Effects of experimental choices and analysis noise on surveys of the "Rare Biosphere"', *Applied and Environmental Microbiology*, 75(10), pp. 3263–3270. Available at: https://doi.org/10.1128/AEM.01931-08.

Han, D. *et al.* (2021) 'Survey of bacterial phylogenetic diversity during the glacier melting season in an Arctic fjord', *Microbial Ecology*, 81(3), pp. 579–591. Available at: https://doi.org/10.1007/s00248-020-01616-4.

Han, D. *et al.* (2024) 'Interaction between phytoplankton and heterotrophic bacteria in Arctic fjords during the glacial melting season as revealed by eDNA metabarcoding', *FEMS Microbiology Ecology*, 100(5), pp. 1-12. Available at: https://doi.org/10.1093/femsec/fiae059.

Handler, E.R. *et al.* (2023) 'Seasonality in land-ocean connectivity and local processes control sediment bacterial community structure and function in a High Arctic tidal flat', *FEMS Microbiology Ecology*, 100(1), pp. 1–20. Available at: https://doi.org/10.1093/femsec/fiad162.

Hanson, C.A. *et al.* (2012) 'Beyond biogeographic patterns: processes shaping the microbial landscape', *Nature Reviews Microbiology*, 10(7), pp. 497–506. Available at: https://doi.org/10.1038/nrmicro2795.

Hardge, K. *et al.* (2018) 'Impact of sequence processing and taxonomic classification approaches on eukaryotic community structure from environmental samples with emphasis on diatoms', *Molecular Ecology Resources*, 18(2), pp. 204-216. Available at: https://doi.org/10.1111/1755-0998.12726.

Hardoim, C.C.P. and Costa, R. (2014) 'Temporal dynamics of prokaryotic communities in the marine sponge *Sarcotragus spinosulus*', *Molecular Ecology*, 23(12), pp. 3097–3112. Available at: https://doi.org/10.1111/mec.12789.

Haro-Moreno, J.M., Rodriguez-Valera, F. and López-Pérez, M. (2019) 'Prokaryotic population dynamics and viral predation in a marine succession experiment using metagenomics', *Frontiers in Microbiology*, 10(2926), pp. 1–14. Available at: https://doi.org/10.3389/fmicb.2019.02926.

Hausmann, B. *et al.* (2016) 'Consortia of low-abundance bacteria drive sulfate reduction-dependent degradation of fermentation products in peat soil microcosms', *ISME Journal*, 10(10), pp. 2365–2375. Available at: https://doi.org/10.1038/ismej.2016.42.

Hausmann, B. *et al.* (2019) 'Long-term transcriptional activity at zero growth of a cosmopolitan rare biosphere member', *mBio*, 10(1), pp. 1–16. Available at: https://doi.org/10.1128/mBio.02189-18.

He, X. *et al.* (2015) 'Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle', *Proceedings of the National Academy of Sciences*, 112(1), pp. 244–249. Available at: https://doi.org/10.1073/PNAS.1419038112.

Henry, L. and Wickham, H. (2024) 'rlang: Functions for Base Types and Core R and "Tidyverse" Features.' R package version 1.1.4. Available at: https://rlang.r-lib.org/.

Herath, D. *et al.* (2017) 'Assessing species diversity using metavirome data: methods and challenges', *Computational and Structural Biotechnology Journal*, 15, pp. 447–455. Available at: https://doi.org/10.1016/j.csbj.2017.09.001.

Hol, W.H.G. *et al.* (2010) 'Reduction of rare soil microbes modifies plant-herbivore interactions', *Ecology Letters*, 13(3), pp. 292–301. Available at: https://doi.org/10.1111/j.1461-0248.2009.01424.x.

Hol, W.H.G. *et al.* (2015) 'Context dependency and saturating effects of loss of rare soil microbes on plant productivity', *Frontiers in Plant Science*, 6(485), pp. 1–10. Available at: https://doi.org/10.3389/fpls.2015.00485.

Holmfeldt, K. *et al.* (2013) 'Twelve previously unknown phage genera are ubiquitous in global oceans', *Proceedings of the National Academy of Sciences*, 110(31), pp. 12798–12803. Available at: https://doi.org/10.1073/pnas.1305956110.

Honeyman, A.S. *et al.* (2022) 'Statistical learning and uncommon soil microbiota explain biogeochemical responses after wildfire', *Applied and Environmental Microbiology*, 88(13), pp. 1-18. Available at: https://doi.org/10.1128/aem.00343-22.

Hooper, P.M. *et al.* (2024) 'Arctic cyanobacterial mat community diversity decreases with latitude across the Canadian Arctic', *FEMS Microbiology Ecology*, 100(6). Available at: https://doi.org/10.1093/femsec/fiae067.

Hörstmann, C. *et al.* (2024) 'Biogeographic gradients of picoplankton diversity indicate increasing dominance of prokaryotes in warmer Arctic fjords', *Communications Biology*, 7(1), pp. 1–11. Available at: https://doi.org/10.1038/s42003-024-05946-8.

Hou, J. *et al.* (2020) 'Biogeography and diversity patterns of abundant and rare bacterial communities in rice paddy soils across China', *Science of the Total Environment*, 730. Available at: https://doi.org/10.1016/j.scitotenv.2020.139116.

Hu, B. *et al.* (2020) 'High-throughput single-cell cultivation reveals the underexplored rare biosphere in deep-sea sediments along the Southwest Indian Ridge', *Lab on a Chip*, 20(2), pp. 363–372. Available at: https://doi.org/10.1039/c9lc00761j.

Hubbell, S.P. (2005) 'Neutral theory in community ecology and the hypothesis of functional equivalence', *Functional Ecology*, 19(1), pp. 166–172. Available at: https://doi.org/10.1111/j.0269-8463.2005.00965.x.

Hubert, C. *et al.* (2009) 'A constant flux of diverse thermophilic bacteria into the cold Arctic seabed', *Science*, 325(5947), pp. 1541–1544. Available at: https://doi.org/10.1126/science.1174012.

Hug, L.A. *et al.* (2016) 'A new view of the tree of life', *Nature Microbiology*, 1(5), p. 16048. Available at: https://doi.org/10.1038/nmicrobiol.2016.48.

Huggett, M.J. and Apprill, A. (2019) 'Coral microbiome database: integration of sequences reveals high diversity and relatedness of coral-associated microbes', *Environmental Microbiology Reports*, 11(3), pp. 372–385. Available at: https://doi.org/https://doi.org/10.1111/1758-2229.12686.

Hugoni, M. *et al.* (2013) 'Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters', *Proceedings of the National Academy of Sciences*, 110(15), pp. 6004–6009. Available at: https://doi.org/10.1073/pnas.1216863110.

Hunt, D.E. *et al.* (2013) 'Relationship between abundance and specific activity of bacterioplankton in open ocean surface waters', *Applied and Environmental Microbiology*, 79(1), pp. 177–184. Available at: https://doi.org/10.1128/AEM.02155-12.

Hur, M. and Park, S.J. (2019) 'Identification of microbial profiles in heavy-metal-contaminated soil from full-length 16S rRNA reads sequenced by a PacBio system', *Microorganisms*, 7(9), pp. 1-15. Available at: https://doi.org/10.3390/microorganisms7090357.

Huse, S.M. *et al.* (2008) 'Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing', *PLoS Genetics*, 4(11), p. e1000255. Available at: https://doi.org/10.1371/journal.pgen.1000255.

Huse, S.M. *et al.* (2010) 'Ironing out the wrinkles in the rare biosphere through improved OTU clustering', *Environmental Microbiology*, 12(7), pp. 1889–1898. Available at: https://doi.org/10.1111/j.1462-2920.2010.02193.x.

Idris, H. *et al.* (2017) 'Actinobacterial rare biospheres and dark matter revealed in habitats of the Chilean Atacama Desert', *Scientific Reports*, 7(1), pp. 1–11. Available at: https://doi.org/10.1038/s41598-017-08937-4.

İnceoğlu, Ö. *et al.* (2015) 'Vertical distribution of functional potential and active microbial communities in meromictic lake Kivu', *Microbial Ecology*, 70(3), pp. 596-611. Available at: https://doi.org/10.1007/s00248-015-0612-9.

von Jackowski, A. *et al.* (2022) 'Variations of microbial communities and substrate regimes in the eastern Fram Strait between summer and fall', *Environmental Microbiology*, 24(9), pp. 4124–4136. Available at: https://doi.org/10.1111/1462-2920.16036.

Jafari, M. and Ansari-Pour, N. (2019) 'Why, when and how to adjust your P values?', *Cell Journal*, 20(4), pp. 604–607. Available at: https://doi.org/10.22074/cellj.2019.5992.

Jaffe, A.L. *et al.* (2020) 'The rise of diversity in metabolic platforms across the Candidate Phyla Radiation', *BMC Biology*, 18(1), p. 69. Available at: https://doi.org/10.1186/s12915-020-00804-5.

Ji, M. *et al.* (2020) 'Distinct assembly mechanisms underlie similar biogeographical patterns of rare and abundant bacteria in Tibetan Plateau grassland soils', *Environmental Microbiology*, 22(6), pp. 2261–2272. Available at: https://doi.org/10.1111/1462-2920.14993.

Jia, X., Dini-Andreote, F. and Falcão Salles, J. (2018) 'Community assembly processes of the microbial rare biosphere', *Trends in Microbiology*, 26(9), pp. 738–747. Available at: https://doi.org/10.1016/j.tim.2018.02.011.

Jia, X., Dini-Andreote, F. and Salles, J.F. (2022) 'Unravelling the interplay of ecological processes structuring the bacterial rare biosphere', *ISME Communications*, 2(1), pp. 1–11. Available at: https://doi.org/10.1038/s43705-022-00177-6.

Jia, Y. *et al.* (2022) 'Sequencing introduced false positive rare taxa lead to biased microbial community diversity, assembly, and interaction interpretation in amplicon studies', *Environmental Microbiome*, 17(1), p. 43. Available at: https://doi.org/10.1186/s40793-022-00436-y.

Jia, Z. *et al.* (2023) 'Enrichment of rare methanogenic Archaea shows their important ecological role in natural high-CO2 terrestrial subsurface environments', *Frontiers in Microbiology*, 14(110529), pp. 1–12. Available at: https://doi.org/10.3389/fmicb.2023.1105259.

Jiang, C.-Y. *et al.* (2016) 'High-throughput single-cell cultivation on microfluidic streak plates', *Applied and Environmental Microbiology*, 82(7), pp. 2210–2218. Available at: https://doi.org/10.1128/AEM.03588-15.

Johnson, J.S. *et al.* (2019) 'Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis', *Nature Communications*, 10(1), pp. 1–11. Available at: https://doi.org/10.1038/s41467-019-13036-1.

Jones, S.E. and Lennon, J.T. (2010) 'Dormancy contributes to the maintenance of microbial diversity', *Proceedings of the National Academy of Sciences*, 107(13), pp. 5881–5886. Available at: https://doi.org/10.1073/pnas.0912765107.

Jousset, A. *et al.* (2017) 'Where less may be more: how the rare biosphere pulls ecosystems strings', *The ISME Journal*, 11(4), pp. 853–862. Available at: https://doi.org/10.1038/ismej.2016.174.

Kalenitchenko, D. *et al.* (2018) 'Ultrarare marine microbes contribute to key sulphur-related ecosystem functions', *Molecular Ecology*, 27(6), pp. 1494–1504. Available at: https://doi.org/10.1111/mec.14513.

Kalvari, I. *et al.* (2018) 'Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families', *Nucleic Acids Research*, 46(D1), pp. D335–D342. Available at: https://doi.org/10.1093/nar/gkx1038.

Kaminsky, R. and Morales, S.E. (2018) 'Conditionally rare taxa contribute but do not account for changes in soil prokaryotic community structure', *Frontiers in Microbiology*, 9(809), pp. 1–6. Available at: https://doi.org/10.3389/fmicb.2018.00809.

Kamono, A. *et al.* (2013) 'Exploring slime mould diversity in high-altitude forests and grasslands by environmental RNA analysis', *FEMS Microbiology Ecology*, 84(1), pp. 98-109. Available at: https://doi.org/10.1111/1574-6941.12042.

Kampouris, I.D. *et al.* (2023) 'Long-term patterns of hydrocarbon biodegradation and bacterial community composition in epipelagic and mesopelagic zones of an Arctic fjord', *Journal of Hazardous Materials*, 446(130656), pp. 1-12. Available at: https://doi.org/10.1016/j.jhazmat.2022.130656.

Kang, S. *et al.* (2016) 'Hill number as a bacterial diversity measure framework with high-throughput sequence data', *Scientific Reports*, 6(38263), pp. 1–4. Available at: https://doi.org/10.1038/srep38263.

Karimi, E. *et al.* (2019) 'Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable *Alphaproteobacteria'*, *Scientific Reports*, 9(1), pp. 1–15. Available at: https://doi.org/10.1038/s41598-019-38737-x.

Karsenti, E. *et al.* (2011) 'A holistic approach to marine eco-systems biology', *PLoS Biology*, 9(10), p. e1001177. Available at: https://doi.org/10.1371/journal.pbio.1001177.

Karsenti, E. (2015) 'The making of Tara Oceans: funding blue skies research for our Blue Planet.', *Molecular systems biology*, 11(5), p. 811. Available at: https://doi.org/10.15252/MSB.20156271.

Karthikeyan, S. *et al.* (2020) 'Integrated omics elucidate the mechanisms driving the rapid biodegradation of deepwater horizon oil in intertidal sediments undergoing oxic–anoxic cycles', *Environmental Science & Technology*, 54(16), pp. 10088–10099. Available at: https://doi.org/10.1021/acs.est.0c02834.

Kassambara, A. (2019) *Practical Statistics in R II - Comparing Groups: Numerical Variables*. Datanovia. Available at: https://www.jaypeedigital.com/book/9789350250037/chapter/ch12.

Kassambara, A. (2022) 'rstatix: Pipe-Friendly Framework for Basic Statistical Tests'. R package version 0.7.0.Available at: https://cran.r-project.org/package=rstatix.

Kaufman, L. and Rousseeuw, P.J. (1987) 'Clustering by means of Medoids', in *Statistical Data Analysis Based on the L1 Norm and Related Methods*, pp. 405–416.

Kaufman, L. and Rousseuw, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Biometrics. Available at: https://doi.org/10.2307/2532178.

Kaur, J. and Madan, N. (2015) 'Association rule mining: a survey', *International Journal of Hybrid Information Technology*, 8(7), pp. 239–242. Available at: https://doi.org/10.14257/ijhit.2015.8.7.22.

Keller-Costa, T. *et al.* (2021) 'Metagenomic insights into the taxonomy, function, and dysbiosis of prokaryotic communities in octocorals', *Microbiome*, 9(1), p. 72. Available at: https://doi.org/https://doi.org/10.1186/s40168-021-01031-y.

Kendall, M.G. (1938) 'A New Measure of Rank Correlation', *Biometrika*, 30(1/2), p. 81. Available at: https://doi.org/10.2307/2332226.

Kim, D.Y. *et al.* (2011) 'Rapid shifts in the structure and composition of a protistan assemblage during bottle incubations affect estimates of total protistan species richness', *Microbial Ecology*, 62(2), pp. 383–398. Available at: https://doi.org/10.1007/s00248-011-9816-9.

Kirchman, D.L., Cottrell, M.T. and Lovejoy, C. (2010) 'The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA

genes', *Environmental Microbiology*, 12(5), pp. 1132–1143. Available at: https://doi.org/10.1111/j.1462-2920.2010.02154.x.

Kirstein, I.V. *et al.* (2019) 'The plastisphere – Uncovering tightly attached plastic "specific" microorganisms', *PLoS ONE*, 14(4), pp. 1–17. Available at: https://doi.org/10.1371/journal.pone.0215859.

Klein Jan, H. *et al.* (2017) 'Exploring the cultivable *Ectocarpus* microbiome', *Frontiers in Microbiology*, 8(2456), pp. 1–13. Available at: https://doi.org/10.3389/fmicb.2017.02456.

Klemetsen, T., Willassen, N.P. and Karlsen, C.R. (2019) 'Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis', *MicrobiologyOpen*, 8(10), p. e898. Available at: https://doi.org/10.1002/mbo3.898.

Knowles, B. *et al.* (2016) 'Lytic to temperate switching of viral communities', *Nature*, 531(7595), pp. 466–470. Available at: https://doi.org/10.1038/nature17193.

Kopf, A. *et al.* (2015) 'The ocean sampling day consortium', *GigaScience*, 4(1), p. 27. Available at: https://doi.org/10.1186/s13742-015-0066-5.

Köstner, N. *et al.* (2017) 'High viral abundance as a consequence of low viral decay in the Baltic Sea redoxcline', *PLoS ONE*, 12(6), pp. 1–21. Available at: https://doi.org/10.1371/journal.pone.0178467.

Kraemer, S. *et al.* (2020) 'Diversity and biogeography of SAR11 bacteria from the Arctic Ocean', *ISME Journal*, 14(1), pp. 79–90. Available at: https://doi.org/10.1038/s41396-019-0499-4.

Kraemer, S.A. *et al.* (2024) 'A multiyear time series (2004–2012) of bacterial and archaeal community dynamics in a changing Arctic Ocean', *ISME Communications*, 4(1), pp. 1–4. Available at: https://doi.org/10.1093/ismeco/ycad004.

Kroeger, M.E. *et al.* (2018) 'New biological insights into how deforestation in Amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes', *Frontiers in Microbiology*, 9(1635), pp. 1–13. Available at: https://doi.org/10.3389/fmicb.2018.01635.

Kruskal, W.H. and Wallis, W.A. (1952) 'Use of Ranks in One-Criterion Variance Analysis,' *Journal of the American Statistical Association*, 47(260), pp. 583–621. Available at: https://doi.org/10.1080/01621459.1952.10483441.

Krzywinski, M. *et al.* (2009) 'Circos: an information aesthetic for comparative genomics', *Genome Research*, 19(9), pp. 1639–1645. Available at: https://doi.org/10.1101/gr.092759.109.

Krzywinski, M. and Altman, N. (2014) 'Visualizing samples with box plots', *Nature Methods*, 11(2), pp. 119–120. Available at: https://doi.org/10.1038/nmeth.2813.

Kumar, V. *et al.* (2019) 'Long-read amplicon denoising', *Nucleic Acids Research*, 47(18), pp. e104–e104. Available at: https://doi.org/10.1093/nar/gkz657.

Kunin, V. *et al.* (2010) 'Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates', *Environmental Microbiology*, 12(1), pp. 118-123. Available at: https://doi.org/10.1111/j.1462-2920.2009.02051.x.

Kunin, W.E. and Gaston, K.J. (1993) 'The biology of rarity: patterns, causes and consequences.', *Trends in ecology & evolution*, 8(8), pp. 298–301. Available at: https://doi.org/10.1016/0169-5347(93)90259-R.

Kurm, V. *et al.* (2017) 'Low abundant soil bacteria can be metabolically versatile and fast growing', *Ecology*, 98(2), pp. 555–564. Available at: https://doi.org/10.1002/ecy.1670.

Kurm, V., van der Putten, W.H. and Hol, W.H.G. (2019) 'Cultivation-success of rare soil bacteria is not influenced by incubation time and growth medium', *PLOS ONE*, 14(1), p. e0210073. Available at: https://doi.org/10.1371/journal.pone.0210073.

Labrie, R. *et al.* (2022) 'Deep ocean microbial communities produce more stable dissolved organic matter through the succession of rare prokaryotes', *Sci. Adv*, 8, p. 35. Available at: https://doi.org/eabn0035.

Lavergne, C. *et al.* (2018) 'First evidence of the presence and activity of archaeal C3 group members in an Atlantic intertidal mudflat', *Scientific Reports*, 8(1), p. 11790. Available at: https://doi.org/10.1038/s41598-018-30222-1.

Lawton, J.H. (1989) 'What is the relationship between population density and body size in animals?', *Oikos*, 55(3), p. 429. Available at: https://doi.org/10.2307/3565606.

Lee, C.K. *et al.* (2012) 'The inter-valley soil comparative survey: the ecology of Dry Valley edaphic microbial communities', *The ISME Journal*, 6(5), pp. 1046–1057. Available at: https://doi.org/10.1038/ismej.2011.170.

Li, C. *et al.* (2023) 'Destabilized microbial networks with distinct performances of abundant and rare biospheres in maintaining networks under increasing salinity stress', *iMeta*, 2(1), pp. 1–16. Available at: https://doi.org/10.1002/imt2.79.

Li, J. *et al.* (2018) 'The effect of filtration method on the efficiency of environmental DNA capture and quantification via metabarcoding', *Molecular Ecology Resources*, 18(5), pp. 1102–1114. Available at: https://doi.org/10.1111/1755-0998.12899.

Liang, Y. *et al.* (2020) 'Differentiation strategies of soil rare and abundant microbial taxa in response to changing climatic regimes', *Environmental Microbiology*, 22(4), pp. 1327–1340. Available at: https://doi.org/10.1111/1462-2920.14945.

Liang, Z. and Keeley, A. (2013) 'Filtration recovery of extracellular DNA from environmental water samples', *Environmental Science and Technology*, 47(16), pp. 9324–9331. Available at: https://doi.org/10.1021/es401342b.

Liu, L. *et al.* (2015) 'The biogeography of abundant and rare bacterioplankton in the lakes and reservoirs of China', *The ISME Journal*, 9(9), pp. 2068–2077. Available at: https://doi.org/10.1038/ismej.2015.29.

Liu, M., Xue, Y. and Yang, J. (2019) 'Rare plankton subcommunities are far more affected by DNA extraction kits than abundant plankton', *Frontiers in Microbiology*, 10(454), pp. 1–12. Available at: https://doi.org/10.3389/fmicb.2019.00454.

Locey, K.J. and Lennon, J.T. (2016) 'Scaling laws predict global microbial diversity', *Proceedings of the National Academy of Sciences of the United States of America*, 113(21), pp. 5970–5975. Available at: https://doi.org/10.1073/pnas.1521291113.

Loganathachetti, D.S. *et al.* (2022) 'Total nitrogen influence bacterial community structure of active layer permafrost across summer and winter seasons in Ny-Ålesund, Svalbard', *World Journal of Microbiology and Biotechnology*, 38(2), pp. 1–13. Available at: https://doi.org/10.1007/s11274-021-03210-3.

Logares, R. *et al.* (2013) 'Biogeography of bacterial communities exposed to progressive long-term environmental change', *The ISME Journal*, 7(5), pp. 937–948. Available at: https://doi.org/10.1038/ismej.2012.168.

Logares, R. *et al.* (2014) 'Patterns of rare and abundant marine microbial eukaryotes', *Current Biology*, 24(8), pp. 813–821. Available at: https://doi.org/10.1016/j.cub.2014.02.050.

Logares, R., Mangot, J.F. and Massana, R. (2015) 'Rarity in aquatic microbes: placing protists on the map', *Research in Microbiology*, 166(10), pp. 831–841. Available at: https://doi.org/10.1016/j.resmic.2015.09.009.

Long, R.A. and Azam, F. (2001) 'Antagonistic interactions among marine pelagic bacteria', *Applied and Environmental Microbiology*, 67(11), pp. 4975-4983. Available at: https://doi.org/10.1128/aem.67.11.4975-4983.2001.

López-Pérez, M. *et al.* (2016) 'Not all particles are equal: the selective enrichment of particle-associated bacteria from the Mediterranean Sea', *Frontiers in Microbiology*, 7(996), pp. 1-15. Available at: https://doi.org/10.3389/fmicb.2016.00996.

Losos, J.B. (2008) 'Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species', *Ecology Letters*, 11(10), pp. 995–1003. Available at: https://doi.org/10.1111/j.1461-0248.2008.01229.x.

Lovejoy, C., Galand, P.E. and Kirchman, D.L. (2011) 'Picoplankton diversity in the Arctic Ocean and surrounding seas', *Marine Biodiversity*, 41(1), pp. 5–12. Available at: https://doi.org/10.1007/s12526-010-0062-z.

Lowe, W.H. and McPeek, M.A. (2014) 'Is dispersal neutral?', *Trends in Ecology & Evolution*, 29(8), pp. 444–450. Available at: https://doi.org/10.1016/j.tree.2014.05.009.

Luan, L. *et al.* (2023) 'Biogeographical patterns of abundant and rare bacterial biospheres in paddy soils across East Asia', *Environmental Microbiology*, 25(2), pp. 294–305. Available at: https://doi.org/10.1111/1462-2920.16281.

Lynch, M.D.J., Bartram, A.K. and Neufeld, J.D. (2012) 'Targeted recovery of novel phylogenetic diversity from next-generation sequence data', *ISME Journal*, 6(11), pp. 2067–2077. Available at: https://doi.org/10.1038/ismej.2012.50.

Lynch, M.D.J. and Neufeld, J.D. (2015) 'Ecology and exploration of the rare biosphere.', *Nature reviews. Microbiology*, 13(4), pp. 217–229. Available at: https://doi.org/10.1038/nrmicro3400.

MacQueen, J. (1967) 'Some methods for classification and analysis of multivariate observations', in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, pp. 281-297. Available at: http://projecteuclid.org/euclid.bsmsp/1200512992.

Maechler, M. *et al.* (2023) 'cluster: Cluster Analysis Basics and Extensions'. R package version 2.1.6. Available at: https://cran.r-project.org/package=cluster.

Magurran, A.E. and Henderson, P.A. (2003) 'Explaining the excess of rare species in natural species abundance distributions', *Nature*, 422(6933), pp. 714–716. Available at: https://doi.org/10.1038/nature01547.

Maidak, B.L. *et al.* (1996) 'The Ribosomal Database Project (RDP)', *Nucleic Acids Research*, 24(1), pp. 82–85. Available at: https://doi.org/10.1093/nar/24.1.82.

Mallmin, E., Traulsen, A. and De Monte, S. (2024) 'Chaotic turnover of rare and abundant species in a strongly interacting model community', *Proceedings of the National Academy of Sciences*, 121(11), pp. 1–12. Available at: https://doi.org/10.1073/pnas.2312822121.

De Mandal, S. *et al.* (2016) 'MiSeq HV4 16S rRNA gene analysis of bacterial community composition among the cave sediments of Indo-Burma biodiversity hotspot', *Environmental Science and Pollution Research*, 23(12), pp. 12216-26. Available at: https://doi.org/10.1007/s11356-016-6423-9.

Mars Brisbin, M., Conover, A.E. and Mitarai, S. (2020) 'Influence of regional oceanography and hydrothermal activity on protist diversity and community structure in the Okinawa trough', *Microbial Ecology*, 80(4), pp. 746–761. Available at: https://doi.org/10.1007/s00248-020-01583-w.

Martin, K. *et al.* (2021) 'The biogeographic differentiation of algal microbiomes in the upper ocean from pole to pole', *Nature Communications*, 12(1), pp. 1–15. Available at: https://doi.org/10.1038/s41467-021-25646-9.

Martinez-Gutierrez, C.A. and Aylward, F.O. (2019) 'Strong purifying selection is associated with genome streamlining in epipelagic *Marinimicrobia'*, *Genome Biology and Evolution*, 11(10), pp. 2887–2894. Available at: https://doi.org/10.1093/gbe/evz201.

Maruyama, F. *et al.* (2008) 'Application of real-time long and short polymerase chain reaction for sensitive monitoring of the fate of extracellular plasmid DNA introduced into river waters', *Microbes and environments*, 23(3), pp. 229–236. Available at: https://doi.org/10.1264/jsme2.23.229.

Massana, R. *et al.* (2015) 'Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing', *Environmental Microbiology*, 17(10), pp. 4035–4049. Available at: https://doi.org/10.1111/1462-2920.12955.

Matias Rodrigues, J.F. *et al.* (2017) 'MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis', *Bioinformatics*, 33(23), pp. 3808–3810. Available at: https://doi.org/10.1093/bioinformatics/btx517.

Matsui, K. *et al.* (2004) 'Use of the SYBR Green I fluorescent dye and a centrifugal filter device for rapid determination of dissolved DNA concentration in fresh water', *Aquatic Microbial Ecology*, 36(1), pp. 99-105. Available at: https://doi.org/10.3354/ame036099.

McDonald, D. *et al.* (2012) 'An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea', *ISME Journal*, 6(3), pp. 610–618. Available at: https://doi.org/10.1038/ismej.2011.139.

McGill, B.J. *et al.* (2007) 'Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework', *Ecology Letters*, 10(10), pp. 995–1015. Available at: https://doi.org/10.1111/j.1461-0248.2007.01094.x.

McLaren, M.R., Willis, A.D. and Callahan, B.J. (2019) 'Consistent and correctable bias in metagenomic sequencing experiments', *eLife*, 8(e46923), pp. 1–31. Available at: https://doi.org/10.7554/elife.46923.

McMurdie, P.J. and Holmes, S. (2014) 'Waste not, want not: why rarefying microbiome data is inadmissible', *PLoS Computational Biology*, 10(4), p. e1003531. Available at: https://doi.org/10.1371/journal.pcbi.1003531.

Meisner, A. *et al.* (2018) 'Drought legacy effects on the composition of soil fungal and prokaryote communities', *Frontiers in Microbiology*, 9(294), pp. 1–12. Available at: https://doi.org/10.3389/fmicb.2018.00294.

Meyer, A. *et al.* (2017) 'Winter to summer oceanographic observations in the Arctic Ocean north of Svalbard', *Journal of Geophysical Research: Oceans*, 122(8), pp. 6218–6237. Available at: https://doi.org/10.1002/2016JC012391.

Miksch, S. *et al.* (2024) 'Taxonomic and functional stability overrules seasonality in polar benthic microbiomes', *The ISME Journal*, 18(1), pp. 1–14. Available at: https://doi.org/10.1093/ismejo/wrad005.

Milanese, A. *et al.* (2019) 'Microbial abundance, activity and population genomic profiling with mOTUs2', *Nature Communications*, 10(1), p. 1014. Available at: https://doi.org/10.1038/s41467-019-08844-4.

Mioduchowska, M. *et al.* (2023) 'Contrasting marine microbial communities of the Fram Strait with the first confirmed record of Cyanobacteria *Prochlorococcus marinus* in the Arctic region', *Biology*, 12(1246), pp. 1-20. Available at: https://doi.org/10.3390/biology12091246.

Miranda-Sánchez, F., Rivera, J. and Vinuesa, P. (2016) 'Diversity patterns of *Rhizobiaceae* communities inhabiting soils, root surfaces and nodules reveal a strong selection of rhizobial partners by legumes', *Environmental microbiology*, 18(8), pp. 2375-91. Available at: https://doi.org/10.1111/1462-2920.13061.

Mitchell, A.L. *et al.* (2019) 'MGnify: the microbiome analysis resource in 2020', *Nucleic Acids Research*, 48(D1), pp. D570–D578. Available at: https://doi.org/10.1093/nar/gkz1035.

Mizuno, C.M. *et al.* (2013) 'Expanding the marine virosphere using metagenomics', *PLoS Genetics*, 9(12), p. e1003987. Available at: https://doi.org/10.1371/journal.pgen.1003987.

Mo, Y. *et al.* (2018) 'Biogeographic patterns of abundant and rare bacterioplankton in three subtropical bays resulting from selective and neutral processes', *The ISME Journal*, 12(9), pp. 2198–2210. Available at: https://doi.org/10.1038/s41396-018-0153-6.

Moitinho-Silva, L. *et al.* (2017) 'The sponge microbiome project', *GigaScience*, 6(10), pp. 1–13. Available at: https://doi.org/10.1093/gigascience/gix077.

Molina, V. *et al.* (2016) 'Bacterial active community cycling in response to solar radiation and their influence on nutrient changes in a high-altitude wetland', *Frontiers in Microbiology*, 7(1823), pp. 1–15. Available at: https://doi.org/10.3389/fmicb.2016.01823.

Mora-Ruiz, M.D.R. *et al.* (2016) 'Endophytic microbial diversity of the halophyte *Arthrocnemum macrostachyum* across plant compartments', *FEMS Microbiology Ecology*, 92(fiw145), pp. 1-10. Available at: https://doi.org/10.1093/femsec/fiw145.

Morris, R.M. *et al.* (2002) 'SAR11 clade dominates ocean surface bacterioplankton communities', *Nature*, 420(6917), pp. 806–810. Available at: https://doi.org/10.1038/nature01240.

Müller, O. *et al.* (2018) 'Spatiotemporal dynamics of ammonia-oxidizing *Thaumarchaeota* in distinct Arctic water masses', *Frontiers in Microbiology*, 9. Available at: https://doi.org/10.3389/fmicb.2018.00024.

Muyzer, G., de Waal, E.C. and Uitterlinden, A.G. (1993) 'Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA.', *Applied and environmental*

microbiology, 59(3), pp. 695–700. Available at: https://doi.org/0099-2240/93/030695-06$02.00/0.

Myer, P.R. *et al.* (2016) 'Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers', *Journal of Microbiological Methods*, 127, pp. 132–140. Available at: https://doi.org/10.1016/j.mimet.2016.06.004.

Mylnikov, A.P. *et al.* (2015) '*Massisteria marina* has a sister: *Massisteria voersi* sp. nov., a rare species isolated from coastal waters of the Baltic Sea', *European Journal of Protistology*, 51(4), pp. 299-310. Available at: https://doi.org/10.1016/j.ejop.2015.05.002.

Nawrocki, E.P. and Eddy, S.R. (2013) 'Infernal 1.1: 100-fold faster RNA homology searches', *Bioinformatics*, 29(22), pp. 2933–2935. Available at: https://doi.org/10.1093/bioinformatics/btt509.

Needham, D.M., Sachdeva, R. and Fuhrman, J.A. (2017) 'Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters', *ISME Journal*, 11(7), pp. 1614–1629. Available at: https://doi.org/10.1038/ismej.2017.29.

Nelson, W.C. and Stegen, J.C. (2015) 'The reduced genomes of *Parcubacteria* (OD1) contain signatures of a symbiotic lifestyle', *Frontiers in Microbiology*, 6(713), pp. 1–14. Available at: https://doi.org/10.3389/fmicb.2015.00713.

Neufeld, J.D., Li, J. and Mohn, W.W. (2008) 'Scratching the surface of the rare biosphere with ribosomal sequence tag primers', *FEMS Microbiology Letters*, 283(2), pp. 146–153. Available at: https://doi.org/10.1111/j.1574-6968.2008.01124.x.

Nicolas, A.M. *et al.* (2021) 'Soil candidate phyla radiation bacteria encode components of aerobic metabolism and co-occur with *Nanoarchaea* in the rare biosphere of rhizosphere grassland communities', *mSystems*, 6(4), pp. 1–7. Available at: https://doi.org/10.1128/msystems.01205-20.

Ning, D. *et al.* (2019) 'A general framework for quantitatively assessing ecological stochasticity', *Proceedings of the National Academy of Sciences*, 116(34), pp. 16892–16898. Available at: https://doi.org/10.1073/pnas.1904623116.

Niño-García, J.P., Ruiz-González, C. and del Giorgio, P.A. (2016) 'Landscape-scale spatial abundance distributions discriminate core from random components of boreal

lake bacterioplankton', *Ecology Letters*, 19(12), pp. 1506–1515. Available at: https://doi.org/10.1111/ele.12704.

Niño-García, J.P., Ruiz-González, C. and del Giorgio, P.A. (2017) 'Exploring the Ecological Coherence between the Spatial and Temporal Patterns of Bacterioplankton in Boreal Lakes', *Frontiers in Microbiology*, 8(636), pp. 1–12. Available at: https://doi.org/10.3389/fmicb.2017.00636.

Nolte, V. *et al.* (2010) 'Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity', *Molecular Ecology*, 19(14), pp. 2908–2915. Available at: https://doi.org/10.1111/j.1365-294X.2010.04669.x.

Ofiteru, I.D. *et al.* (2010) 'Combined niche and neutral effects in a microbial wastewater treatment community', *Proceedings of the National Academy of Sciences*, 107(35), pp. 15345–15350. Available at: https://doi.org/10.1073/pnas.1000604107.

Ogier, J.C. *et al.* (2019) 'RpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing', *BMC Microbiology*, 19(1), pp. 1–16. Available at: https://doi.org/10.1186/s12866-019-1546-z.

Oksanen, J. *et al.* (2018) 'vegan: Community Ecology Package', R package version 2.5.3. Available at: http://cran.r-project.org/package=vegan.

Oksanen, J. *et al.* (2025) 'vegan: Community Ecology Package', R package version 2.7.0. Available at: https://vegandevs.github.io/vegan/.

Omae, K. *et al.* (2021) 'Diversity analysis of thermophilic hydrogenogenic carboxydotrophs by carbon monoxide dehydrogenase amplicon sequencing using new primers', *Extremophiles*, 25(1), pp. 61–76. Available at: https://doi.org/10.1007/s00792-020-01211-y.

Oren, A. *et al.* (2023) 'International code of nomenclature of prokaryotes. Prokaryotic Code (2022 Revision)', *International Journal of Systematic and Evolutionary Microbiology*, 73(5), pp. 1–88. Available at: https://doi.org/10.1099/ijsem.0.005585.

Oren, A. and Garrity, G.M. (2021) 'Valid publication of the names of forty-two phyla of prokaryotes', *International Journal of Systematic and Evolutionary Microbiology*, 71(10), pp. 1-7. Available at: https://doi.org/10.1099/ijsem.0.005056.

Ortega-Retuerta, E. *et al.* (2013) 'Spatial variability of particle-attached and free-living bacterial diversity in surface waters from the Mackenzie River to the Beaufort Sea

(Canadian Arctic)', *Biogeosciences*, 10(4), pp. 2747–2759. Available at: https://doi.org/10.5194/bg-10-2747-2013.

Overgaard, C.K. *et al.* (2022) 'Application of ecosystem-specific reference databases for increased taxonomic resolution in soil microbial profiling', *Frontiers in Microbiology*, 13(94239), pp. 1-15. Available at: https://doi.org/10.3389/fmicb.2022.942396.

Padilla, C.C. *et al.* (2015) 'Standard filtration practices may significantly distort planktonic microbial diversity estimates', *Frontiers in Microbiology*, 6(547), pp. 1–10. Available at: https://doi.org/10.3389/fmicb.2015.00547.

Padilla-Gamiño, J.L. *et al.* (2012) 'From parent to gamete: vertical transmission of *Symbiodinium* (*Dinophyceae*) ITS2 sequence assemblages in the reef building coral *Montipora capitata'*, *PLoS ONE*, 7(e38440), pp. 1-11. Available at: https://doi.org/10.1371/journal.pone.0038440.

Paliy, O. *et al.* (2009) 'High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray', *Applied and Environmental Microbiology*, 75(11), pp. 3572–3579. Available at: https://doi.org/10.1128/AEM.02764-08.

Pan, C. *et al.* (2020) '*Thioamycolamides* A–E, sulfur-containing cycliclipopeptides produced by the rare Actinomycete *Amycolatopsis* sp.', *Organic Letters*, 22(8), pp. 3014–3017. Available at: https://doi.org/10.1021/acs.orglett.0c00776.

Pan, X. *et al.* (2017) 'Highly sensitive detection of low-abundance white spot syndrome virus by a pre-amplification PCR method', *Journal of Microbiology and Biotechnology*, 27(3), pp. 471–479. Available at: https://doi.org/10.4014/jmb.1606.06053.

Papadakis, P. *et al.* (2021) 'Characterization of bacterial microbiota of P.D.O. feta cheese by 16S metagenomic analysis', *Microorganisms*, 9(2377), pp. 1-18. Available at: https://doi.org/10.3390/microorganisms9112377.

Parada, A.E., Needham, D.M. and Fuhrman, J.A. (2016) 'Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples', *Environmental Microbiology*, 18(5), pp. 1403–1414. Available at: https://doi.org/10.1111/1462-2920.13023.

Parks, D.H. *et al.* (2018) 'A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life', *Nature Biotechnology*, 36(10), pp. 996–1004. Available at: https://doi.org/10.1038/nbt.4229.

Parks, D.H. *et al.* (2020) 'A complete domain-to-species taxonomy for Bacteria and Archaea', *Nature Biotechnology*, 38(9), pp. 1079–1086. Available at: https://doi.org/10.1038/s41587-020-0501-8.

Parks, D.H. *et al.* (2022) 'GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy', *Nucleic Acids Research*, 50(D1), pp. D785–D794. Available at: https://doi.org/10.1093/nar/gkab776.

Pascoal, F. *et al.* (2022a) 'Exploration of the types of rarity in the Arctic Ocean from the perspective of multiple methodologies', *Microbial Ecology*, 84(1), pp. 59–72. Available at: https://doi.org/10.1007/s00248-021-01821-9.

Pascoal, F. *et al.* (2022b) 'The spatial distribution and biogeochemical drivers of nitrogen cycle genes in an Antarctic desert', *Frontiers in Microbiology*, 13(927129), pp. 1-16. Available at: https://doi.org/10.3389/fmicb.2022.927129.

Pascoal, F. *et al.* (2023) 'Inter-comparison of marine microbiome sampling protocols', *ISME Communications*, 3(1), pp. 1-16. Available at: https://doi.org/10.1038/s43705-023-00278-w.

Pascoal, F. *et al.* (2024) 'Full-length 16S rRNA gene sequencing combined with adequate database selection improves the description of Arctic marine prokaryotic communities', *Annals of Microbiology*, 74(29), pp. 1-12. Available at: https://doi.org/10.1186/s13213-024-01767-6.

Pascoal, F. *et al.* (2025a) 'Definition of the microbial rare biosphere through unsupervised machine learning', *Communications Biology* (in peer-review).

Pascoal, F. *et al.* (2025b) 'Architecture and implementation of *ulrb* algorithm in R', *Ecological Informatics.* Article in press. Available at: https://doi.org/10.1016/j.ecoinf.2025.103229.

Pascoal, F., Costa, R. and Magalhães, C. (2021) 'The microbial rare biosphere: current concepts, methods and ecological principles', *FEMS Microbiology Ecology*, 97(fiaa227), pp. 1-15. Available at: https://doi.org/10.1093/femsec/fiaa227.

Pascoal, F., Magalhães, C. and Costa, R. (2020) 'The link between the ecology of the prokaryotic rare biosphere and its biotechnological potential', *Frontiers in Microbiology*, 11(231), pp. 1-8. Available at: https://doi.org/10.3389/fmicb.2020.00231.

Paulsen, M.L. *et al.* (2016) '*Synechococcus* in the Atlantic gateway to the Arctic Ocean', *Frontiers in Marine Science*, 3(191), pp. 1-14. Available at: https://doi.org/10.3389/fmars.2016.00191.

Pedrós-Alió, C. (2006) 'Marine microbial diversity: can it be determined?', *Trends in Microbiology*, 14(6), pp. 257–263. Available at: https://doi.org/10.1016/j.tim.2006.04.007.

Pedrós-Alió, C. (2007) 'Dipping into the rare biosphere', *Science*, 315(5809), pp. 192–193. Available at: https://doi.org/10.1126/science.1135933.

Pedrós-Alió, C. (2012) 'The rare bacterial biosphere', *Annual Review of Marine Science*, 4(1), pp. 449–466. Available at: https://doi.org/10.1146/annurev-marine-120710-100948.

Pedrós-Alió, C., Potvin, M. and Lovejoy, C. (2015) 'Diversity of planktonic microorganisms in the Arctic Ocean', *Progress in Oceanography*, 139, pp. 233–243. Available at: https://doi.org/10.1016/j.pocean.2015.07.009.

Pepe-Ranney, C. *et al.* (2012) 'Cyanobacterial construction of hot spring siliceous stromatolites in Yellowstone National Park', *Environmental Microbiology*, 14(5), pp. 1182-97. Available at: https://doi.org/10.1111/j.1462-2920.2012.02698.x.

Perez-Mon, C., Frey, B. and Frossard, A. (2020) 'Functional and structural responses of Arctic and Alpine soil prokaryotic and fungal communities under freeze-thaw cycles of different frequencies', *Frontiers in Microbiology*, 11(982), pp. 1–14. Available at: https://doi.org/10.3389/fmicb.2020.00982.

Pesant, S. *et al.* (2015) 'Open science resources for the discovery and analysis of Tara Oceans data', *Scientific Data*, 2(150023), pp. 1–16. Available at: https://doi.org/10.1038/sdata.2015.23.

Pesant, S. (2017) 'Registry of all samples from the EuroMarine inter-comparison of marine plankton metagenome analysis methods (EMOSE 2017 edition)'. PANGEA Dataset. Available at: https://doi.org/https://doi.pangaea.de/10.1594/PANGAEA.879516.

Pessi, I.S. *et al.* (2018) 'Community structure and distribution of benthic cyanobacteria in Antarctic lacustrine microbial mats', *FEMS Microbiology Ecology*, 94(5), pp. 1–13. Available at: https://doi.org/10.1093/femsec/fiy042.

Pester, M. *et al.* (2010) 'A "rare biosphere" microorganism contributes to sulfate reduction in a peatland', *The ISME Journal*, 4(12), pp. 1591–1602. Available at: https://doi.org/10.1038/ismej.2010.75.

Pester, M. *et al.* (2012) 'Sulfate-reducing microorganisms in wetlands - fameless actors in carbon cycling and climate change', *Frontiers in Microbiology*, 3(72), pp. 1-19. Available at: https://doi.org/10.3389/fmicb.2012.00072.

Peter, H. *et al.* (2018) 'Changes in bacterioplankton community structure during early lake ontogeny resulting from the retreat of the Greenland Ice Sheet', *The ISME Journal*, 12(2), pp. 544–555. Available at: https://doi.org/10.1038/ismej.2017.191.

Pinseel, E. *et al.* (2020) 'Global radiation in a rare biosphere soil diatom', *Nature Communications*, 11(1), pp. 1–12. Available at: https://doi.org/10.1038/s41467-020-16181-0.

Piredda, R. *et al.* (2018) 'Diatom diversity through HTS-metabarcoding in coastal European seas', *Scientific Reports*, 8(1), pp. 1–12. Available at: https://doi.org/10.1038/s41598-018-36345-9.

Połka, J. *et al.* (2015) 'Bacterial diversity in typical Italian salami at different ripening stages as revealed by high-throughput sequencing of 16S rRNA amplicons', *Food Microbiology*, 46, pp. 342-356. Available at: https://doi.org/10.1016/j.fm.2014.08.023.

Polyakov, I. V. *et al.* (2017) 'Greater role for Atlantic inflows on sea-ice loss in the Eurasian Basin of the Arctic Ocean', *Science*, 356(6335), pp. 285–291. Available at: https://doi.org/10.1126/science.aai8204.

Polyakov, I. V. *et al.* (2020) 'Borealization of the Arctic Ocean in response to anomalous advection from sub-arctic seas', *Frontiers in Marine Science*, 7(491), pp. 1-32. Available at: https://doi.org/10.3389/fmars.2020.00491.

Pootakham, W. *et al.* (2019) 'Heat-induced shift in coral microbiome reveals several members of the *Rhodobacteraceae* family as indicator species for thermal stress in *Porites lutea'*, *Microbiology Open*, 8(e935), pp. 1-20. Available at: https://doi.org/10.1002/mbo3.935.

Pootakham, W. *et al.* (2021) 'Taxonomic profiling of *Symbiodiniaceae* and bacterial communities associated with Indo-Pacific corals in the Gulf of Thailand using PacBio sequencing of full-length ITS and 16S rRNA genes', *Genomics*, 113(4), pp. 2717–2729. Available at: https://doi.org/10.1016/j.ygeno.2021.06.001.

Preston, F.W. (1948) 'The commonness and rarity of species', *Ecology*, 29(3), pp. 254–283. Available at: http://www.jstor.org/stable/1930989.

Priest, T. *et al.* (2021) 'Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms', *PeerJ*, 9, pp. 1–19. Available at: https://doi.org/10.7717/peerj.11721.

Priest, T. *et al.* (2023) 'Atlantic water influx and sea-ice cover drive taxonomic and functional shifts in Arctic marine bacterial communities', *ISME Journal*, 17(10), pp. 1612–1625. Available at: https://doi.org/10.1038/s41396-023-01461-6.

Proctor, L.M. *et al.* (2019) 'The integrative human microbiome project', *Nature*, 569(7758), pp. 641–648. Available at: https://doi.org/10.1038/s41586-019-1238-8.

Prodan, A. *et al.* (2020) 'Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing', *PLoS ONE*, 15(1), pp. 1–19. Available at: https://doi.org/10.1371/journal.pone.0227434.

Promińska, A., Cisek, M. and Walczowski, W. (2017) 'Kongsfjorden and Hornsund hydrography – comparative study based on a multiyear survey in fjords of west Spitsbergen', *Oceanologia*, 59(4), pp. 397–412. Available at: https://doi.org/10.1016/j.oceano.2017.07.003.

Pruesse, E. *et al.* (2007) 'SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB', *Nucleic Acids Research*, 35(21), pp. 7188–7196. Available at: https://doi.org/10.1093/nar/gkm864.

Pust, M.M. and Tümmler, B. (2022) 'Bacterial low-abundant taxa are key determinants of a healthy airway metagenome in the early years of human life', *Computational and Structural Biotechnology Journal*, 20, pp. 175–186. Available at: https://doi.org/10.1016/j.csbj.2021.12.008.

Qin, J. *et al.* (2010) 'A human gut microbial gene catalogue established by metagenomic sequencing', *Nature*, 464(7285), pp. 59–65. Available at: https://doi.org/10.1038/nature08821.

Qin, Z. *et al.* (2022) 'Responses of abundant and rare prokaryotic taxa in a controlled organic contaminated site subjected to vertical pollution-induced disturbances', *Science of the Total Environment*, 853(158625), pp. 1-9. Available at: https://doi.org/10.1016/j.scitotenv.2022.158625.

Quast, C. *et al.* (2012) 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools', *Nucleic Acids Research*, 41(D1), pp. D590–D596. Available at: https://doi.org/10.1093/nar/gks1219.

Quero, G.M. *et al.* (2020) 'Inorganic and organic carbon uptake processes and their connection to microbial diversity in meso- and bathypelagic Arctic waters (Eastern Fram Strait)', *Microbial Ecology*, 79(4), pp. 823–839. Available at: https://doi.org/10.1007/s00248-019-01451-2.

Quero, G.M. and Luna, G.M. (2014) 'Diversity of rare and abundant bacteria in surface waters of the Southern Adriatic Sea', *Marine Genomics*, 17, pp. 9–15. Available at: https://doi.org/10.1016/j.margen.2014.04.002.

Quigley, K.M. *et al.* (2014) 'Deep-sequencing method for quantifying background abundances of *Symbiodinium* types: exploring the rare *Symbiodinium* biosphere in reef-building corals', *PLoS ONE*, 9(4), p. e94297. Available at: https://doi.org/10.1371/journal.pone.0094297.

R Core Team (2020) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

R Core Team (2023) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Rabelo-Ruiz, M. *et al.* (2021) 'Allium-based phytobiotic enhances egg production in laying hens through microbial composition changes in ileum and cecum', *Animals*, 11(2), p. 448. Available at: https://doi.org/10.3390/ani11020448.

Rabinowitz, D. (1981) 'Seven forms of rarity', in *The Biological Aspects of Rare Plant Conservation*. John Wiley & Sons, Inc., pp. 205–217. Available at: https://www.esf.edu/efb/parry/Invert_Cons_14_Readings/Rabinowitz_1981.pdf.

Ramírez, G.A. *et al.* (2019) 'Ecology of subseafloor crustal biofilms', *Frontiers in Microbiology*, 10(1983), pp. 1–17. Available at: https://doi.org/10.3389/fmicb.2019.01983.

Ramond, P. *et al.* (2023) 'Assembly processes and functional diversity of marine protists and their rare biosphere', *Environmental Microbiome*, 18(1), pp. 1–14. Available at: https://doi.org/10.1186/s40793-023-00513-w.

Rantanen, M. *et al.* (2022) 'The Arctic has warmed nearly four times faster than the globe since 1979', *Communications Earth and Environment*, 3(1), pp. 1–10. Available at: https://doi.org/10.1038/s43247-022-00498-3.

Rego, A. *et al.* (2019) 'Actinobacteria and Cyanobacteria diversity in terrestrial Antarctic microenvironments evaluated by culture-dependent and independent methods',

Frontiers in Microbiology, 10(1018), pp. 1–19. Available at: https://doi.org/10.3389/fmicb.2019.01018.

Ren, Z., Luo, W. and Zhang, C. (2022) 'Rare bacterial biosphere is more environmental controlled and deterministically governed than abundant one in sediment of thermokarst lakes across the Qinghai-Tibet Plateau', *Frontiers in Microbiology*, 13(944646), pp. 1-12. Available at: https://doi.org/10.3389/fmicb.2022.944646.

Renner, A.H.H., Dodd, P.A. and Fransson, A. (2018) *An assessment of MOSJ - The state of the marine environment around Svalbard and Jan Mayen*. Tromsø: Norwegian Polar Institute, Fram Centre. Available at: http://hdl.handle.net/11250/2500141.

Reveillaud, J. *et al.* (2014) 'Host-specificity among abundant and rare taxa in the sponge microbiome', *The ISME Journal*, 8(6), pp. 1198–1209. Available at: https://doi.org/10.1038/ismej.2013.227.

Richa, K. *et al.* (2017) 'Distribution, community composition, and potential metabolic activity of bacterioplankton in an urbanized Mediterranean Sea coastal zone', *Applied and Environmental Microbiology*, 83(17), pp. 1–17. Available at: https://doi.org/10.1128/AEM.00494-17.

Richter-Heitmann, T. *et al.* (2020) 'Stochastic dispersal rather than deterministic selection explains the spatio-temporal distribution of soil bacteria in a temperate grassland', *Frontiers in Microbiology*, 11(1391), pp. 1–19. Available at: https://doi.org/10.3389/fmicb.2020.01391.

Riley, R. *et al.* (2023) 'Terabase-scale coassembly of a tropical soil microbiome', *Microbiology Spectrum*, 11(4), pp. 1–13. Available at: https://doi.org/10.1128/spectrum.00200-23.

Rocca, J.D. *et al.* (2020) 'Rare microbial taxa emerge when communities collide: freshwater and marine microbiome responses to experimental mixing', *Ecology*, 101(3956), p. e02956. Available at: https://doi.org/10.1002/ecy.2956.

Rodriguez-Lanetty, M. *et al.* (2013) 'Ecological inferences from a deep screening of the complex bacterial consortia associated with the coral, *Porites astreoides'*, *Molecular Ecology*, 22(16), pp. 4349–4362. Available at: https://doi.org/10.1111/mec.12392.

Rodríguez-Pérez, H., Ciuffreda, L. and Flores, C. (2022) 'NanoRTax, a real-time pipeline for taxonomic and diversity analysis of nanopore 16S rRNA amplicon sequencing data',

Computational and Structural Biotechnology Journal, 20, pp. 5350–5354. Available at: https://doi.org/10.1016/j.csbj.2022.09.024.

Rousseeuw, P.J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20(C), pp. 53–65. Available at: https://doi.org/10.1016/0377-0427(87)90125-7.

Royo-Llonch, M. *et al.* (2017) 'Exploring microdiversity in novel *Kordia* sp. (*Bacteroidetes*) with proteorhodopsin from the tropical Indian Ocean via single amplified genomes', *Frontiers in Microbiology*, 8(1317), pp. 1–14. Available at: https://doi.org/10.3389/fmicb.2017.01317.

Rusch, D.B. *et al.* (2007) 'The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific', *PLoS Biology*, 5(3), p. e77. Available at: https://doi.org/10.1371/journal.pbio.0050077.

Sabbagh, E.I. *et al.* (2020) 'Weekly variations of viruses and heterotrophic nanoflagellates and their potential impact on bacterioplankton in shallow waters of the central Red Sea', *FEMS Microbiology Ecology*, 96(fiaa033), pp. 1–16. Available at: https://doi.org/10.1093/femsec/fiaa033.

Salazar, G. *et al.* (2016) 'Global diversity and biogeography of deep-sea pelagic prokaryotes', *ISME Journal*, 10(3), pp. 596–608. Available at: https://doi.org/10.1038/ismej.2015.137.

Salazar, G. *et al.* (2019) 'Gene expression changes and community turnover differentially shape the global ocean metatranscriptome', *Cell*, 179(5), pp. 1068-1083.e21. Available at: https://doi.org/10.1016/j.cell.2019.10.014.

Sánchez-Paz, A. (2010) 'White spot syndrome virus: an overview on an emergent concern', *Veterinary Research*, 41(6), p. 43. Available at: https://doi.org/10.1051/vetres/2010015.

Santi, I. *et al.* (2021) *European Marine Omics Biodiversity Observation Network (EMO BON) Handbook*. Paris: EMBRC-ERIC. France. Available at: https://repository.oceanbestpractices.org/handle/11329/1738%0Awww.embrc.eu.

Santos, H.F. *et al.* (2012) 'Comparison of different protocols for the extraction of microbial DNA from reef corals', *Brazilian Journal of Microbiology*, 43(2), pp. 517–527. Available at: https://doi.org/10.1590/S1517-83822012000200012.

Sanz-Sáez, I. *et al.* (2023) 'Top abundant deep ocean heterotrophic bacteria can be retrieved by cultivation', *ISME Communications*, 3(1), pp. 1–13. Available at: https://doi.org/10.1038/s43705-023-00290-0.

Sauret, C. *et al.* (2014) '"Rare biosphere" bacteria as key phenanthrene degraders in coastal seawaters', *Environmental Pollution*, 194, pp. 246–253. Available at: https://doi.org/10.1016/j.envpol.2014.07.024.

Scales, B.S. *et al.* (2021) 'Cross-hemisphere study reveals geographically ubiquitous , unexplored biosphere', *American Society for Microbiology*, 6(3), pp. 1-17. Available at: https://doi.org/10.1128/msphere.00851-20.

Schiaffino, M.R. *et al.* (2016) 'Microbial eukaryote communities exhibit robust biogeographical patterns along a gradient of Patagonian and Antarctic lakes', *Environmental Microbiology*, 18(12), pp. 5249–5264. Available at: https://doi.org/10.1111/1462-2920.13566.

Schloss, P.D. *et al.* (2016) 'Status of the archaeal and bacterial census: an update', *mBio*, 7(3), pp. 1–10. Available at: https://doi.org/10.1128/mBio.00201-16.

Schreiber, L. *et al.* (2019) 'Potential for microbially mediated natural attenuation of diluted bitumen on the coast of British Columbia (Canada)', *Applied and Environmental Microbiology*, 85(10), pp. 1–19. Available at: https://doi.org/10.1128/AEM.00086-19.

Schultz, J. *et al.* (2022) 'Polyphasic analysis reveals potential petroleum hydrocarbon degradation and biosurfactant production by rare biosphere thermophilic bacteria from Deception Island, an active Antarctic volcano', *Frontiers in Microbiology*, 13(885557), pp. 1–19. Available at: https://doi.org/10.3389/fmicb.2022.885557.

Sebastián, M. *et al.* (2024) 'The active free-living bathypelagic microbiome is largely dominated by rare surface taxa', *ISME Communications*, 4(1), pp. 1-13. Available at: https://doi.org/10.1093/ismeco/ycae015.

Selje, N., Simon, M. and Brinkhoff, T. (2004) 'A newly discovered Roseobacter cluster in temperate and polar oceans', *Nature*, 427(6973), pp. 445-448. Available at: https://doi.org/10.1038/nature02272.

Semedo, M. *et al.* (2021) 'Depth profile of nitrifying archaeal and bacterial communities in the remote oligotrophic waters of the North Pacific', *Frontiers in Microbiology*, 12(3), pp. 1–18. Available at: https://doi.org/10.3389/fmicb.2021.624071.

Seol, D. *et al.* (2022) 'Microbial identification using rRNA operon region: database and tool for metataxonomics with long-read sequence', *Microbiology Spectrum*, 10(2), pp. 1-17. Available at: https://doi.org/10.1128/spectrum.02017-21.

Ser-Giacomi, E. *et al.* (2018) 'Ubiquitous abundance distribution of non-dominant plankton across the global ocean', *Nature Ecology & Evolution*, 2(8), pp. 1243–1249. Available at: https://doi.org/10.1038/s41559-018-0587-2.

Shade, A. *et al.* (2012) 'Culturing captures members of the soil rare biosphere', *Environmental Microbiology*, 14(9), pp. 2247–2252. Available at: https://doi.org/10.1111/j.1462-2920.2012.02817.x.

Shade, A. *et al.* (2014) 'Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity', *mBio*, 5(4), pp. 1–9. Available at: https://doi.org/10.1128/mBio.01371-14.

Shade, A. and Gilbert, J.A. (2015) 'Temporal patterns of rarity provide a more complete view of microbial diversity', *Trends in Microbiology*, 23(6), pp. 335–340. Available at: https://doi.org/10.1016/j.tim.2015.01.007.

Shaffer, J.P. *et al.* (2022) 'A comparison of six DNA extraction protocols for 16S, ITS and shotgun metagenomic sequencing of microbial communities', *BioTechniques*, 73(1), pp. 34–46. Available at: https://doi.org/10.2144/btn-2022-0032.

Shaw, G.T.W. *et al.* (2019) 'A systematic approach re-analyzing the effects of temperature disturbance on the microbial community of mesophilic anaerobic digestion', *Scientific Reports*, 9(1), pp. 1–14. Available at: https://doi.org/10.1038/s41598-019-42987-0.

Shu, D. *et al.* (2021) 'Rare prokaryotic sub-communities dominate the complexity of ecological networks and soil multinutrient cycling during long-term secondary succession in China's Loess Plateau', *Science of the Total Environment*, 774, p. 145737. Available at: https://doi.org/10.1016/j.scitotenv.2021.145737.

Sieburth, J.McN., Smetacek, V. and Lenz, J. (1978) 'Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions 1', *Limnology and Oceanography*, 23(6), pp. 1256–1263. Available at: https://doi.org/https://doi.org/10.4319/lo.1978.23.6.1256.

Siles, J.A., Hendrickson, A.J. and Terry, N. (2022) 'Coupling of metataxonomics and culturing improves bacterial diversity characterization and identifies a novel Rhizorhapis

sp. with metal resistance potential in a multi-contaminated waste sediment', *Journal of Environmental Management*, 322(116132), pp. 1-11. Available at: https://doi.org/10.1016/j.jenvman.2022.116132.

Silva, S.G. *et al.* (2022) 'Insights into the antimicrobial activities and metabolomes of *Aquimarina* (*Flavobacteriaceae*, *Bacteroidetes*) species from the rare marine biosphere', *Marine Drugs*, 20(7), p. 423. Available at: https://doi.org/10.3390/md20070423.

Sipos, R. *et al.* (2007) 'Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis', *FEMS Microbiology Ecology*, 60(2), pp. 341–350. Available at: https://doi.org/10.1111/j.1574-6941.2007.00283.x.

Sjöstedt, J. *et al.* (2012) 'Recruitment of members from the rare biosphere of marine bacterioplankton communities after an environmental disturbance', *Applied and Environmental Microbiology*, 78(5), pp. 1361–1369. Available at: https://doi.org/10.1128/AEM.05542-11.

Skogseth, R. *et al.* (2020) 'Variability and decadal trends in the Isfjorden (Svalbard) ocean climate and circulation – An indicator for climate change in the European Arctic', *Progress in Oceanography*, 187(102394), pp. 1-31. Available at: https://doi.org/10.1016/j.pocean.2020.102394.

Skoog, E.J. and Bosak, T. (2024) 'Predicted metabolic roles and stress responses provide insights into candidate phyla Hydrogenedentota and Sumerlaeota as members of the rare biosphere in biofilms from various environments', *Environmental Microbiology Reports*, 16(1), pp. 1–15. Available at: https://doi.org/10.1111/1758-2229.13228.

Sogin, M.L. *et al.* (2006) 'Microbial diversity in the deep sea and the underexplored "rare biosphere"', *Proceedings of the National Academy of Sciences*, 103(32), pp. 12115–12120. Available at: https://doi.org/10.1073/pnas.0605127103.

de Sousa, A.G.G. *et al.* (2019) 'Diversity and composition of pelagic prokaryotic and protist communities in a thin Arctic sea-ice regime', *Microbial Ecology*, 78(2), pp. 388–408. Available at: https://doi.org/10.1007/s00248-018-01314-2.

Sperling, M. *et al.* (2012) 'Differential effects of hydrographic and biogeochemical properties on the SAR11 clade and *Roseobacter* RCA cluster in the North Sea', *Aquatic Microbial Ecology*, 67(1), pp. 25–34. Available at: https://doi.org/10.3354/ame01580.

Stegen, J.C. *et al.* (2012) 'Stochastic and deterministic assembly processes in subsurface microbial communities', *ISME Journal*, 6(9), pp. 1653–1664. Available at: https://doi.org/10.1038/ismej.2012.22.

Stegen, J.C. *et al.* (2013) 'Quantifying community assembly processes and identifying features that impose them', *The ISME Journal*, 7(11), pp. 2069–2079. Available at: https://doi.org/10.1038/ismej.2013.93.

Stegen, J.C. *et al.* (2015) 'Estimating and mapping ecological processes influencing microbial community assembly', *Frontiers in Microbiology*, 6(370), pp. 1–15. Available at: https://doi.org/10.3389/fmicb.2015.00370.

Stegen, J.C. *et al.* (2016) 'Groundwater–surface water mixing shifts ecological assembly processes and stimulates organic carbon turnover', *Nature Communications*, 7(1), p. 11237. Available at: https://doi.org/10.1038/ncomms11237.

Steiner, P.A. *et al.* (2019) 'Highly variable mRNA half-life time within marine bacterial taxa and functional genes', *Environmental Microbiology*, 21(10), pp. 3873–3884. Available at: https://doi.org/10.1111/1462-2920.14737.

Steven, B. *et al.* (2017) 'Simulated rRNA/DNA ratios show potential to misclassify active populations as dormant', *Applied and Environmental Microbiology*, 83(11), pp. 1–11. Available at: https://doi.org/10.1128/AEM.00696-17.

Stingl, U. *et al.* (2007) 'The SAR92 clade: an abundant coastal clade of culturable marine bacteria possessing proteorhodopsin', *Applied and Environmental Microbiology*, 73(7), pp. 2290–2296. Available at: https://doi.org/10.1128/AEM.02559-06.

Stoeck, T. *et al.* (2009) 'Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities', *BMC Biology*, 7(i), p. 72. Available at: https://doi.org/10.1186/1741-7007-7-72.

Stoeck, T. *et al.* (2010) 'Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water', *Molecular Ecology*, 19(Suppl. 1), pp. 21–31. Available at: https://doi.org/10.1111/j.1365-294X.2009.04480.x.

Stoeck, T. *et al.* (2014) 'A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology', *Environmental Microbiology*, 16(2), pp. 430-444. Available at: https://doi.org/10.1111/1462-2920.12194.

Sunagawa, S. *et al.* (2015) 'Structure and function of the global ocean microbiome', *Science*, 348(1261359), pp. 1-9. Available at: https://doi.org/10.1126/science.1261359.

Sunagawa, S. *et al.* (2020) 'Tara Oceans: towards global ocean ecosystems biology', *Nature Reviews Microbiology*, 18(8), pp. 428–445. Available at: https://doi.org/10.1038/s41579-020-0364-5.

Sunagawa, S., Woodley, C.M. and Medina, M. (2010) 'Threatened corals provide underexplored microbial habitats', *PLoS ONE*, 5(3), p. e9554. Available at: https://doi.org/10.1371/journal.pone.0009554.

Swan, B.K. *et al.* (2011) 'Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean', *Science*, 333(6047), pp. 1296-1300. Available at: https://doi.org/10.1126/science.1203690.

Swenson, N.G. (2014) *Functional and Phylogenetic Ecology in R*, *Use R!* New York, NY: Springer New York (Use R!). Available at: https://doi.org/10.1007/978-1-4614-9542-0.

Tada, Y. *et al.* (2013) 'Growth and distribution patterns of *Roseobacter*/*Rhodobacter*, SAR11, and *Bacteroidetes* lineages in the Southern Ocean', *Polar Biology*, 36(5), pp. 691–704. Available at: https://doi.org/10.1007/s00300-013-1294-8.

Tan, P.N., Kumar, V. and Srivastava, J. (2004) 'Selecting the right objective measure for association analysis', *Information Systems*, 29(4), pp. 293–313. Available at: https://doi.org/10.1016/S0306-4379(03)00072-3.

Tang, L. *et al.* (2023) 'Plant Community Associates with Rare Rather than Abundant Fungal Taxa in Alpine Grassland Soils', *Applied and Environmental Microbiology*, 89(1), pp. 1–15. Available at: https://doi.org/10.1128/aem.01862-22.

Tara Ocean Foundation, Tara Oceans and European Molecular Biology Laboratory (2022) 'Priorities for ocean microbiome research', *Nature Microbiology*, 7(7), pp. 937–947. Available at: https://doi.org/10.1038/s41564-022-01145-5.

Taylor, M.W. *et al.* (2013) '"Sponge-specific" bacteria are widespread (but rare) in diverse marine environments', *The ISME Journal*, 7(2), pp. 438–443. Available at: https://doi.org/10.1038/ismej.2012.111.

Taylor, T.B. and Buckling, A. (2010) 'Competition and dispersal in *Pseudomonas aeruginosa'*, *The American Naturalist*, 176(1), pp. 83–89. Available at: https://doi.org/10.1086/652995.

Tedersoo, L. *et al.* (2021) 'Perspectives and benefits of high-throughput long-read sequencing in microbial ecology', *Applied and Environmental Microbiology*, 87(17), pp. 1–19. Available at: https://doi.org/10.1128/AEM.00626-21.

Teeling, H. *et al.* (2016) 'Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms', *eLife*, 5(e11888), pp. 1–31. Available at: https://doi.org/10.7554/eLife.11888.

Telling, J. *et al.* (2015) 'Rock comminution as a source of hydrogen for subglacial ecosystems', *Nature Geoscience*, 8(11), pp. 851-855. Available at: https://doi.org/10.1038/ngeo2533.

Tesi, T. *et al.* (2021) 'Rapid Atlantification along the Fram Strait at the beginning of the 20th century', *Science Advances*, 7(48), pp. 1-10. Available at: https://doi.org/10.1126/sciadv.abj2946.

Thiele, S. *et al.* (2022) 'A winter-to-summer transition of bacterial and archaeal communities in Arctic sea ice', *Microorganisms*, 10(8), p. 1618. Available at: https://doi.org/10.3390/microorganisms10081618.

Thingstad, T.F. (2000) 'Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems', *Limnology and Oceanography*, 45(6), pp. 1320–1328. Available at: https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.4319/lo.2000.45.6.1320.

Thingstad, T.F. *et al.* (2014) 'A theoretical analysis of how strain-specific viruses can control microbial species diversity', *Proceedings of the National Academy of Sciences of the United States of America*, 111(21), pp. 7813–7818. Available at: https://doi.org/10.1073/pnas.1400909111.

Thomas, F.A., Mohan, M. and Krishnan, K.P. (2021) 'Bacterial diversity and their metabolic profiles in the sedimentary environments of Ny-Ålesund, Arctic', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 114(9), pp. 1339–1360. Available at: https://doi.org/10.1007/s10482-021-01604-9.

Thomas, F.A., Sinha, R.K. and Krishnan, K.P. (2020) 'Bacterial community structure of a glacio-marine system in the Arctic (Ny-Ålesund, Svalbard)', *Science of the Total Environment*, 718, p. 135264. Available at: https://doi.org/10.1016/j.scitotenv.2019.135264.

Thomas, T. *et al.* (2016) 'Diversity, structure and convergent evolution of the global sponge microbiome', *Nature Communications*, 7(11870), pp. 1-12. Available at: https://doi.org/10.1038/ncomms11870.

Tian, C. *et al.* (2023) 'Microbial community structure and metabolic potential at the initial stage of soil development of the glacial forefields in Svalbard', *Microbial Ecology*, 86(2), pp. 933–946. Available at: https://doi.org/10.1007/s00248-022-02116-3.

Toubes-Rodrigo, M. *et al.* (2021) 'Active microbial ecosystem in glacier basal ice fueled by iron and silicate comminution-derived hydrogen', *Microbiology Open*, 10(4), pp. 1–13. Available at: https://doi.org/10.1002/mbo3.1200.

Tréguer, P. *et al.* (2018) 'Influence of diatom diversity on the ocean biological carbon pump', *Nature Geoscience*, 11(1), pp. 27–37. Available at: https://doi.org/10.1038/s41561-017-0028-x.

Tréguer, P.J. *et al.* (2021) 'Reviews and syntheses: The biogeochemical cycle of silicon in the modern ocean', *Biogeosciences*, 18(4), pp. 1269–1289. Available at: https://doi.org/10.5194/bg-18-1269-2021.

Tripp, H.J. (2013) 'The unique metabolism of SAR11 aquatic bacteria', *Journal of Microbiology*, 51(2), pp. 147–153. Available at: https://doi.org/10.1007/s12275-013-2671-2.

Troussellier, M. *et al.* (2017) 'Sustaining rare marine microorganisms: macroorganisms as repositories and dispersal agents of microbial diversity', *Frontiers in Microbiology*, 8(947), pp. 1–17. Available at: https://doi.org/10.3389/fmicb.2017.00947.

Tsementzi, D. *et al.* (2016) 'SAR11 bacteria linked to ocean anoxia and nitrogen loss', *Nature*, 536(7615), pp. 179–183. Available at: https://doi.org/10.1038/nature19068.

Tverberg, V. *et al.* (2019) "The Kongsfjorden Transect: Seasonal and Inter-annual Variability in Hydrography," in *The Ecosystem of Kongsfjorden, Svalbard*. Springer, pp. 49–104. Available at: https://doi.org/10.1007/978-3-319-46425-1_3.

Tytgat, B. *et al.* (2014) 'Bacterial diversity assessment in antarctic terrestrial and aquatic microbial mats: A comparison between bidirectional pyrosequencing and cultivation', *PLoS ONE*, 9(6), p. e97564. Available at: https://doi.org/10.1371/journal.pone.0097564.

Umbach, A.K., Stegelmeier, A.A. and Neufeld, J.D. (2021) 'Archaea are rare and uncommon members of the mammalian skin microbiome', *mSystems*, 6(4), pp. 1–16. Available at: https://doi.org/10.1128/msystems.00642-21.

Våge, S., Storesund, J.E. and Thingstad, T.F. (2013) 'SAR11 viruses and defensive host strains', *Nature*, 499(7459), pp. E3–E4. Available at: https://doi.org/10.1038/nature12387.

Varela, M.M. *et al.* (2020) 'Changes in activity and community composition shape bacterial responses to size-fractionated marine DOM', *Frontiers in Microbiology*, 11(586148), pp. 1–14. Available at: https://doi.org/10.3389/fmicb.2020.586148.

Vázquez-Rosas-Landa, M. *et al.* (2020) 'Population genomics of *Vibrionaceae* isolated from an endangered oasis reveals local adaptation after an environmental perturbation', *BMC genomics*, 21(1), p. 418. Available at: https://doi.org/10.1186/s12864-020-06829-y.

Vellend, M. (2010) 'Conceptual synthesis in community ecology', *The Quarterly Review of Biology*, 85(2), pp. 183–206. Available at: https://doi.org/10.1086/652373.

Venter, J.C. *et al.* (2004) 'Environmental genome shotgun sequencing of the Sargasso Sea', *Science*, 304(5667), pp. 66–74. Available at: https://doi.org/10.1126/science.1093857.

Vergin, K. *et al.* (2013) 'Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean', *Aquatic Microbial Ecology*, 71(1), pp. 1–13. Available at: https://doi.org/10.3354/ame01661.

Vigneron, A. *et al.* (2018) 'Beyond the tip of the iceberg; a new view of the diversity of sulfite- and sulfate-reducing microorganisms', *The ISME Journal*, 12(8), pp. 2096–2099. Available at: https://doi.org/10.1038/s41396-018-0155-4.

Vihtakari, M. (2020) 'PlotSvalbard: PlotSvalbard - Plot research'. R package Version 0.9.2. Available at: https://github.com/MikkoVihtakari/PlotSvalbard.

Vihtakari, M. (2024) 'ggOceanMaps: plot data on oceanographic maps using "ggplot2"'. R package Version 2.2.0. Available at: https://mikkovihtakari.github.io/ggOceanMaps/.

Villegas-Plazas, M. *et al.* (2019) 'Variations in microbial diversity and metabolite profiles of the tropical marine sponge *Xestospongia muta* with season and depth', *Microbial Ecology*, 78(1), pp. 243–256. Available at: https://doi.org/10.1007/s00248-018-1285-y.

Vinod, H.D. (1969) 'Integer programming and the theory of grouping', *Journal of the American Statistical Association*, 64(326), pp. 506–519. Available at: https://doi.org/10.1080/01621459.1969.10500990.

Vipindas, P.V. *et al.* (2023) 'Water mass controlled vertical stratification of bacterial and archaeal communities in the Western Arctic Ocean during summer sea-ice melting', *Microbial Ecology*, 85(4), pp. 1150–1163. Available at: https://doi.org/10.1007/s00248-022-01992-z.

Waleron, M. *et al.* (2007) 'Allochthonous inputs of riverine picocyanobacteria to coastal waters in the Arctic Ocean', *FEMS Microbiology Ecology*, 59(2), pp. 356–365. Available at: https://doi.org/10.1111/j.1574-6941.2006.00236.x.

Walesiak, M. and Dudek, A. (2020) 'The choice of variable normalization method in cluster analysis', in K.S. Soliman (ed.) *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges*. International Business Information Management Association (IBIMA), pp. 325–340. Available at:

Walsh, D.A., Zaikova, E. and Hallam, S.J. (2009) 'Large Volume (20L+) Filtration of Coastal Seawater Samples', *Journal of Visualized Experiments* [Preprint], (28). Available at: https://doi.org/10.3791/1161.

Wang, Q. *et al.* (2007) 'Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy', *Applied and Environmental Microbiology*, 73(16), pp. 5261–5267. Available at: https://doi.org/10.1128/AEM.00062-07.

Wang, S. *et al.* (2022) 'Microbial richness of marine biofilms revealed by sequencing full-length 16S rRNA genes', *Genes*, 13(6), p. 1050. Available at: https://doi.org/10.3390/genes13061050.

Wang, Y. *et al.* (2017) 'Quantifying the importance of the rare biosphere for microbial community response to organic pollutants in a freshwater ecosystem', *Applied and Environmental Microbiology*, 83(8), pp. e03321-16. Available at: https://doi.org/10.1128/AEM.03321-16.

Wang, Y. *et al.* (2020) 'Biogeographic pattern of bacterioplanktonic community and potential function in the Yangtze River: roles of abundant and rare taxa', *Science of the Total Environment*, 747, p. 141335. Available at: https://doi.org/10.1016/j.scitotenv.2020.141335.

Ward, D.M., Weller, R. and Bateson, M.M. (1990) '16S rRNA sequences reveal numerous uncultured microorganisms in a natural community', *Nature*, 345(6270), pp. 63-65. Available at: https://doi.org/10.1038/345063a0.

Webb, C.O. (2000) 'Exploring the phylogenetic structure of ecological communities: an example for rain forest trees', *The American Naturalist*, 156(2), pp. 145–155. Available at: https://doi.org/10.1086/303378.

Webb, C.O. *et al.* (2002) 'Phylogenies and community ecology', *Annual Review of Ecology and Systematics*, 33(1), pp. 475–505. Available at: https://doi.org/10.1146/annurev.ecolsys.33.010802.150448.

Webster, N.S. *et al.* (2009) 'Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts', *Environmental Microbiology*, 12(8), pp. 2070–2082. Available at: https://doi.org/10.1111/j.1462-2920.2009.02065.x.

Webster, N.S. and Thomas, T. (2016) 'The sponge hologenome', *mBio*, 7(2), pp. e00135-16. Available at: https://doi.org/10.1128/mBio.00135-16.

Wei, S.T.-S. *et al.* (2018) 'Microbial functional responses to cholesterol catabolism in denitrifying sludge', *mSystems*, 3(5), pp. 1–19. Available at: https://doi.org/10.1128/msystems.00113-18.

Weisse, T. (2014) 'Ciliates and the rare biosphere – Community ecology and population dynamics', *Journal of Eukaryotic Microbiology*, 61(4), pp. 419–433. Available at: https://doi.org/10.1111/jeu.12123.

Weitz, J.S. and Dushoff, J. (2008) 'Alternative stable states in host – Phage dynamics', *Theoretical Ecology*, 1(1), pp. 13–19. Available at: https://doi.org/10.1007/s12080-007-0001-1.

Wickham, H. (2014) 'Tidy Data', *Journal of Statistical Software*, 59(10), pp. 1–23. Available at: http://www.jstatsoft.org/.

Wickham, H. (2016) 'ggplot2: elegant graphics for data analysis'. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Wickham, H. *et al.* (2022) 'devtools: Tools to Make Developing R Packages Easier'. Available at: https://devtools.r-lib.org/.

Wickham, H. *et al.* (2023) 'dplyr: A Grammar of Data Manipulation'. R package version 1.1.4. Available at: https://dplyr.tidyverse.org.

Wickham, H. (2023) 'stringr: Simple, Consistent Wrappers for Common String Operations'. R package version 1.5.1. Available at: https://cran.r-project.org/web/packages/stringr/index.html.

Wickham, H., Henry, L. and RStudio (2023) 'purrr: Functional Programming Tools'. R package version 1.0.2. Available at: https://cran.r-project.org/web/packages/purrr/index.html.

Wickham, H., Vaughan, D. and Girlich, M. (2023) 'tidyr: Tidy Messy Data'. R package version 1.3.0. Available at: https://tidyr.tidyverse.org.

Wiencke, C. and Hop, H. (2016) 'Ecosystem Kongsfjorden: new views after more than a decade of research', *Polar Biology*, 39(10), pp. 1679–1687. Available at: https://doi.org/10.1007/s00300-016-2032-9.

Wietz, M. *et al.* (2024) 'The Arctic summer microbiome across Fram Strait: depth, longitude, and substrate concentrations structure microbial diversity in the euphotic zone', *Environmental Microbiology*, 26(2), pp. 1–16. Available at: https://doi.org/10.1111/1462-2920.16568.

Wilk, C.O. (2019) *Fundamentals of Data Visualization*. O'Reilly Media. Available at: https://clauswilke.com/dataviz.

Wilson, B. *et al.* (2017) 'Changes in marine prokaryote composition with season and depth over an Arctic polar year', *Frontiers in Marine Science*, 4(95), p. 95. Available at: https://doi.org/10.3389/fmars.2017.00095.

Woese, C.R. and Fox, G.E. (1977) 'Phylogenetic structure of the prokaryotic domain: The primary kingdoms', *Proceedings of the National Academy of Sciences*, 74(11), pp. 5088–5090. Available at: https://doi.org/10.1073/pnas.74.11.5088.

Wold, A., Assmy, P. and Duarte, P. (2024) 'Biogeochemistry data from Kongsfjorden and Rijpfjorden 2011 to 2020'. Norwegian Polar Institute, Dataset. Available at: https://doi.org/10.21334/npolar.2024.4d4de169.

Wu, J., Hua, Z. and Gu, L. (2022) 'Per-, poly-fluoroalkyl substances (PFASs) and planktonic microbiomes: identification of biotic and abiotic regulations in community coalescence and food webs', *Environmental Pollution*, 302(119078), pp. 1-14. Available at: https://doi.org/10.1016/j.envpol.2022.119078.

Wu, W. *et al.* (2017) 'Abundant and rare picoeukaryotic sub-communities present contrasting patterns in the epipelagic waters of marginal seas in the northwestern Pacific Ocean', *Environmental Microbiology*, 19(1), pp. 287–300. Available at: https://doi.org/10.1111/1462-2920.13606.

Xu, Q. *et al.* (2022) 'Rare bacteria assembly in soils is mainly driven by deterministic processes', *Microbial Ecology*, 83(1), pp. 137–150. Available at: https://doi.org/10.1007/s00248-021-01741-8.

Xue, Y. *et al.* (2018) 'Distinct patterns and processes of abundant and rare eukaryotic plankton communities following a reservoir cyanobacterial bloom', *The ISME Journal*, 12(9), pp. 2263–2277. Available at: https://doi.org/10.1038/s41396-018-0159-0.

Yadav, S. *et al.* (2021) 'The physiology and metabolic properties of a novel, low-abundance *Psychrilyobacter* species isolated from the anoxic Black Sea shed light on its ecological role', *Environmental Microbiology Reports*, 13(6), pp. 899–910. Available at: https://doi.org/10.1111/1758-2229.13012.

Yamamoto, K. *et al.* (2019) 'Diversity and geochemical community assembly processes of the living rare biosphere in a sand-and-gravel aquifer ecosystem in the Midwestern United States', *Scientific Reports*, 9(1), pp. 1–11. Available at: https://doi.org/10.1038/s41598-019-49996-z.

Yan, K. *et al.* (2023) 'Abundant fungi dominate the complexity of microbial networks in soil of contaminated site: high-precision community analysis by full-length sequencing', *Science of The Total Environment*, 861(160563), pp. 1-11. Available at: https://doi.org/10.1016/j.scitotenv.2022.160563.

Yang, B., Wang, Y. and Qian, P.-Y. (2016) 'Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis', *BMC Bioinformatics*, 17(1), p. 135. Available at: https://doi.org/10.1186/s12859-016-0992-y.

Yang, F., Li, Q. and Yin, X. (2024) 'Metagenomic analysis of the effects of salinity on microbial community and functional gene diversity in glacial meltwater estuary, Ny-Alesund, Arctic', *Brazilian Journal of Microbiology*, 55(2), pp. 1587–1599. Available at: https://doi.org/10.1007/s42770-024-01298-x.

Yang, S. *et al.* (2017) 'Community structure of rare methanogenic archaea: insight from a single functional group', *FEMS Microbiology Ecology*, 93(11), pp. 1–10. Available at: https://doi.org/10.1093/femsec/fix126.

Yang, S. *et al.* (2024) 'Rare subcommunity maintains the stability of ecosystem multifunctionality by deterministic assembly processes in subtropical estuaries', *Frontiers in Microbiology*, 15(1365546), pp. 1–12. Available at: https://doi.org/10.3389/fmicb.2024.1365546.

Yeh, Y. *et al.* (2021) 'Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S', *Environmental Microbiology*, 23(6), pp. 3240–3250. Available at: https://doi.org/10.1111/1462-2920.15553.

Yergeau, E. *et al.* (2017) 'Metagenomic survey of the taxonomic and functional microbial communities of seawater and sea ice from the Canadian Arctic', *Scientific Reports*, 7(42242), pp. 1–10. Available at: https://doi.org/10.1038/srep42242.

Yilmaz, P. *et al.* (2014) 'The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks', *Nucleic Acids Research*, 42(D1), pp. D643–D648. Available at: https://doi.org/10.1093/nar/gkt1209.

Yin, Z.-Y. *et al.* (2018) 'Examining the patterns and dynamics of species abundance distributions in succession of forest communities by model selection', *PLoS ONE*, 13(5), p. e0196898. Available at: https://doi.org/10.1371/journal.pone.0196898.

Youssef, N., Steidley, B.L. and Elshahed, M.S. (2012) 'Novel high-rank phylogenetic lineages within a sulfur spring (Zodletone Spring, Oklahoma), revealed using a combined pyrosequencing-sanger approach', *Applied and Environmental Microbiology*, 78(8), pp. 2677–2688. Available at: https://doi.org/10.1128/AEM.00002-12.

Youssef, N.H., Ashlock-Savage, K.N. and Elshahed, M.S. (2012) 'Phylogenetic diversities and community structure of members of the extremely halophilic Archaea (order *Halobacteriales*) in multiple saline sediment habitats', *Applied and Environmental Microbiology*, 78(5), pp. 1332-44. Available at: https://doi.org/10.1128/AEM.07420-11.

Youssef, N.H., Couger, M.B. and Elshahed, M.S. (2010) 'Fine-Scale Bacterial Beta Diversity within a Complex Ecosystem (Zodletone Spring, OK, USA): The Role of the Rare Biosphere', *PLoS ONE*, 5(8), p. e12414. Available at: https://doi.org/10.1371/journal.pone.0012414.

Yu, T. *et al.* (2022) 'Effects of waterlogging on soybean rhizosphere bacterial community using V4, LoopSeq, and PacBio 16S rRNA Sequence', *Microbiology Spectrum*, 10(1), pp. e02011-21. Available at: https://doi.org/10.1128/spectrum.02011-21.

Yuan, C. *et al.* (2015) 'Reconstructing 16S rRNA genes in metagenomic data', *Bioinformatics*, 31(12), pp. i35–i43. Available at: https://doi.org/10.1093/bioinformatics/btv231.

Zehavi, T., Probst, M. and Mizrahi, I. (2018) 'Insights into culturomics of the rumen microbiome', *Frontiers in Microbiology*, 9(1999), pp. 1–10. Available at: https://doi.org/10.3389/fmicb.2018.01999.

Zhang, H. *et al.* (2020) 'Interaction and assembly processes of abundant and rare microbial communities during a diatom bloom process', *Environmental Microbiology*, 22(5), pp. 1707–1719. Available at: https://doi.org/10.1111/1462-2920.14820.

Zhang, N. *et al.* (2021) 'Divergent temporal response of abundant and rare bacterial communities to transient *Escherichia coli* O157:H7 invasion', *Frontiers in Microbiology*, 12(665380), pp. 1–11. Available at: https://doi.org/10.3389/fmicb.2021.665380.

Zhang, W. *et al.* (2018) 'The diversity and biogeography of abundant and rare intertidal marine microeukaryotes explained by environment and dispersal limitation', *Environmental Microbiology*, 20(2), pp. 462–476. Available at: https://doi.org/10.1111/1462-2920.13916.

Zhang, X. *et al.* (2021) 'Bacterial community variations with salinity in the saltwater-intruded estuarine aquifer', *Science of the Total Environment*, 755(142423), pp. 1-11. Available at: https://doi.org/10.1016/j.scitotenv.2020.142423.

Zhang, Y. *et al.* (2018) 'Abundant and rare microbial biospheres respond differently to environmental and spatial factors in Tibetan hot springs', *Frontiers in Microbiology*, 9(2096), pp. 1–16. Available at: https://doi.org/10.3389/fmicb.2018.02096.

Zhang, Y. *et al.* (2021) 'Vertical diversity and association pattern of total, abundant and rare microbial communities in deep-sea sediments', *Molecular Ecology*, 30(12), pp. 2800–2816. Available at: https://doi.org/10.1111/mec.15937.

Zhang, Z. *et al.* (2000) 'A greedy algorithm for aligning DNA sequences', *Journal of Computational Biology*, 7(1-2), pp. 203-214. Available at: https://doi.org/10.1089/10665270050081478.

Zhou, J. and Ning, D. (2017) 'Stochastic community assembly: does it matter in microbial ecology?', *Microbiology and Molecular Biology Reviews*, 81(4), pp. e00002-17. Available at: https://doi.org/10.1128/MMBR.00002-17.

Zhu, J. *et al.* (2018) 'Revealing the anaerobic acclimation of microbial community in a membrane bioreactor for coking wastewater treatment by Illumina Miseq sequencing', *Journal of Environmental Sciences*, 64, pp. 139–148. Available at: https://doi.org/10.1016/j.jes.2017.06.003.

Zhu, M. *et al.* (2023) 'Deciphering the distinct successional patterns and potential roles of abundant and rare microbial taxa of urban riverine plastisphere', *Journal of Hazardous Materials*, 450(131080), pp. 1-14. Available at: https://doi.org/10.1016/j.jhazmat.2023.131080.

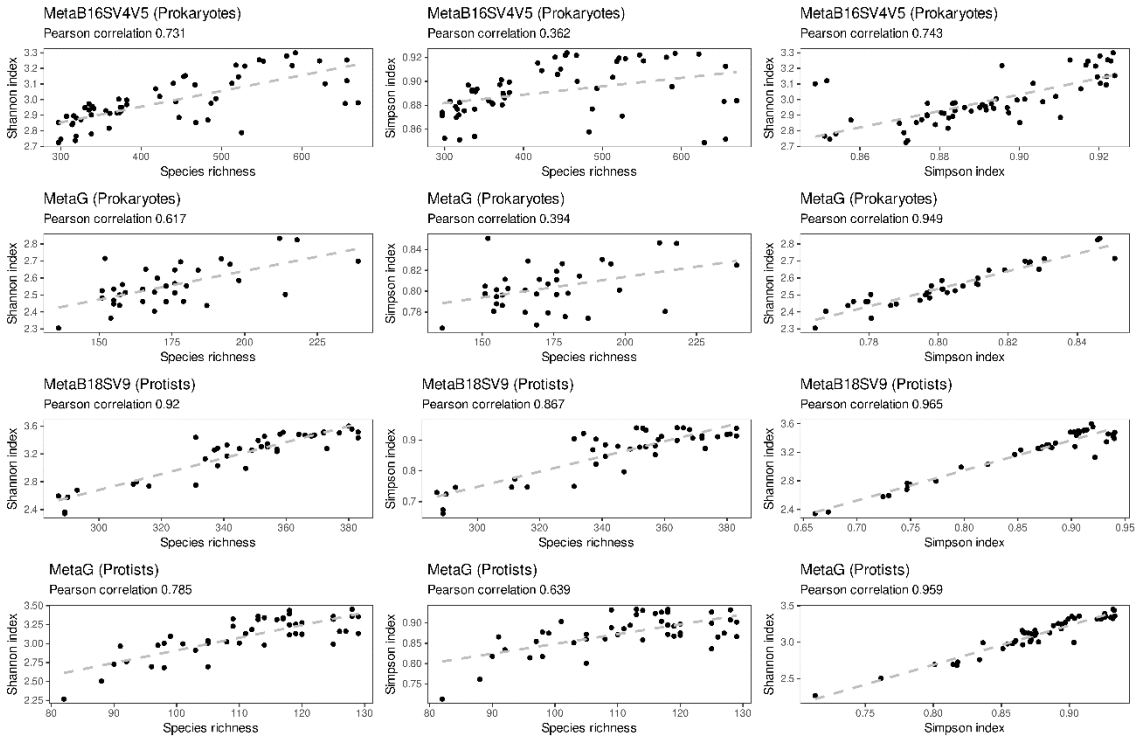
Ziegler, M. *et al.* (2018) 'Rare symbionts may contribute to the resilience of coral-algal assemblages', *ISME Journal*, 12(1), pp. 161–172. Available at: https://doi.org/10.1038/ismej.2017.151.

Zilber-Rosenberg, I. and Rosenberg, E. (2008) 'Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution', *FEMS Microbiology Reviews*, 32(5), pp. 732-735. Available at: https://doi.org/10.1111/j.1574-6976.2008.00123.x.

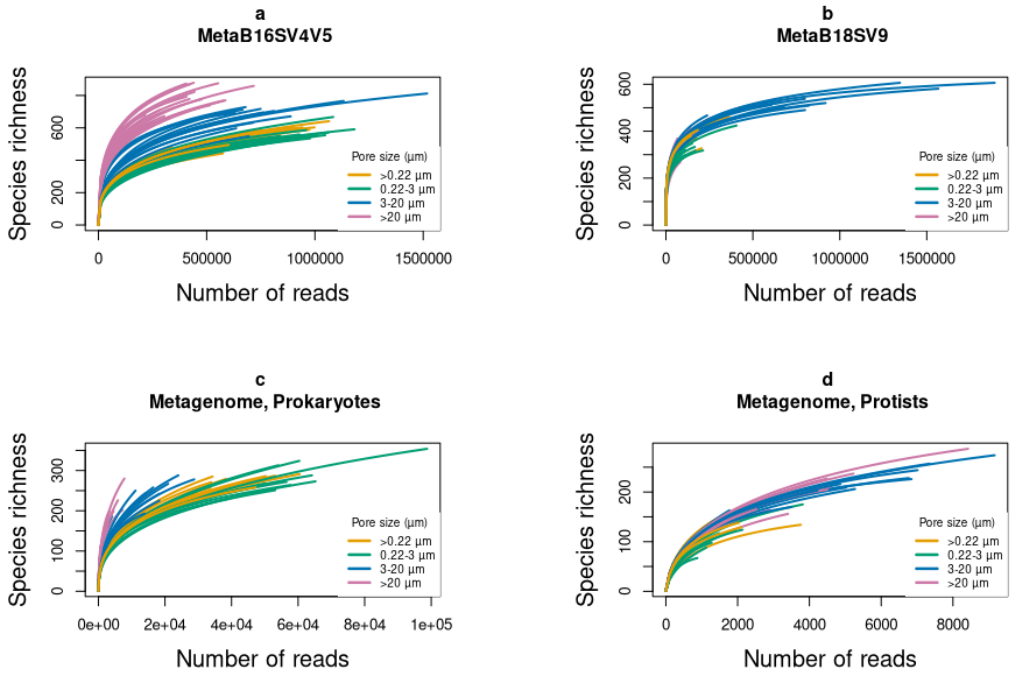
Attachments

10. Supplementary Figures

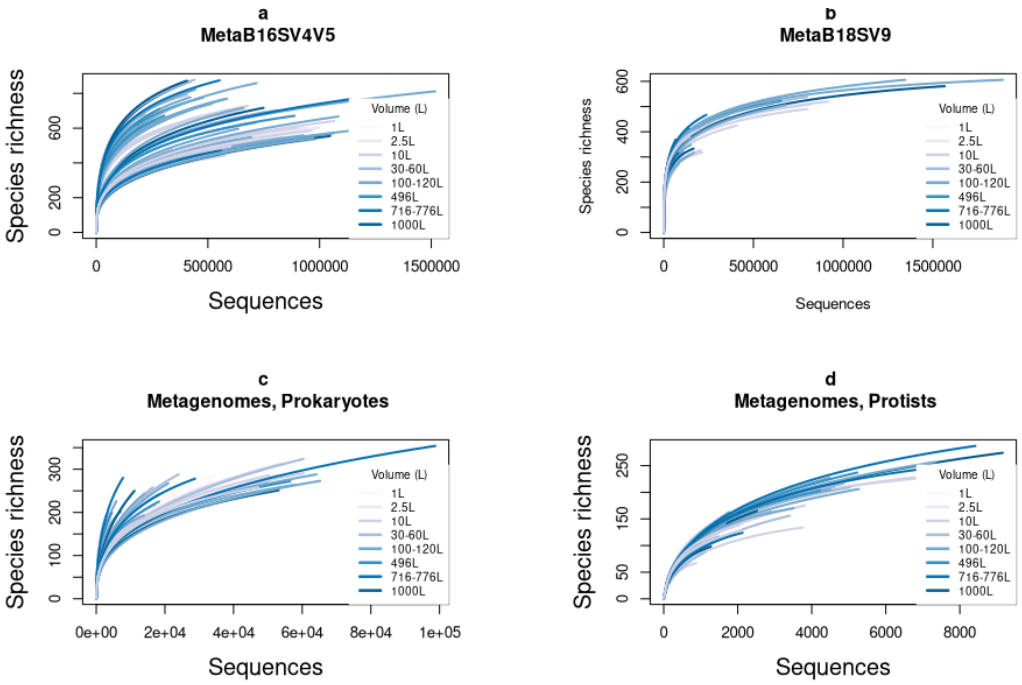
10.1. Supplementary Figures of Chapter 3



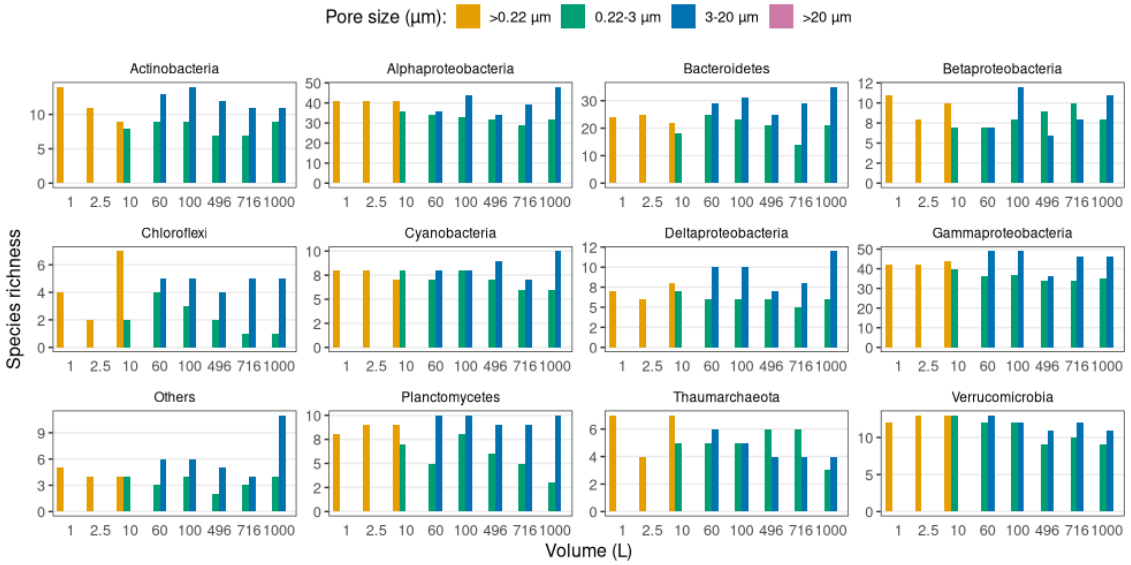
Supplementary Figure 10.1. Pairwise Pearson correlation between alpha metrics, for all sequencing strategies tested here. The metrics that were compared against species richness were Shannon index and Simpson index, for the following combinations: MetaB16SV5V5 (Prokaryotes), MetaB18SV9 (Protists) and metagenomes (MetaG: Prokaryotes and Protists). Tendency lines were added to help visualization.



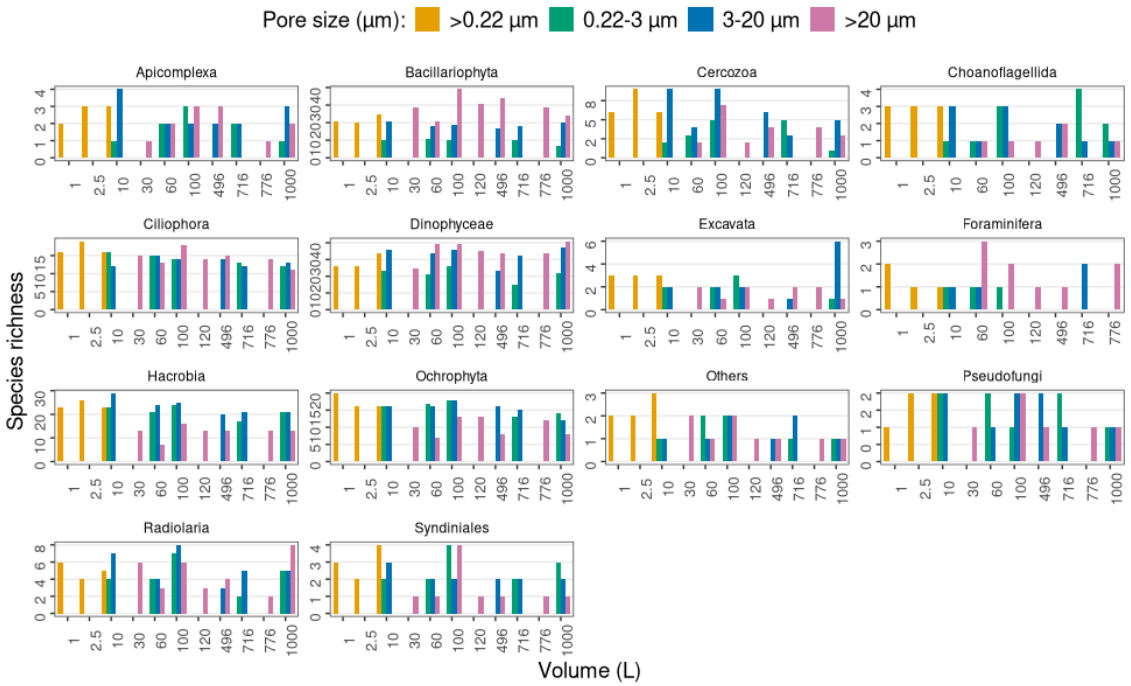
Supplementary Figure 10.2. Rarefaction curves by size fraction. Each grid represents a methodological combination, (a) for MetaB16SV4V5 (Prokaryotes), (b) for MetaB18SV9 (Protists), (c) Metagenomes (Prokaryotes), and (d) Metagenomes (Prokaryotes). Colors illustrate different size fractions.



Supplementary Figure 10.3. Rarefaction curves by filtered volume. Each grid represents a methodological combination, (a) for MetaB16SV4V5 (Prokaryotes), (b) for MetaB18SV9 (Protists), (c) Metagenomes (Prokaryotes), and (d) Metagenomes (Prokaryotes). Colors illustrate different volumes.

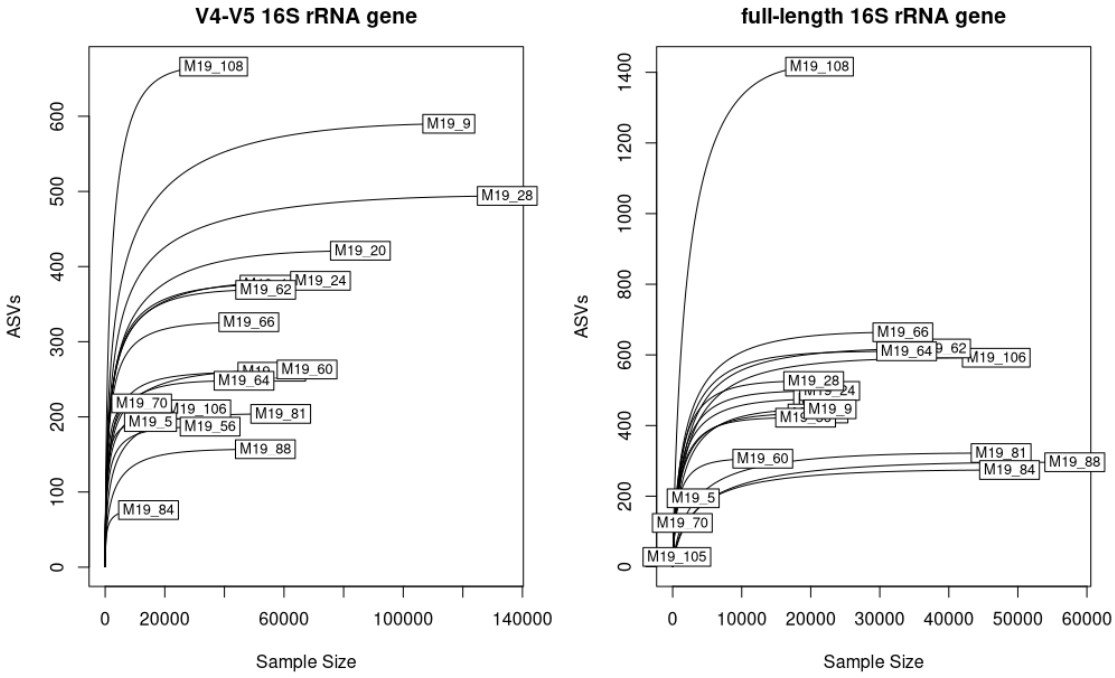


Supplementary Figure 10.4. Prokaryotic species richness per taxonomic group, from metagenomes. Each panel represents the species richness of a specific prokaryotic phyla or class for each volume (1–1000 L). Bar plots indicate species richness and are colored by pore size. The taxonomic group called “Others” includes all phyla that did not reach, at least, 100 taxonomic lineages across all samples, to avoid an excessive amount of uninformative, redundant panels.

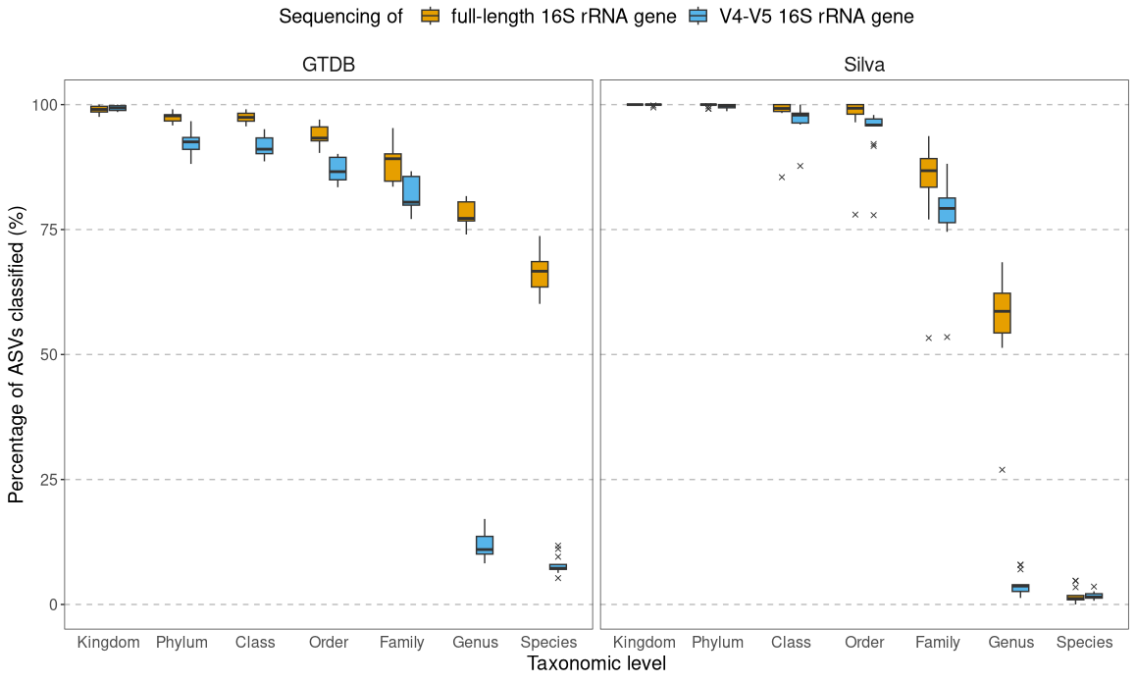


Supplementary Figure 10.5. Protist species richness per taxonomic group, from MetaB18SV9. Each panel represents the species richness of a specific protist group for each volume (1–1000 L). Bar plots indicate species richness and are colored by pore size. Selected taxonomic groups follow a “fake rank” manually curated to highlight groups of interest, with the less representative groups merged into “Others” designation.

10.2. Supplementary Figures of Chapter 4

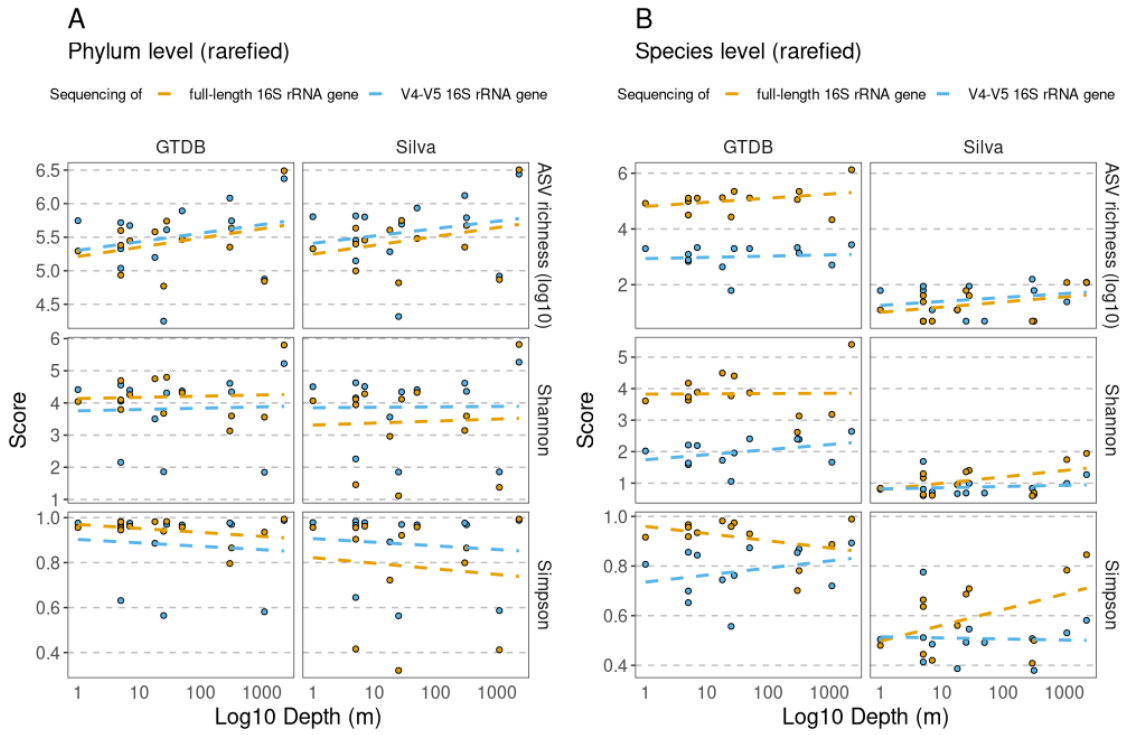


Supplementary Figure 10.6. Rarefaction curves for all samples from V4V5 (left) and full-length (right) 16S rRNA gene amplicon sequencing.

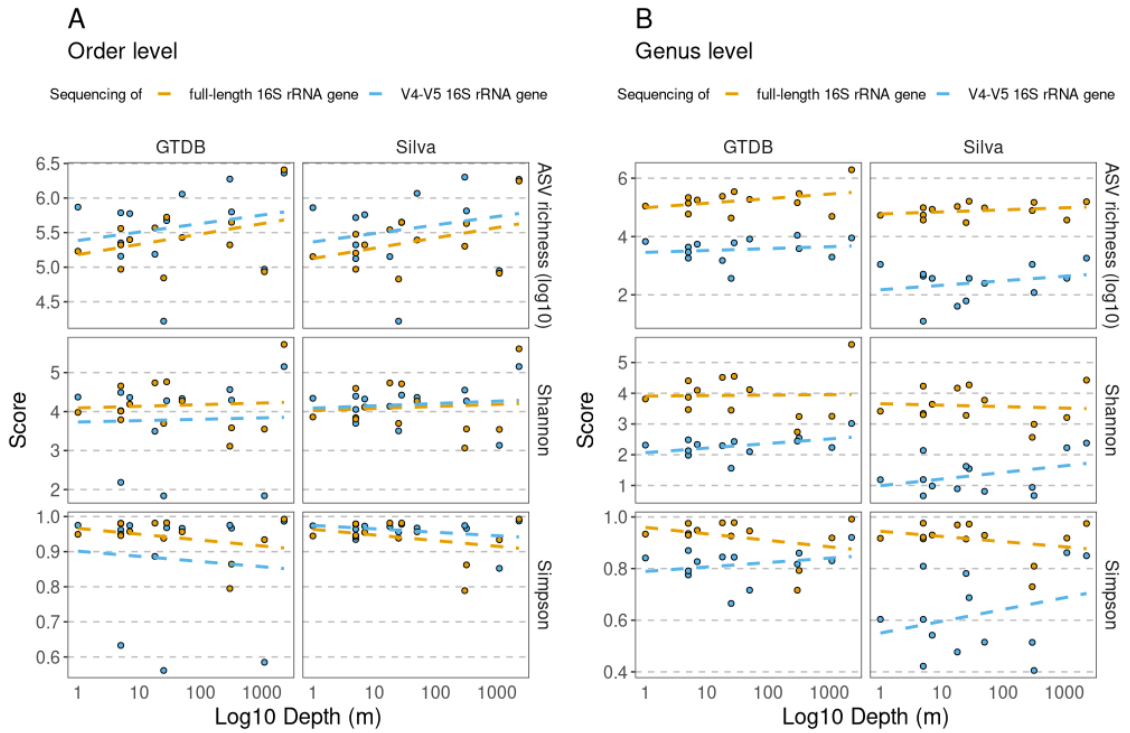


Supplementary Figure 10.7. Rarefied version of Figure 4.2, with percentage of ASVs classified at each taxonomic level. The left panel shows results for the GTDB database, and the right panel shows results for the Silva database. For each database, full-length and V4V5 16S rRNA gene sequencing were compared (orange and blue, respectively). Boxplots were used to illustrate centrality metrics of the percentage of ASVs classified, including crosses representing the outliers

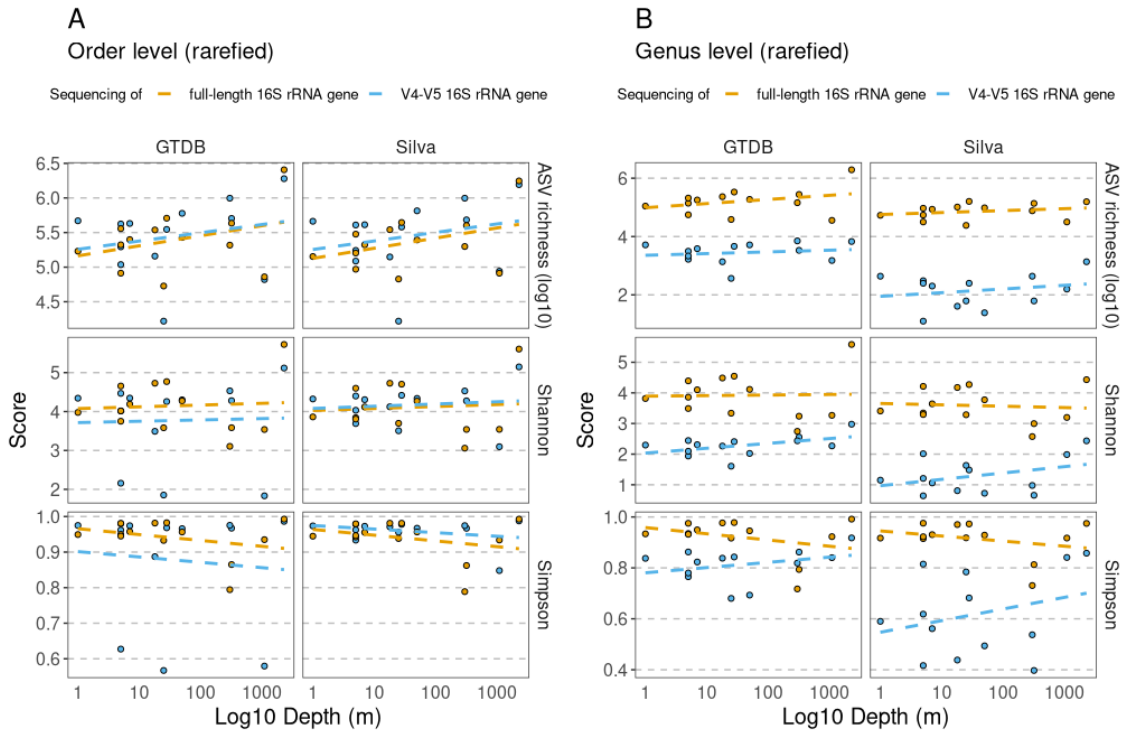
and were calculated with 13 observations, per independent group. Note that the “Kingdom” taxonomic level was maintained to keep consistency with the terminology used by the reference databases used.



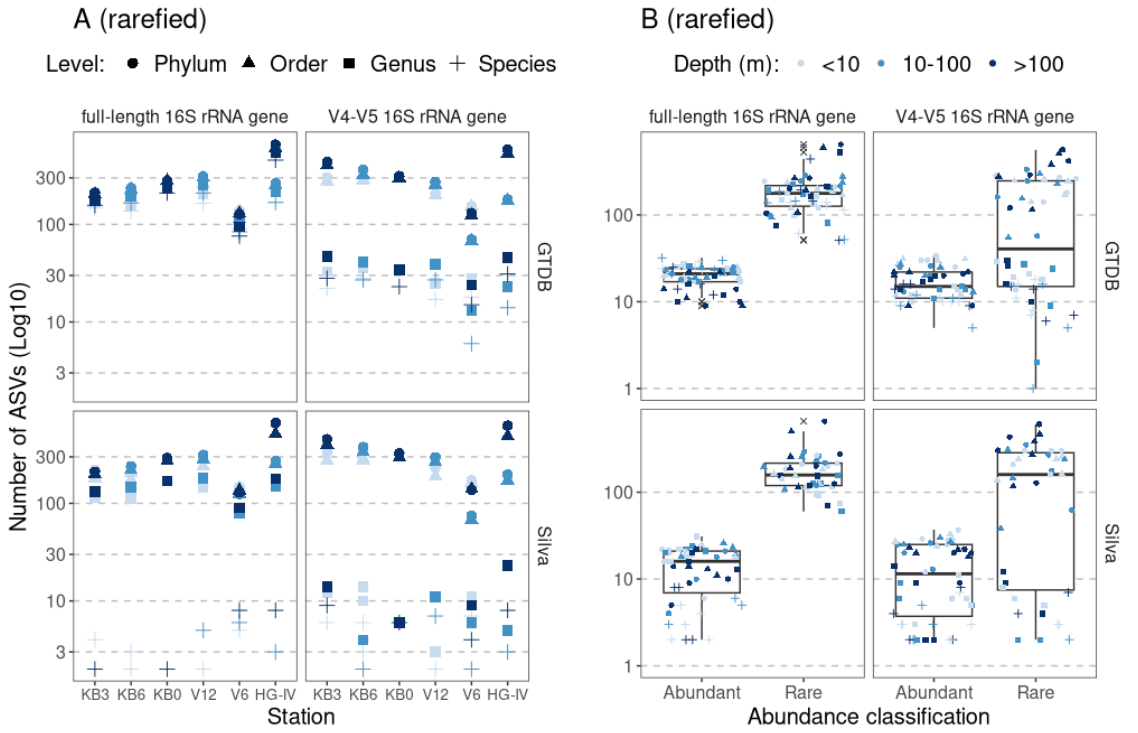
Supplementary Figure 10.8. Rarefied version of Figure 4.4, with alpha diversity scores across depth. ASVs were filtered (A) at phylum level and (B) at species level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help read the figure. Regression equations and statistical support are available in Supplementary Table 11.15.



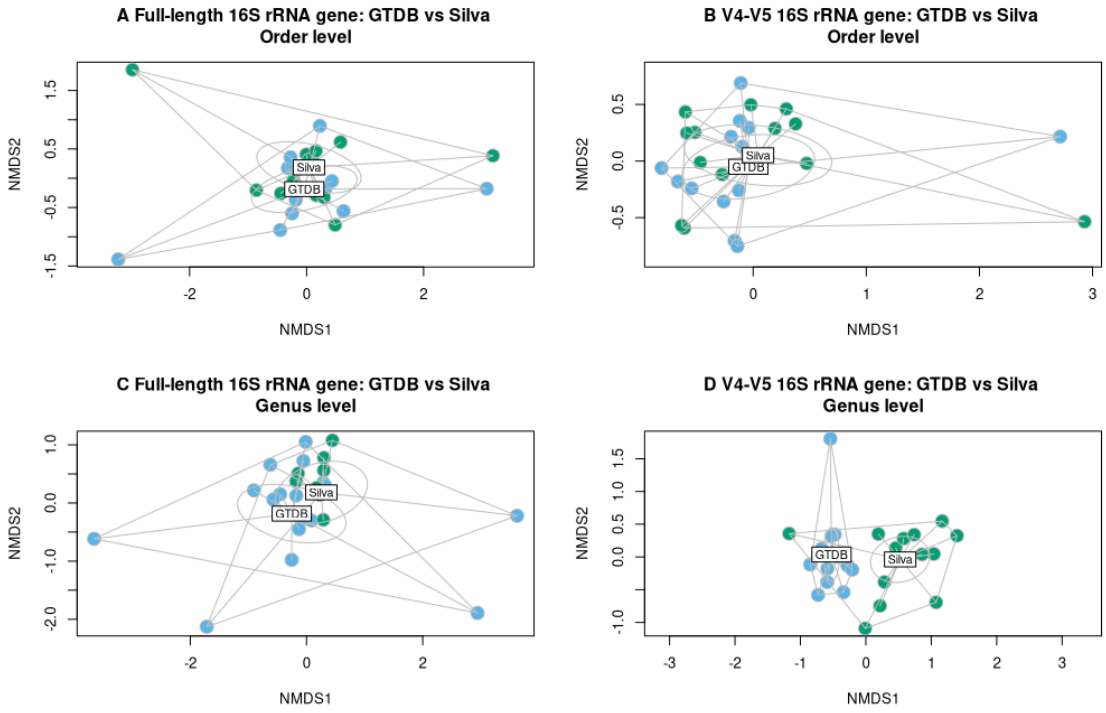
Supplementary Figure 10.9. Alpha diversity scores across depth. ASVs were filtered (A) at order level and (B) at genus level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help read the figure. Regression equations and statistical support are available in Supplementary Table 11.15.



Supplementary Figure 10.10. Rarefied version of Supplementary Figure 4.4, with alpha diversity scores across depth. ASVs were filtered (A) at order level and (B) at genus level. For each specified taxonomic level, only the ASVs that got a taxonomic classification were used. The alpha diversity scores used were the number of ASVs (same as ASV richness), the Shannon index and the Simpson index. Orange was used for full-length and blue for V4V516S rRNA gene sequencing. Columns in the facet grid were used to distinguish taxonomy databases (GTDB and Silva). Tendency lines were added to help read the figure. Regression equations and statistical support are available in Supplementary Table 11.15.

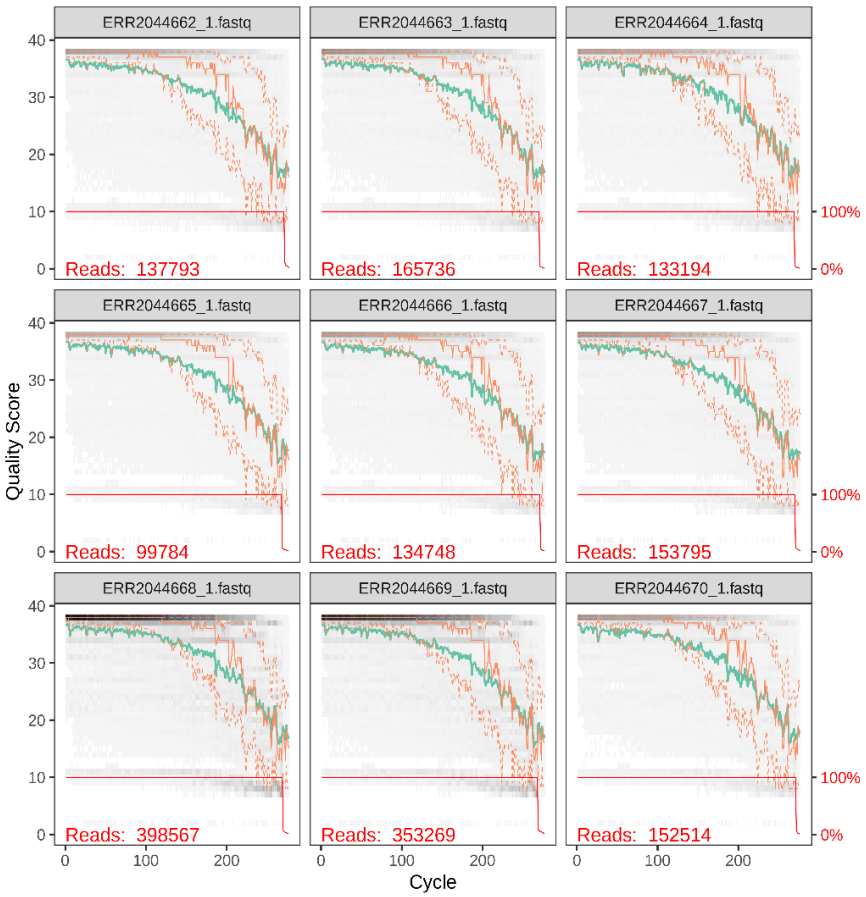


Supplementary Figure 10.11. Rarefied version of Figure 4.5, with ASV richness at phylum, order, genus and species taxonomic levels across methodological combinations. **A**. ASV richness by sampling Station; **B**. ASV richness by abundance classification, with rare ASVs defined as ASVs with less than 0.1% relative abundance, per sample. For both (A) and (B), different taxonomic levels were represented by different shape points. The ASV richness was illustrated in Log10 scale so that lower values could be distinguished; each panel further divided itself into four panes, allowing direct comparison of full-length and V4V5 16S rRNA gene sequencing (columns) and Silva or GTDB databases (rows). In (A) the points were illustrated without centrality metrics, because each station included less than five distinct observations, while in (B) a centrality metric was used with boxplots, because the number of observations (n = 13 per independent group) allowed it to do so. Nevertheless, the original points were plotted on top of the boxplots, so that it was possible to see the direct connection between (A) and (B). Additionally, note that some combinations presented less than 13 observations, which reflected samples without any species classified. The points were colored in different shades of blue, divided by depth: Surface (< 10 m); Middle (10-100 m); and Deep (> 100 m).

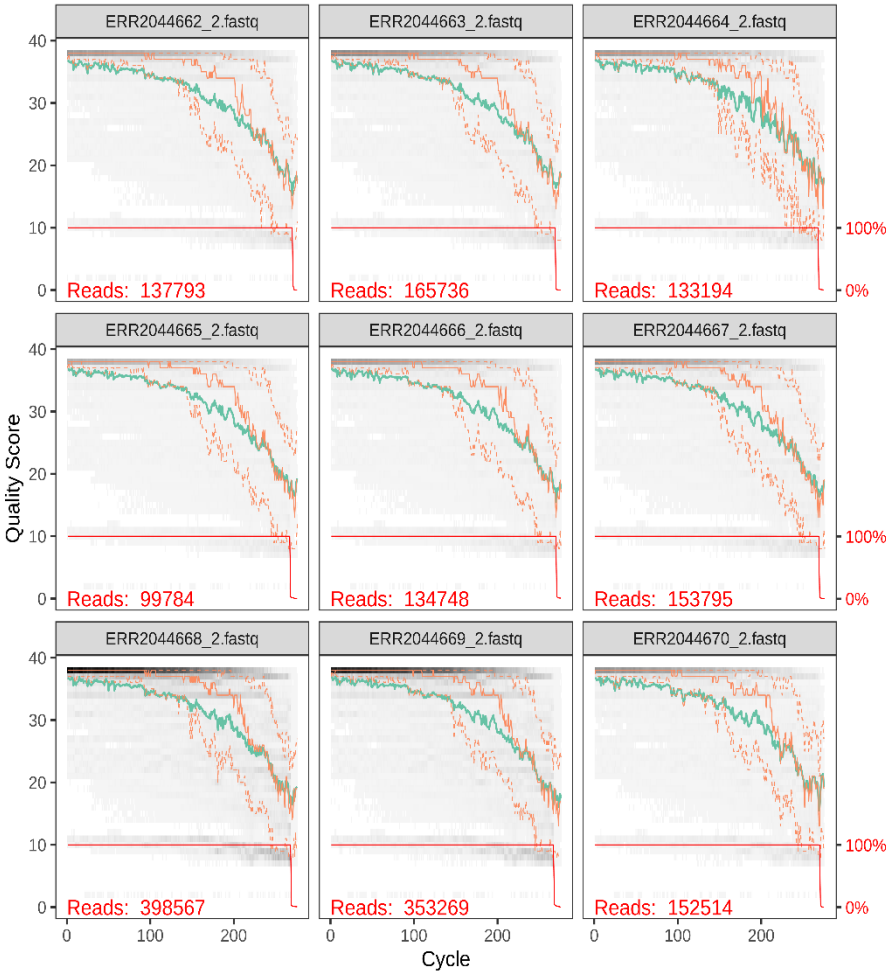


Supplementary Figure 10.12. Beta diversity at order and genus level. Community composition between Silva and GTDB databases for: **A**. Full-length 16S rRNA gene, at order level; **B**. V4V5 region of the 16S rRNA gene, at order level; **C**. full-length 16S rRNA gene, at genus level; **D**. V4V5 region of the 16S rRNA gene, at genus level. Bray-Curtis distances were illustrated with nMDS ordination. Shapes and color were used to illustrate sampling stations and depth, according to the figure label.

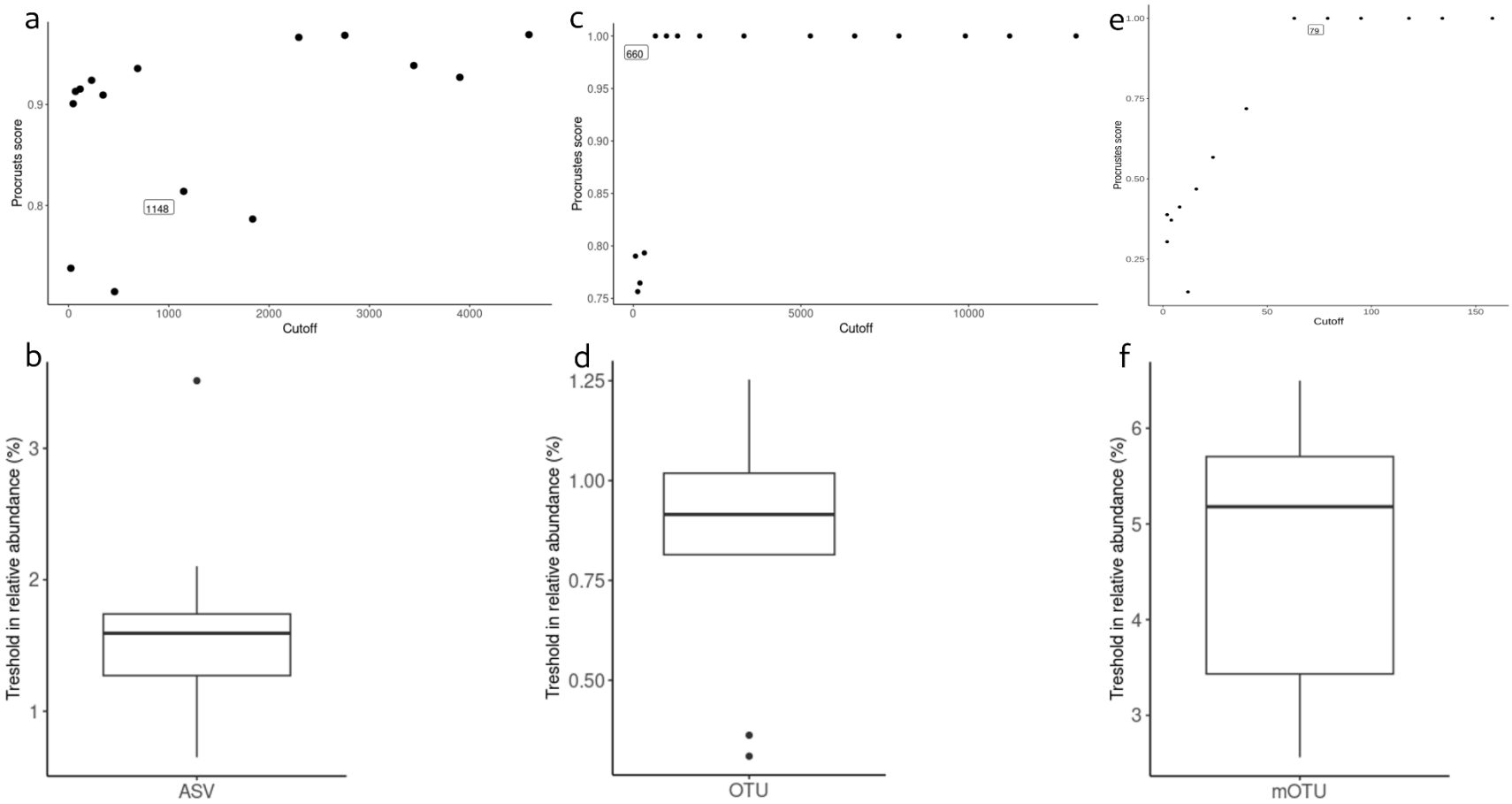
10.3. Supplementary Figures of Chapter 5



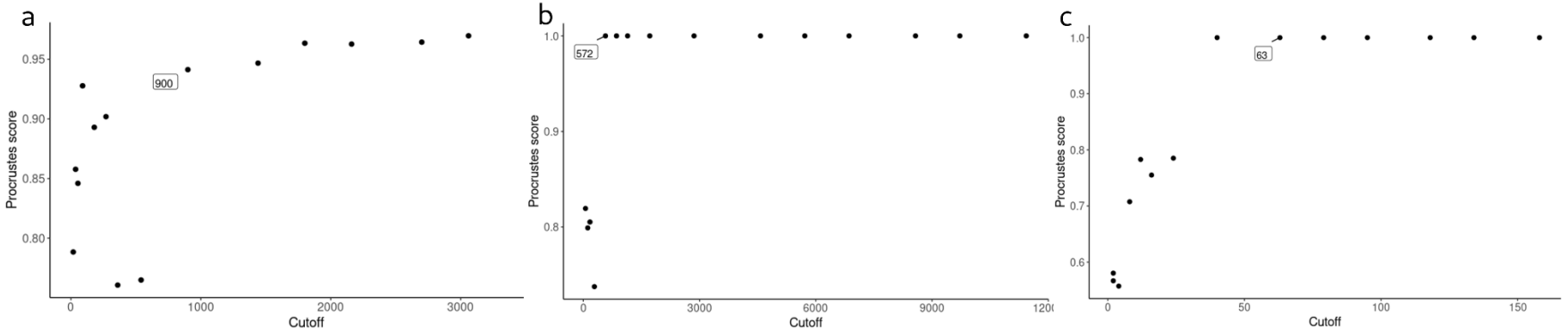
Supplementary Figure 10.13. Quality scores profile for the forward 16S rRNA gene amplicon sequencing reads of N-ICE 2015.



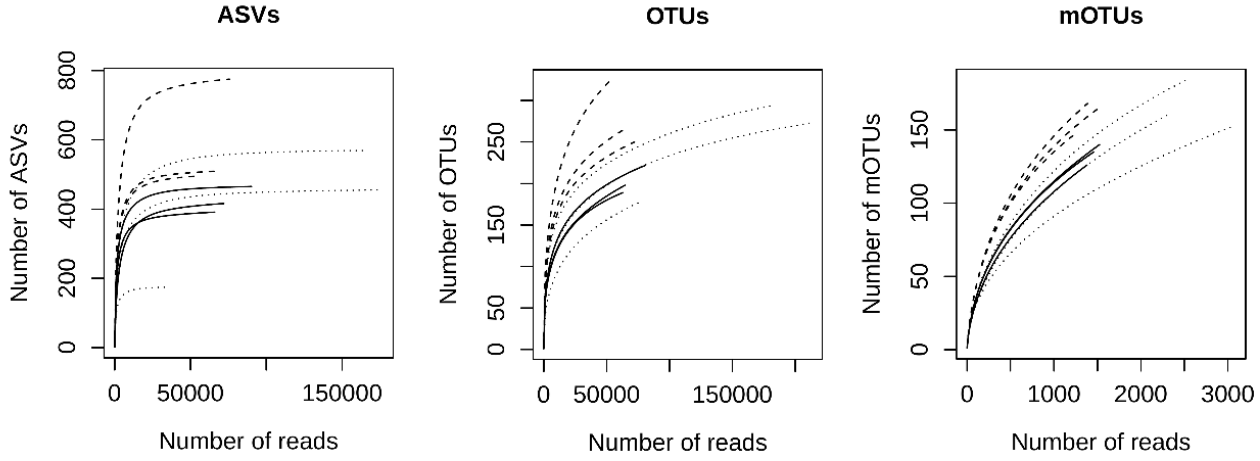
Supplementary Figure 10.14. Quality scores profile for the reverse 16S rRNA gene amplicon sequencing reads of N-ICE 2015.



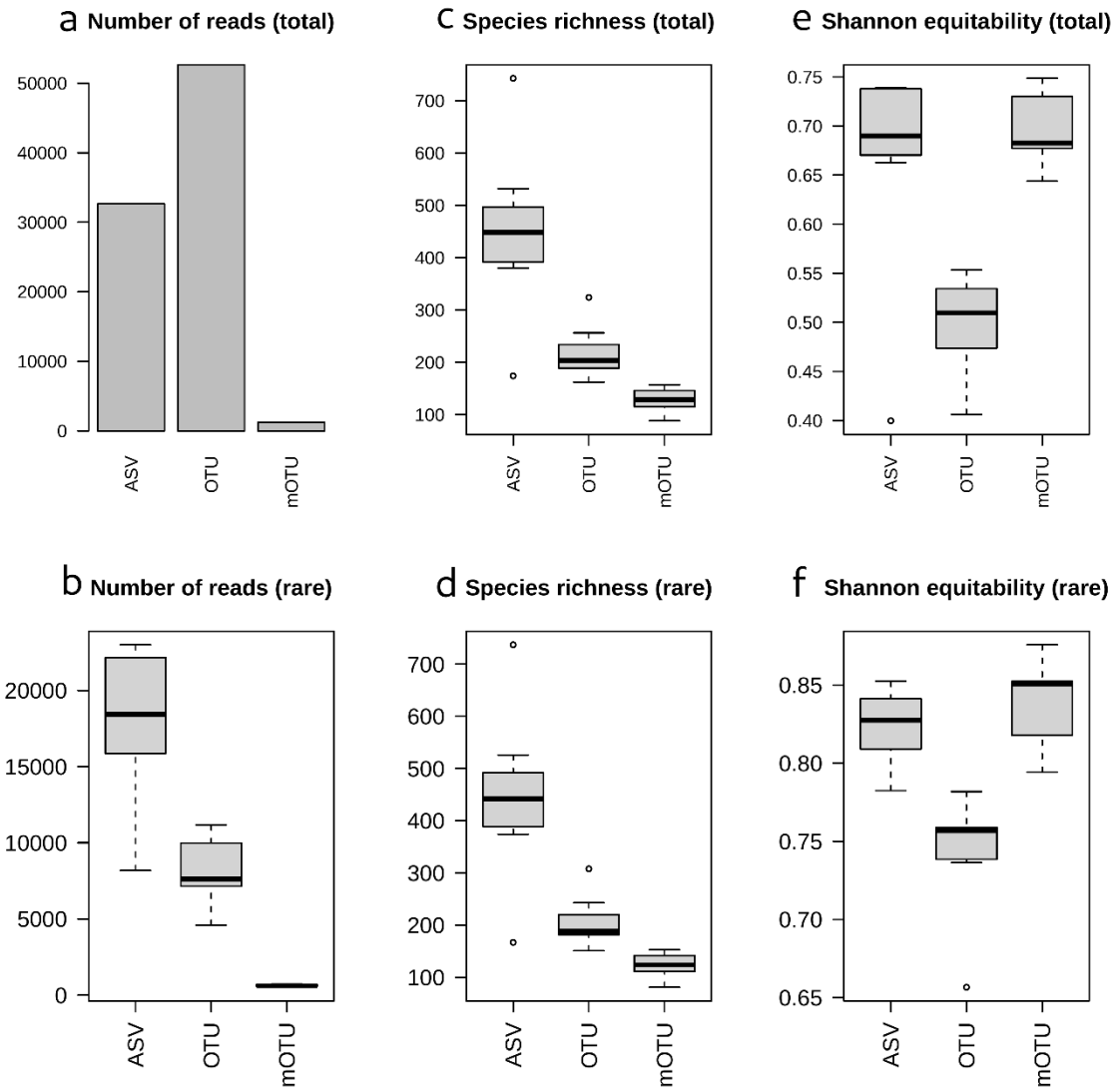
Supplementary Figure 10.15. MultiCoLA results of Procrustes and boxplot with the relative abundance equivalent of the selected cutoff, for non-rarefied data; **a.** Procrustes results for the ASVs; **b.** Boxplots with relative abundance equivalent for each sample, after application of the selected threshold, for the ASVs; **c.** Procrustes results for the OTUs; **d.** Boxplots with relative abundance equivalent for each sample, after application of the selected threshold, for the OTUs; **e.** Procrustes results for the mOTUs; **f.** Boxplots with relative abundance equivalent for each sample, after application of the selected threshold, for the mOTUs. Boxplots in (a), (d), and (f) have n = 9 samples.



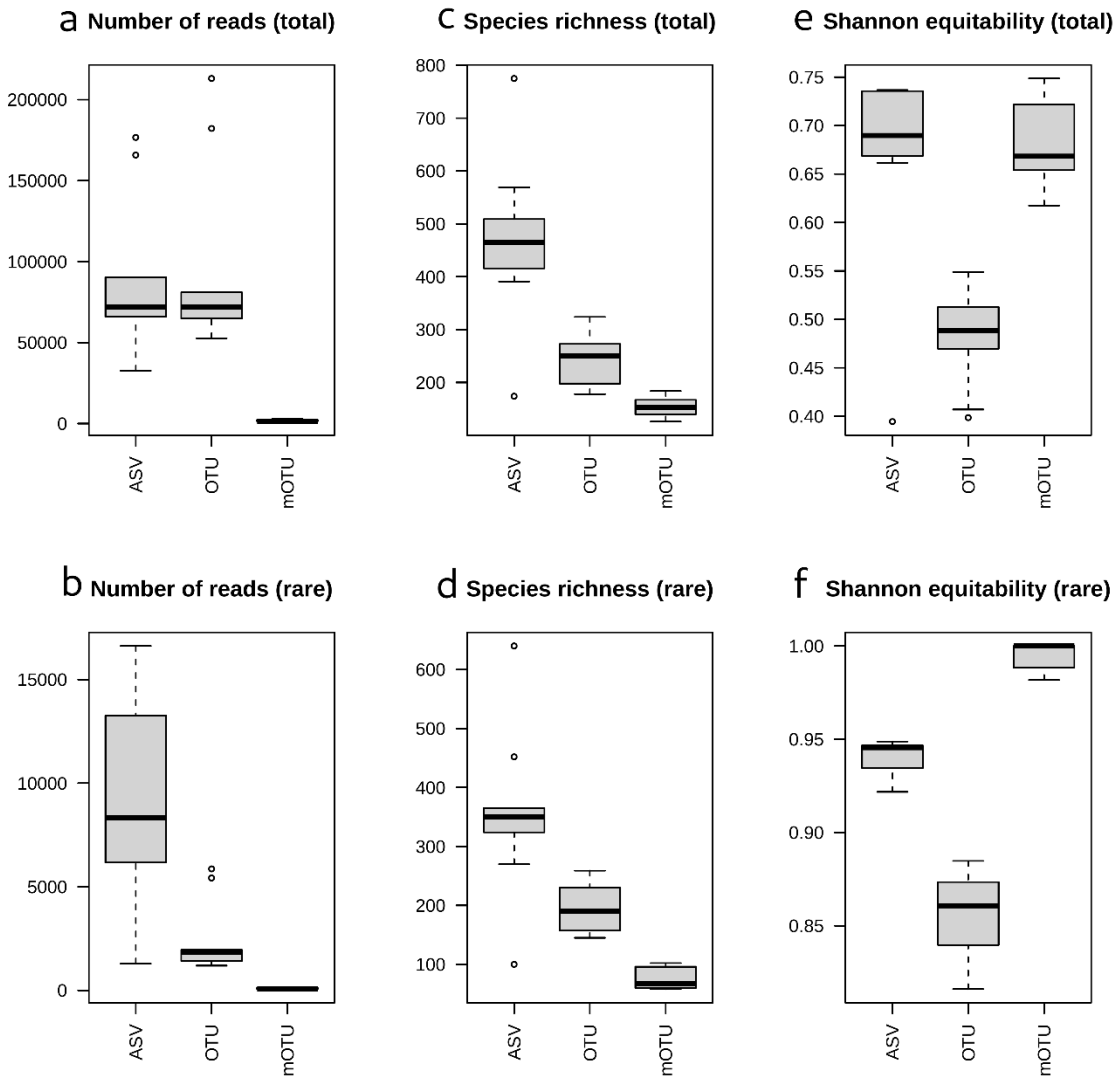
Supplementary Figure 10.16. MultiCoLA results of Procrustes for rarefied data. **a.** Procrustes results for the ASVs; **b.** Procrustes results for the OTUs; **c.** Procrustes results for the mOTUs.



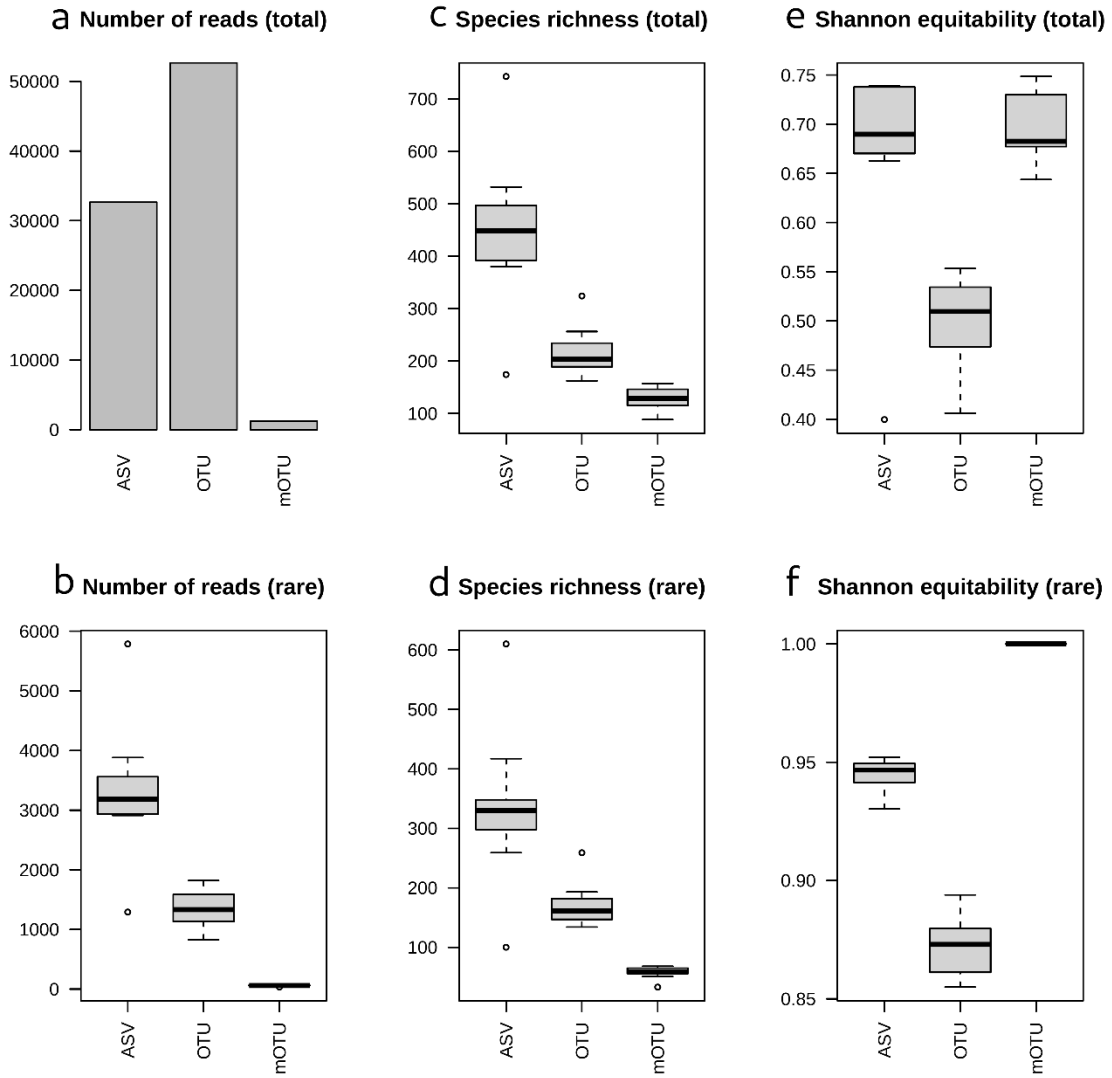
Supplementary Figure 10.17 – Rarefaction curves for ASVs, OTUs, and mOTUs of N-ICE 2015.



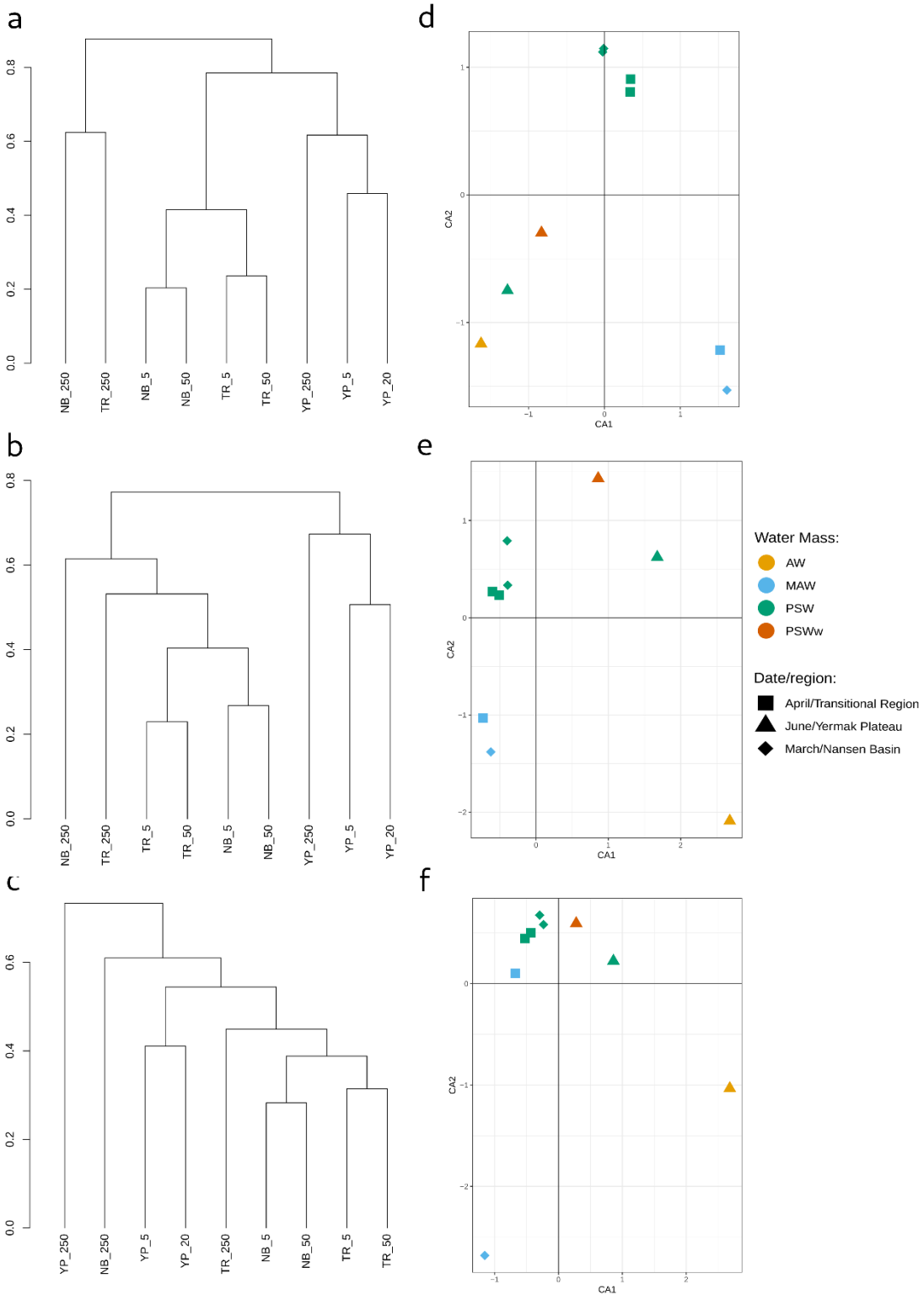
Supplementary Figure 10.18. Alpha diversity for the total community (rarefied) and for the rare biosphere, defined by the MultiCoLA algorithm, across ASVs, OTUs and mOTUs. **a.** Number of reads for the total community; **b.** Number of reads for the rare biosphere; **c**. Species richness for the total community; **d.** Species richness for the rare biosphere; **e**. Shannon equitability for the total community; **f.** Shannon equitability for the rare biosphere.



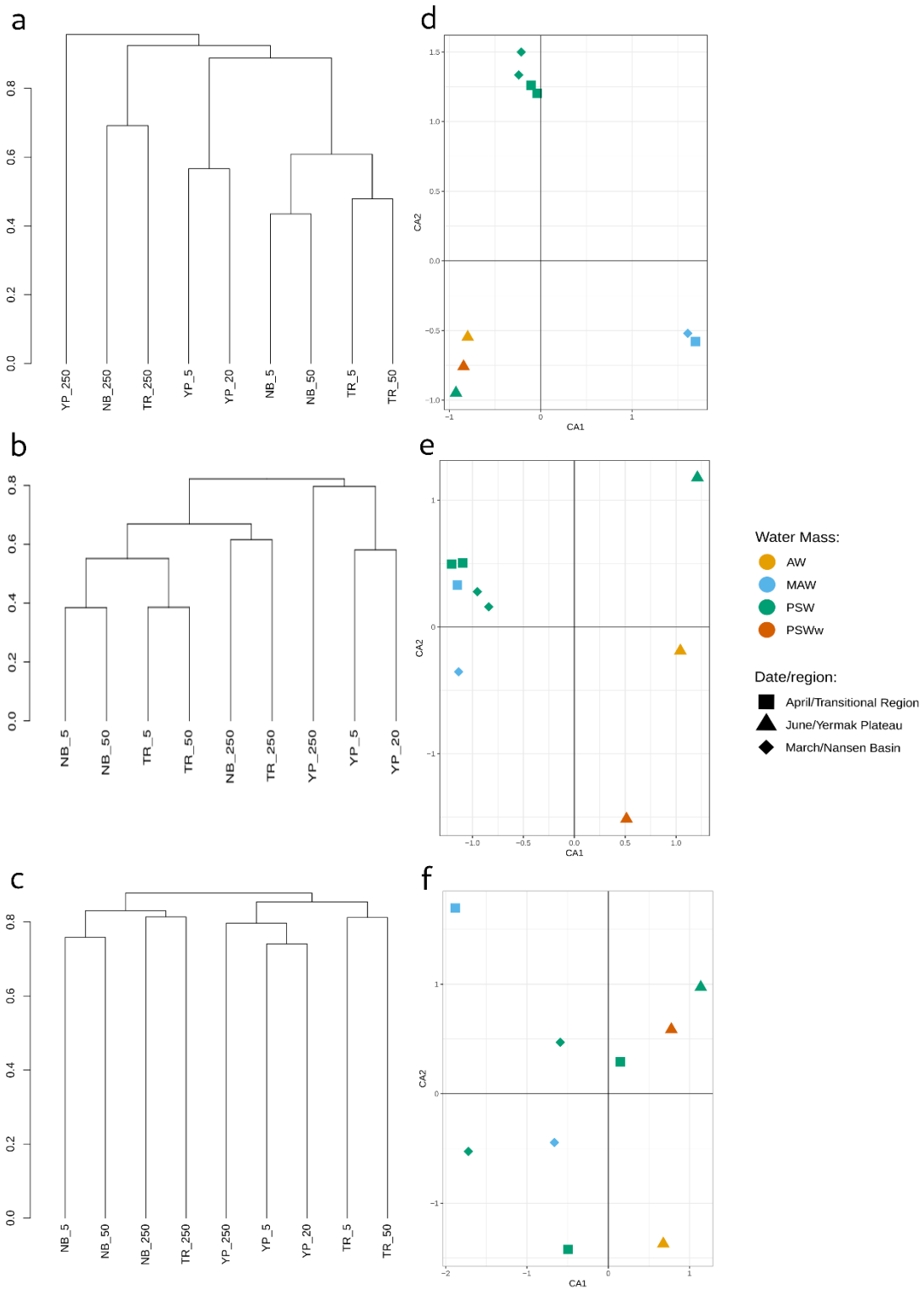
Supplementary Figure 10.19. Alpha diversity for the total community (non-rarefied) and for the rare biosphere, defined by the 0.1% relative abundance threshold, per sample, across ASVs, OTUs and mOTUs. **a**. Number of reads for the total community; **b.** Number of reads for the rare biosphere; **c.** Species richness for the total community. **d.** Species richness for the rare biosphere; **e**. Shannon equitability for the total community; **f.** Shannon equitability for the rare biosphere.



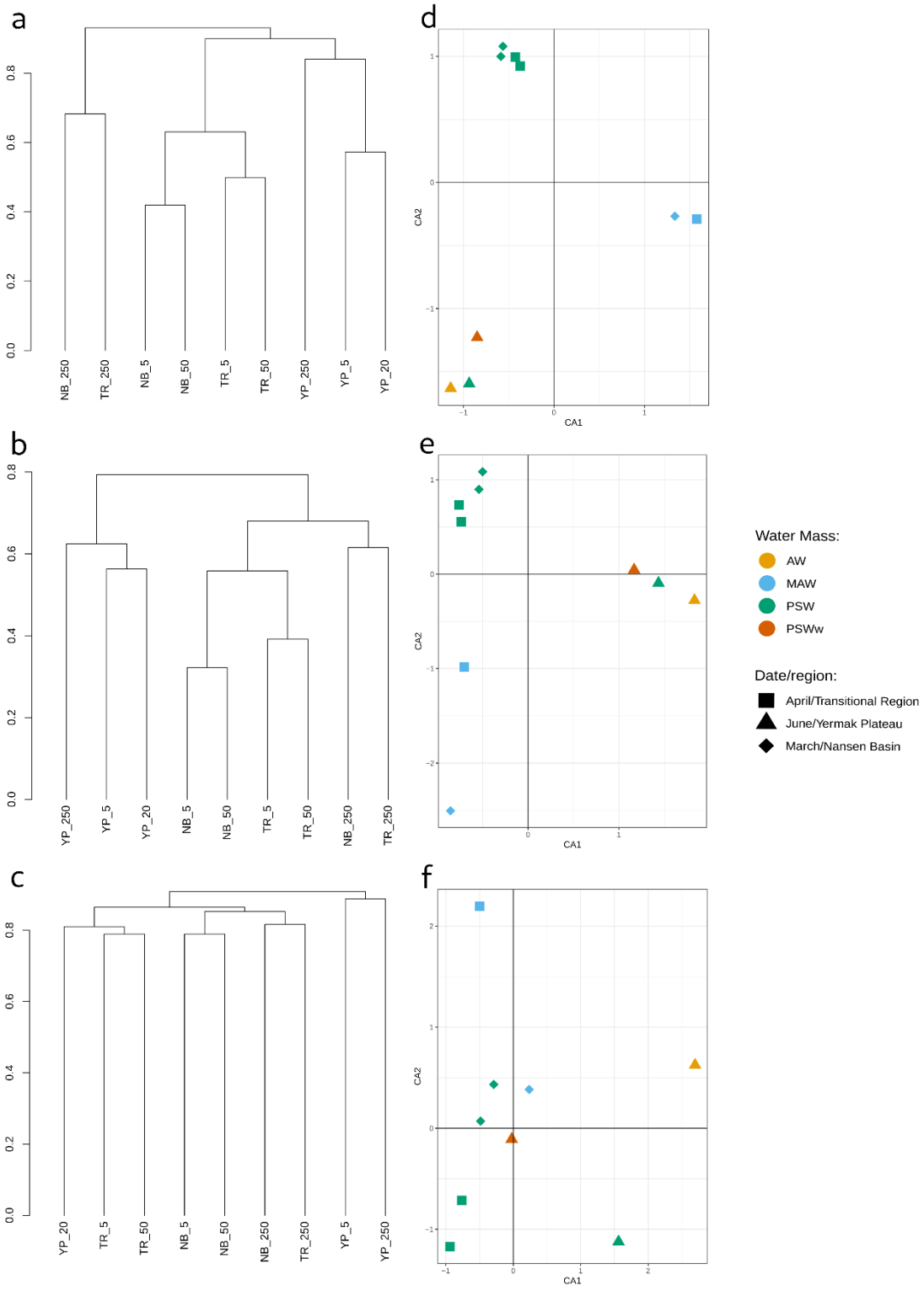
Supplementary Figure 10.20. Alpha diversity for the total community (rarefied) and for the rare biosphere, defined by the 0.1% relative abundance threshold, per sample, across ASVs, OTUs and mOTUs. **a**. Number of reads for the total community; **b.** Number of reads for the rare biosphere; **c.** Species richness for the total community; **d.** Species richness for the rare biosphere; **e.** Shannon equitability for the total community; **f**. Shannon equitability for the rare biosphere.



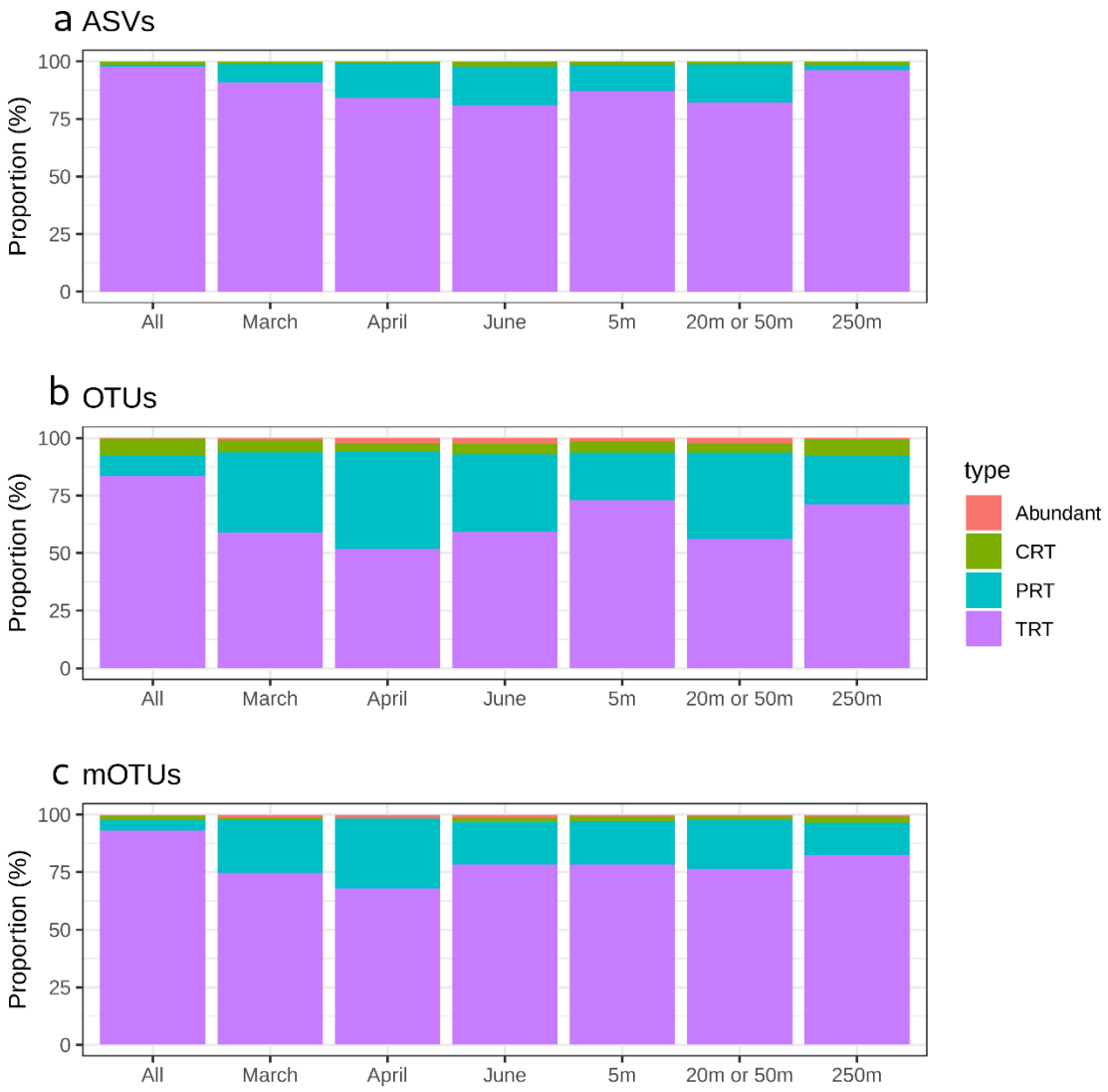
Supplementary Figure 10.21. Dendrograms and multivariate analysis for the rare biosphere, defined by MultiCoLA (rarefied). **a.** Dendrogram of the rare ASVs; **b.** Dendrogram of the rare OTUs; **c.** Dendrogram of the rare mOTUs; **d.** CA for the rare ASVs; **e.** CA for the rare OTUs; **f**. CA for the rare mOTUs.



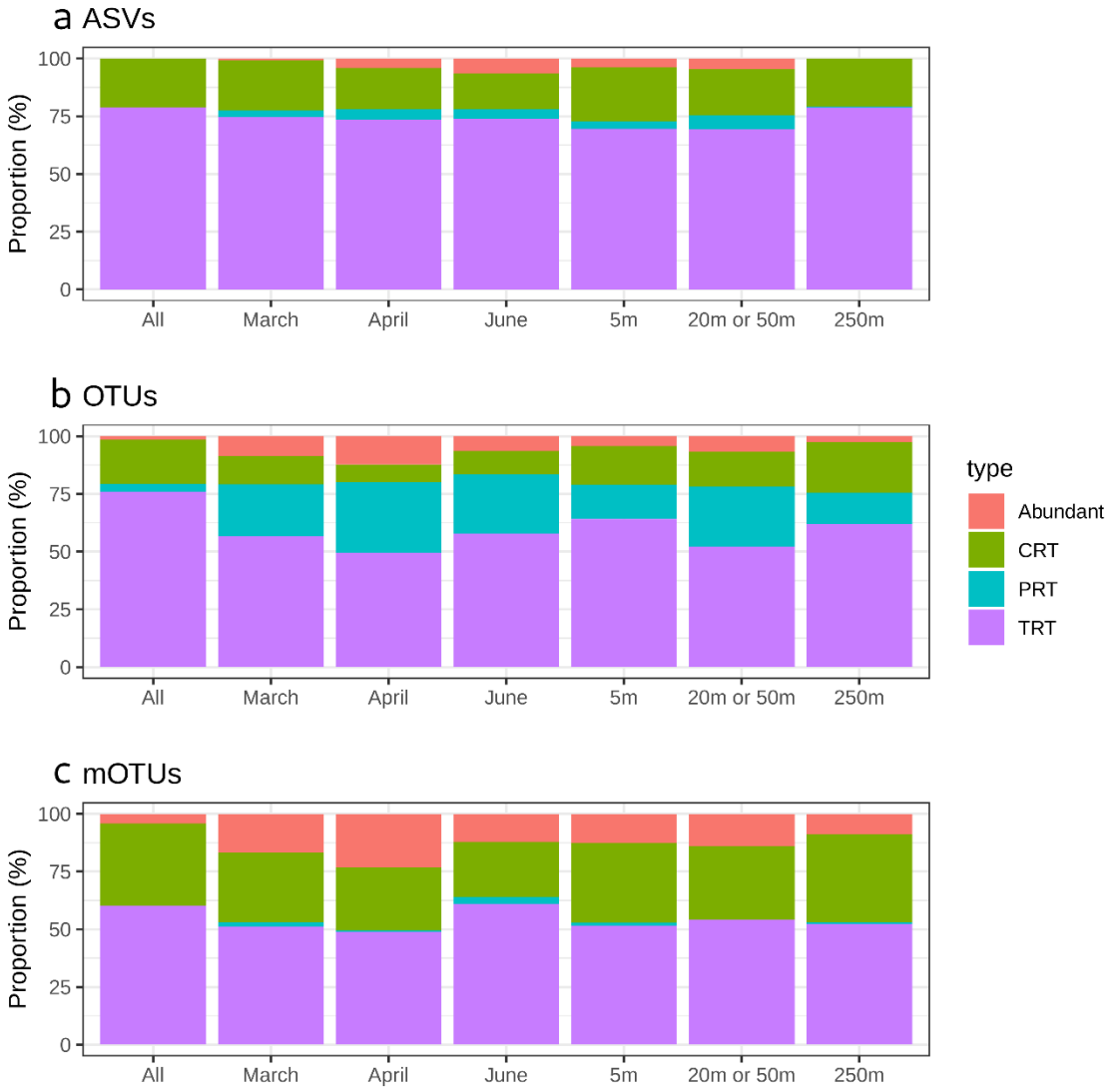
Supplementary Figure 10.22. Dendrograms and multivariate analysis for the rare biosphere, defined by 0.1% relative abundance threshold, per sample (non-rarefied). **a.** Dendrogram of the rare ASVs; **b.** Dendrogram of the rare OTUs; **c.** Dendrogram of the rare mOTUs; **d.** CA for the rare ASVs; **e.** CA for the rare OTUs; **f.** CA for the rare mOTUs.



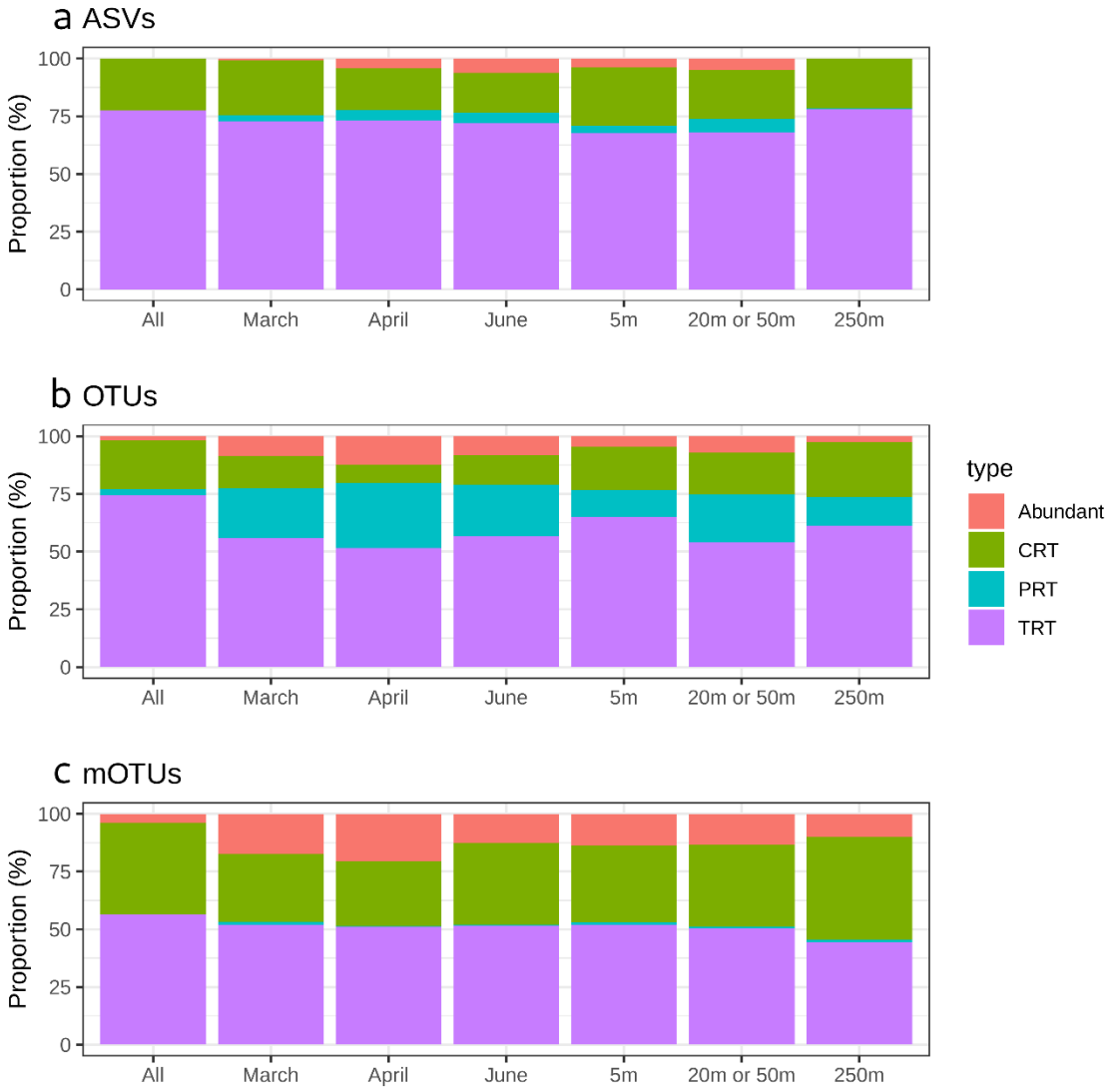
Supplementary Figure 10.23. Dendrograms and multivariate analysis for the rare biosphere, defined by 0.1% relative abundance threshold, per sample (rarefied). **a.** Dendrogram of the rare ASVs; **b.** Dendrogram of the rare OTUs; **c.** Dendrogram of the rare mOTUs; **d.** CA for the rare ASVs; **e.** CA for the rare OTUs; **f**. CA for the rare mOTUs.



Supplementary Figure 10.24. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. Rare biosphere was defined by the MultiCoLA algorithm, from rarefied data. **a.** The proportions for the rare ASVs; **b.** The proportions for the rare OTUs; **c.** The proportions for the rare mOTUs. The stacked bar plots are colored with purple for transiently rare taxa (TRT), blue for permanently rare taxa (PRT), green for conditionally rare taxa (CRT) and red for the abundant taxa. Within each dataset, different groups of samples were selected, specifically: “All” is for all samples; “March” is the group of samples from the month of March and they represent different depths (5 m, 50 m and 250 m); “April” is the group of samples from the month of April and they represent different depths (5 m, 50 m and 250 m); “June” is the group of samples from the month of June and they represent different depths (5 m, 20 m and 250 m); “5 m” is the group of samples taken at 5 m of depth and they represent the different months (March, April and June); “20 m or 50 m” is the group of samples taken at 20 m or 50 m of depth and they represent the different months (March, April and June); “250 m” is the group of samples taken at 250 m of depth and they represent the different months (March, April and June).

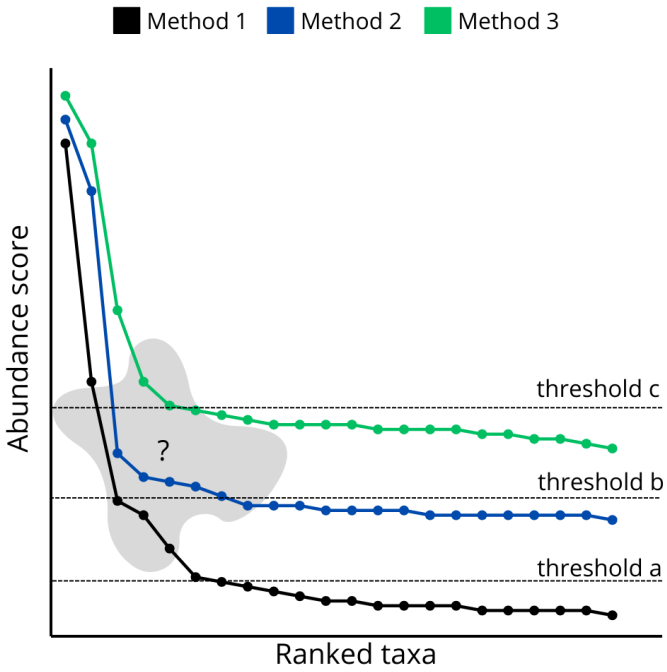


Supplementary Figure 10.25. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. Rare biosphere was defined by the 0.1% relative abundance threshold, per sample, from non-rarefied data. **a.** The proportions for the rare ASVs; **b**. The proportions for the rare OTUs; **c**. The proportions for the rare mOTUs. The stacked bar plots are colored with purple for transiently rare taxa (TRT), blue for permanently rare taxa (PRT), green for conditionally rare taxa (CRT) and red for the abundant taxa. Within each dataset, different groups of samples were selected, specifically: “All” is for all samples; “March” is the group of samples from the month of March and they represent different depths (5 m, 50 m and 250 m); “April” is the group of samples from the month of April and they represent different depths (5 m, 50 m and 250 m); “June” is the group of samples from the month of June and they represent different depths (5 m, 20 m and 250 m); “5 m” is the group of samples taken at 5 m of depth and they represent the different months (March, April and June); “20 m or 50 m” is the group of samples taken at 20 m or 50 m of depth and they represent the different months (March, April and June); “250 m” is the group of samples taken at 250 m of depth and they represent the different months (March, April and June).

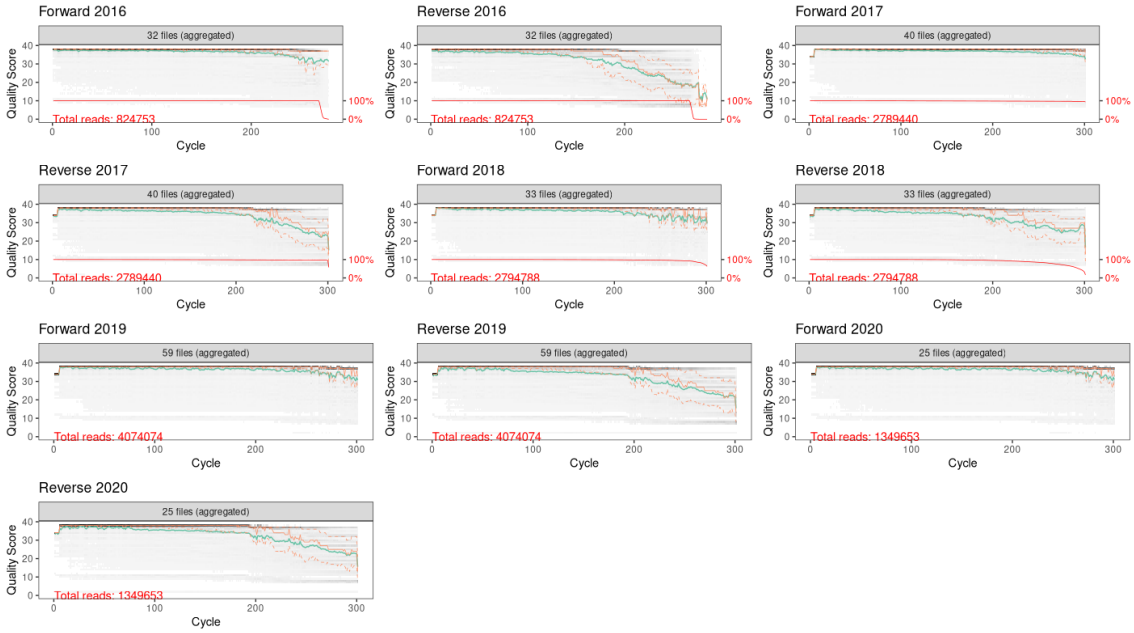


Supplementary Figure 10.26. Stacked bar plots with the proportion of taxa for each type of rarity, if rare, and for the proportion of abundant taxa. Rare biosphere was defined by the 0.1% relative abundance threshold, per sample, from rarefied data. **a.** The proportions for the rare ASVs; **b.** The proportions for the rare OTUs; **c.** The proportions for the rare mOTUs. The stacked bar plots are colored with purple for transiently rare taxa (TRT), blue for permanently rare taxa (PRT), green for conditionally rare taxa (CRT) and red for the abundant taxa. Within each dataset, different groups of samples were selected, specifically: “All” is for all samples; “March” is the group of samples from the month of March and they represent different depths (5 m, 50 m and 250 m); “April” is the group of samples from the month of April and they represent different depths (5 m, 50 m and 250 m); “June” is the group of samples from the month of June and they represent different depths (5 m, 20 m and 250 m); “5 m” is the group of samples taken at 5 m of depth and they represent the different months (March, April and June); “20 m or 50 m” is the group of samples taken at 20 m or 50 m of depth and they represent the different months (March, April and June); “250 m” is the group of samples taken at 250 m of depth and they represent the different months (March, April and June).

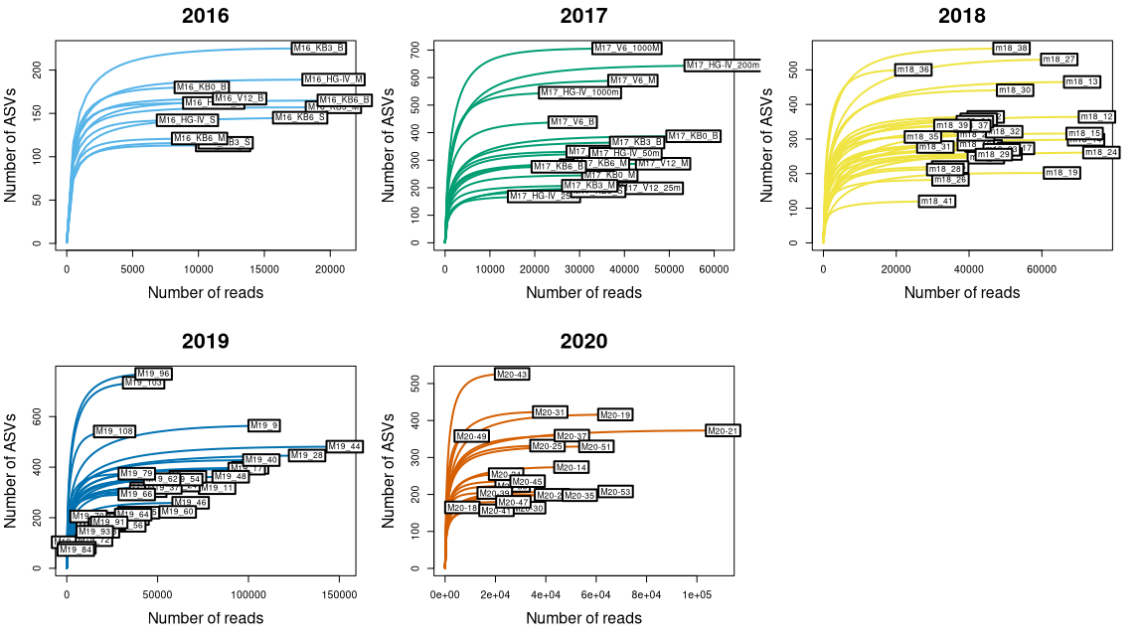
10.4. Supplementary Figures of Chapter 6



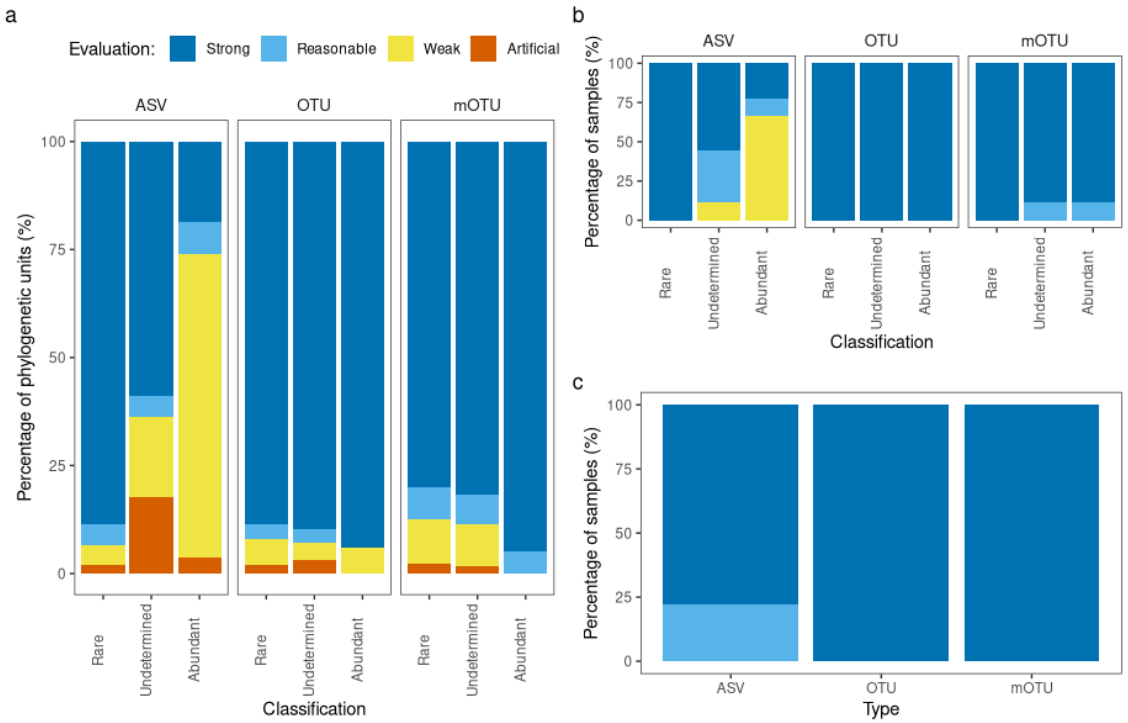
Supplementary Figure 10.27. – Schematic representation of RAC and of the problem with threshold-based definitions of the rare biosphere. Three different RAC are illustrated, representing different methodologies with abundance scores in different orders of magnitude. The grey area with the question mark represents the area of the curve where a threshold is decided, and three different thresholds are illustrated, but neither can work equally well for all three curves.



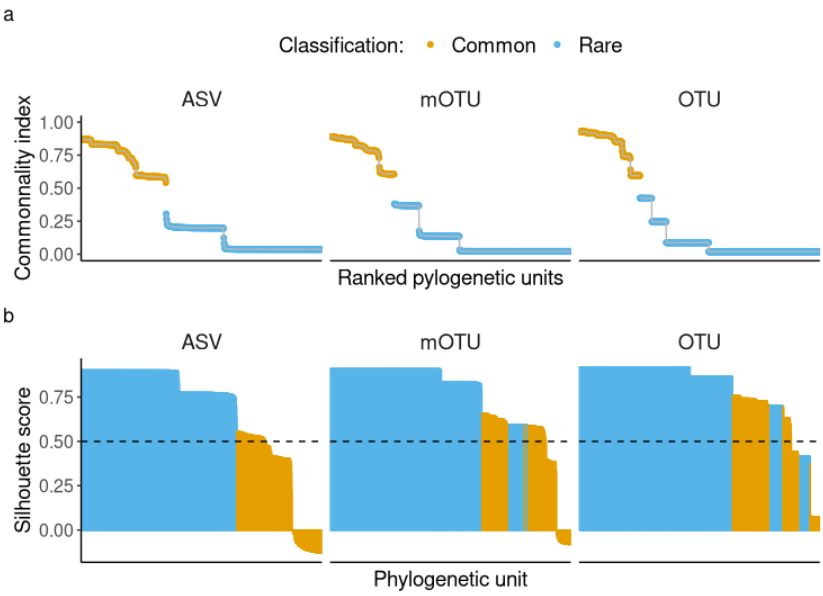
Supplementary Figure 10.28. Aggregate quality profiles of FASTQ files (Forward and Reverse) with V4V5 16S rRNA gene amplicon sequencing results from MOSJ 2016-2020.



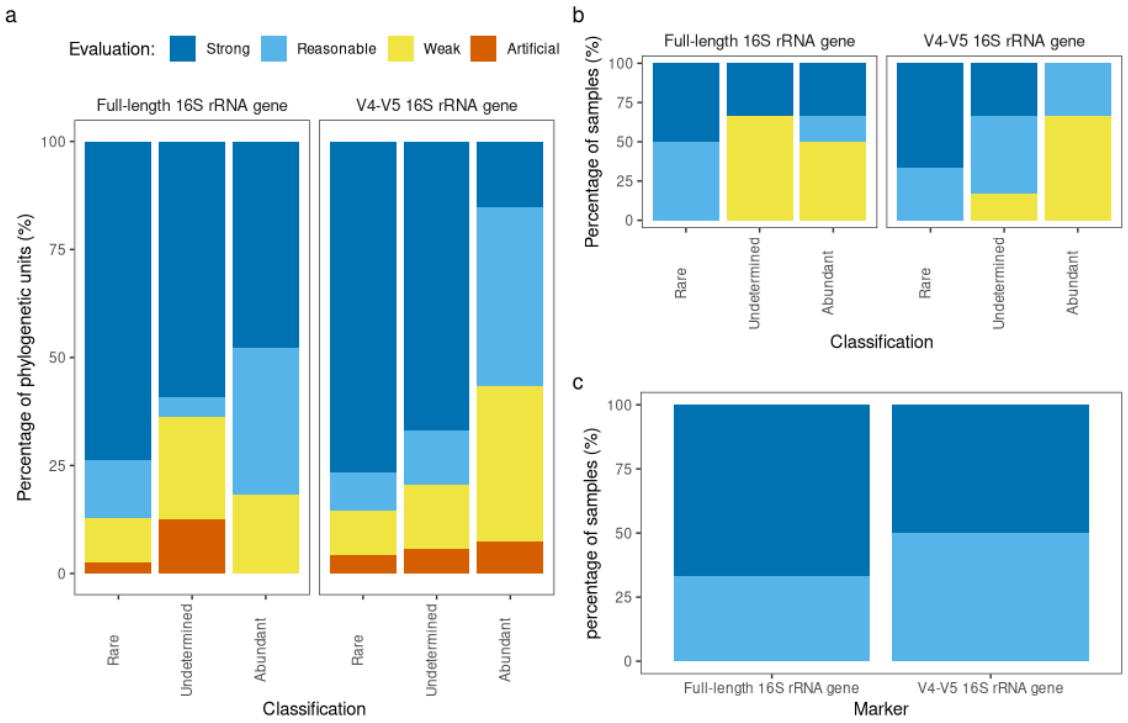
Supplementary Figure 10.29. Rarefaction curves for the entire MOSJ dataset, grouped by years (from 2016 to 2020, totaling 117 samples).



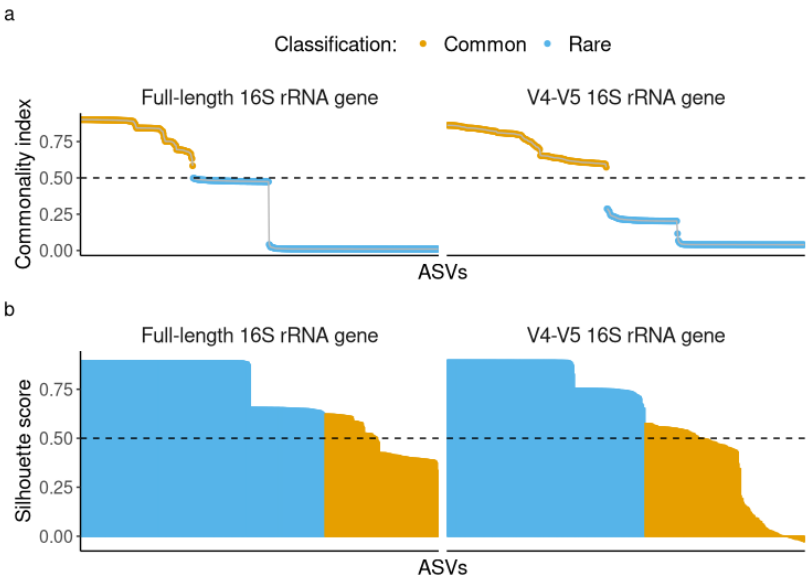
Supplementary Figure 10.30. Evaluation of clusters obtained for ASVs, OTUs, and mOTUs, based on Silhouette scores. **a**. Proportion of ASVs with each evaluation; **b**. Percentage of samples with each evaluation; **c**. Percentage of samples with each evaluation for the clustering result obtained for the whole dataset.



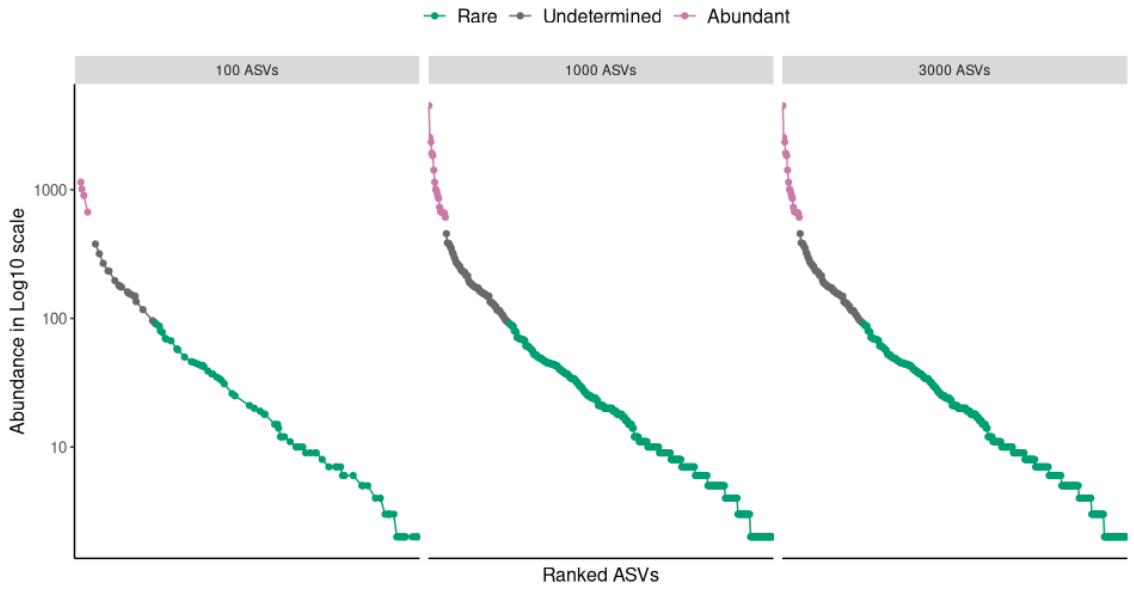
Supplementary Figure 10.31. *FuzzyQ* results for different phylogenetic units (ASV, mOTU and OTU). **a**. Commonality index (all samples combined). By default, an index above 0.5 is considered "Common" and below 0.5 is considered "Rare". **b**. Silhouette scores. Colors divide "Common" and "Rare" phylogenetic units. All scores were plotted in descending order.



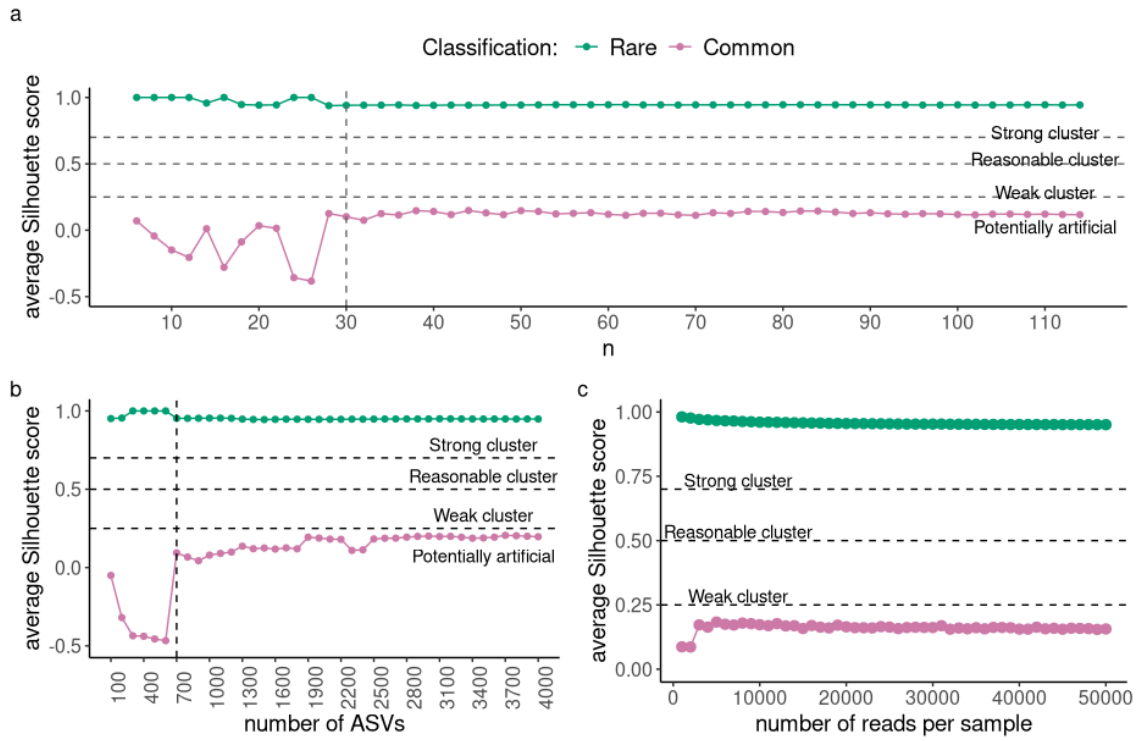
Supplementary Figure 10.32. Evaluation of clusters obtained for ASVs derived from V4V5 and full-length 16S rRNA gene, based on Silhouette scores. **a**. Proportion of ASVs with each evaluation; **b**. Percentage of samples with each evaluation; **c**. Percentage of samples with each evaluation for the clustering result obtained for the whole dataset.



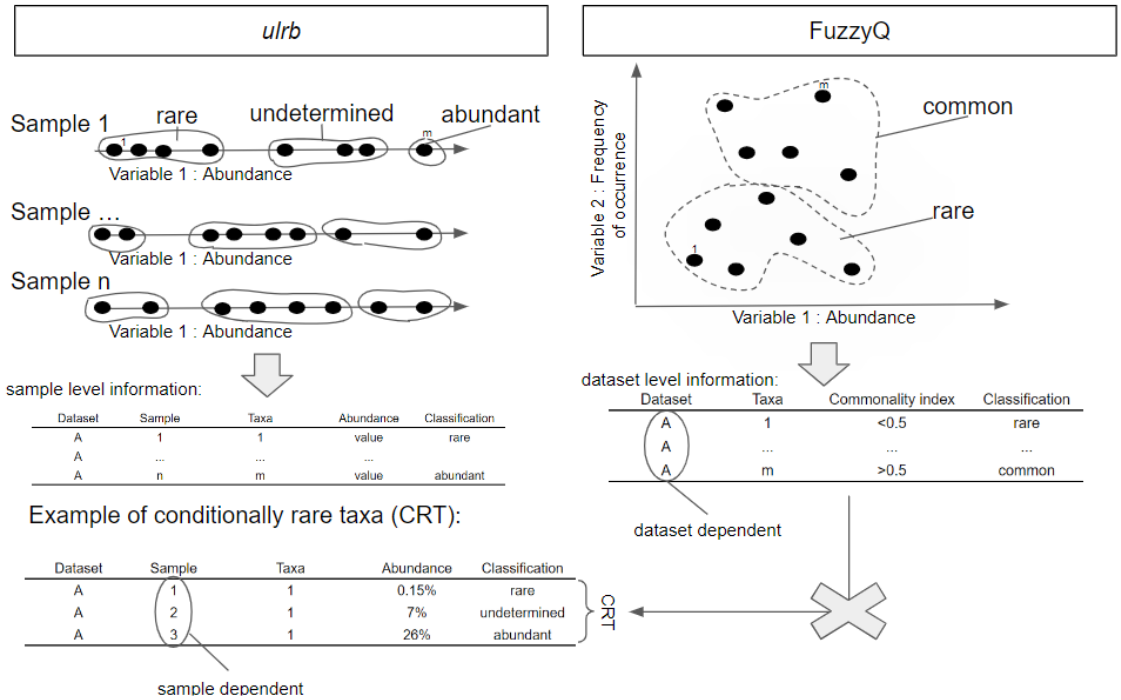
Supplementary Figure 10.33. FuzzyQ results for V4V5 and full-length 16S rRNA gene sequencing. **a**. Commonality index (all samples combined); **b**. Silhouette scores. Abundance classifications were colored coded. All scores were plotted in descending order.



Supplementary Figure 10.34. Illustration of RAC obtained by different number of randomly selected ASVs. A random sample (m18_33) was selected for illustrative purposes.

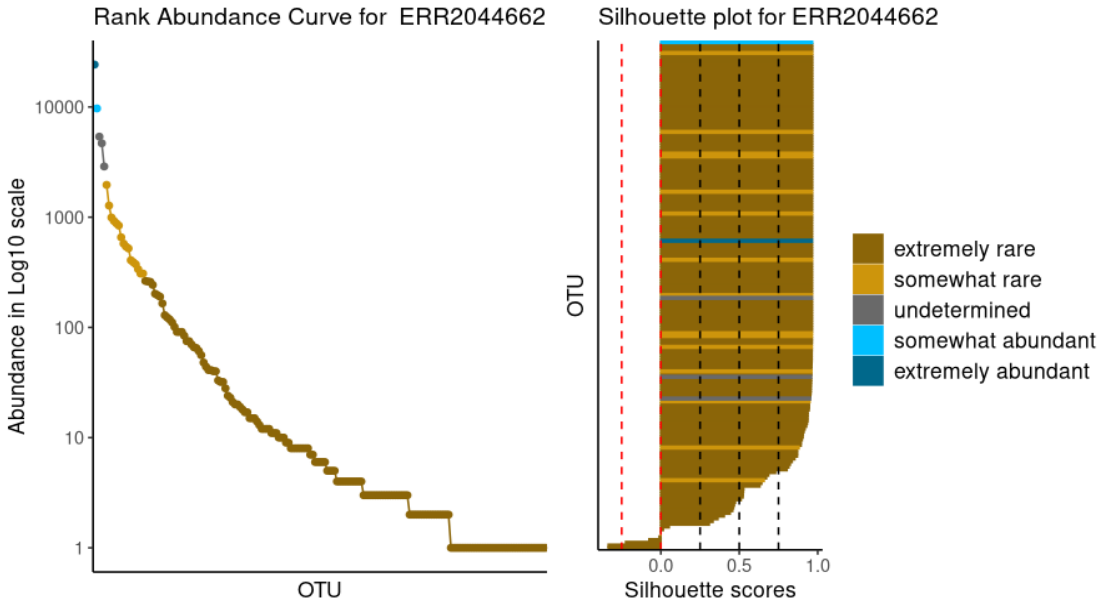


Supplementary Figure 10.35. Quality of *FuzzyQ* clustering measured by the average Silhouette score as a function of number of samples, ASVs, and sequencing depth. **a**. Number of samples (n), ranging from 6 samples up to 114 samples, in increments of two-by-two samples; **b**. Number of ASVs, ranging from 100 to 4000 ASVs per sample, in increments of 300 by 300 ASVs; **c**. Number of reads per sample, in increments of 1000 by 1000 reads. For the response variable, there is one average Silhouette score per observation and the classifications were grouped by different colors, as illustrated in the figure legend.



Supplementary Figure 10.36. Schematic representation of the main differences between *ulrb* and *FuzzyQ*. As illustrated, *ulrb* applies its algorithm in each sample separately, attributing a classification for each taxon. Then, the output table will provide information for each taxon in each sample. The output from *ulrb* allows for the characterization of types of rarity, as illustrated by an example of conditionally rare taxa. As for *FuzzyQ*, since it considers simultaneously abundance and frequency of occurrence, it provides information for each taxon at the whole dataset level. Thus, it does not allow for the same characterization of the types of rarity as is possible with *ulrb*.

10.5. Supplementary Figures of Chapter 7



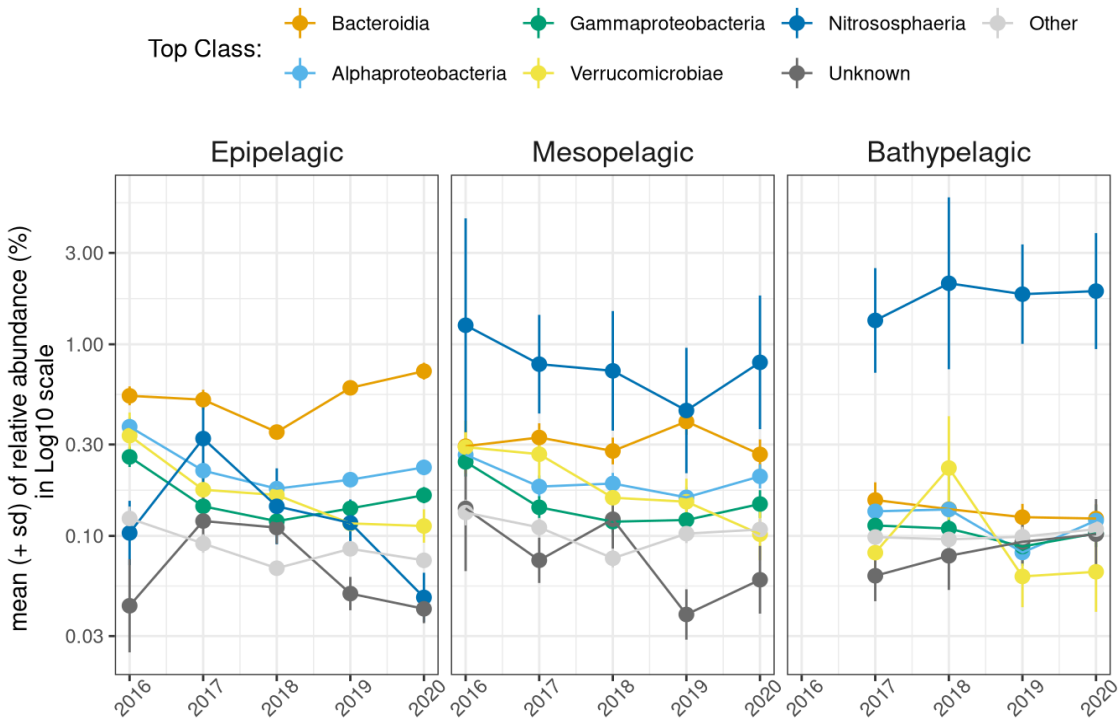
Supplementary Figure 10.37. Example of plot_*ulrb*() output for a single sample, using five classifications. The left grid shows the abundance of each OTU, ordered from most to least abundant. The right grid shows the Silhouette score of each OTU, ordered from highest to lowest. OTUs are colored according to their classification.

```
# A tibble: 72 × 6
# Groups:   Sample [9]
   Sample     data                    DB     CH average_Silhouette     k
   <chr>      <list>               <dbl>  <dbl>              <dbl> <int>
 1 ERR2044662 <tibble [187 × 10]> 0.372  1821.              0.952     3
 2 ERR2044662 <tibble [187 × 10]> 0.527  2055.              0.882     4
 3 ERR2044662 <tibble [187 × 10]> 0.413  4934.              0.856     5
 4 ERR2044662 <tibble [187 × 10]> 0.429  5465.              0.840     6
 5 ERR2044662 <tibble [187 × 10]> 0.335 17589.              0.848     7
 6 ERR2044662 <tibble [187 × 10]> 0.389 17180.              0.784     8
 7 ERR2044662 <tibble [187 × 10]> 0.395 18083.              0.774     9
 8 ERR2044662 <tibble [187 × 10]> 0.329 30332.              0.770    10
 9 ERR2044663 <tibble [220 × 10]> 0.391  1491.              0.955     3
10 ERR2044663 <tibble [220 × 10]> 0.516  1495.              0.891     4
# i 62 more rows
# i Use `print(n = ...)` to see more rows
```

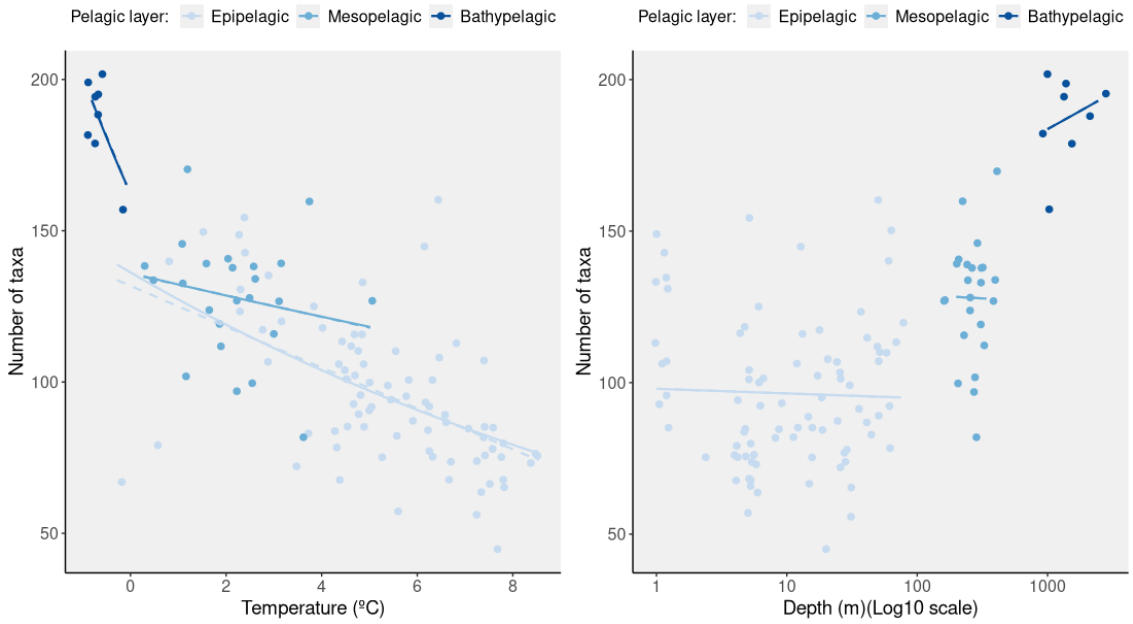
Supplementary Figure 10.38. Print screen of data frame obtained by setting with_plot = FALSE in the evaluate_k() function. The data frame includes the scores for each index obtained for each k value, in each sample.

10.6. Supplementary Figures of Chapter 8

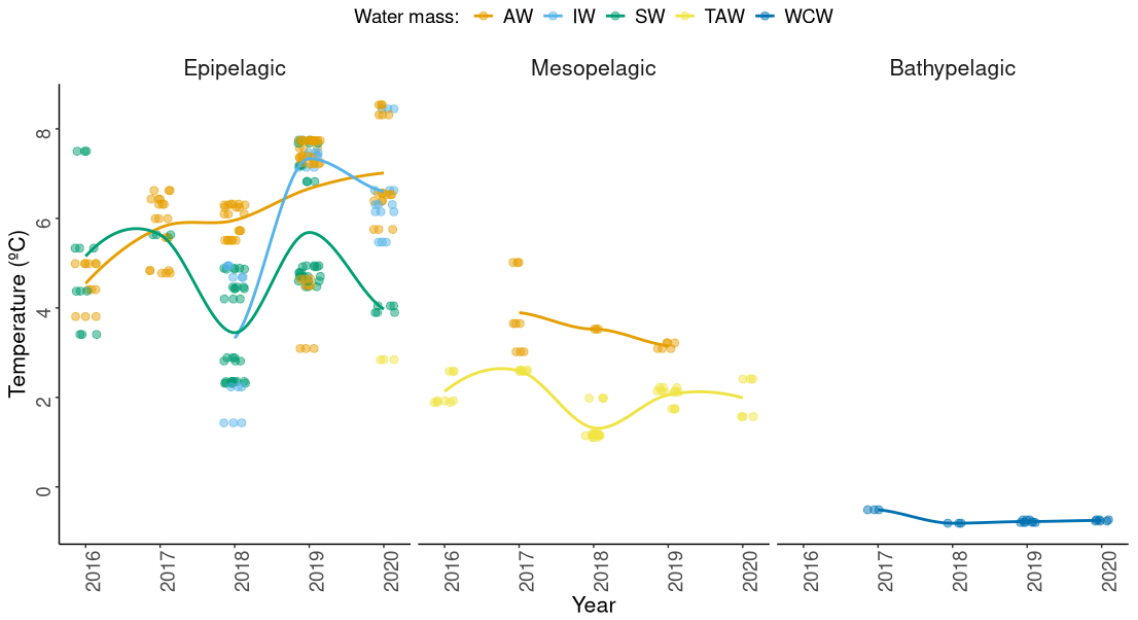
Supplementary Figure 10.39. Relative abundance (%) of taxonomic groups, in each sample. The x axis shows, for each sample, all taxonomic groups present (vertically). The left grid only shows taxonomic groups that appeared in a single sample and the right grid shows the taxonomic groups that were present in more than one sample. The x axis ticks were omitted, because they overlap excessively. Thus, each vertical line represents the taxonomic groups present in a given sample. For example, if several taxonomic groups were present in a single sample, then they would appear in the vertical line of that sample in the left grid. Red dots indicate singleton taxonomic groups (represented by a single read). A red line indicates the 0.01% relative abundance threshold.



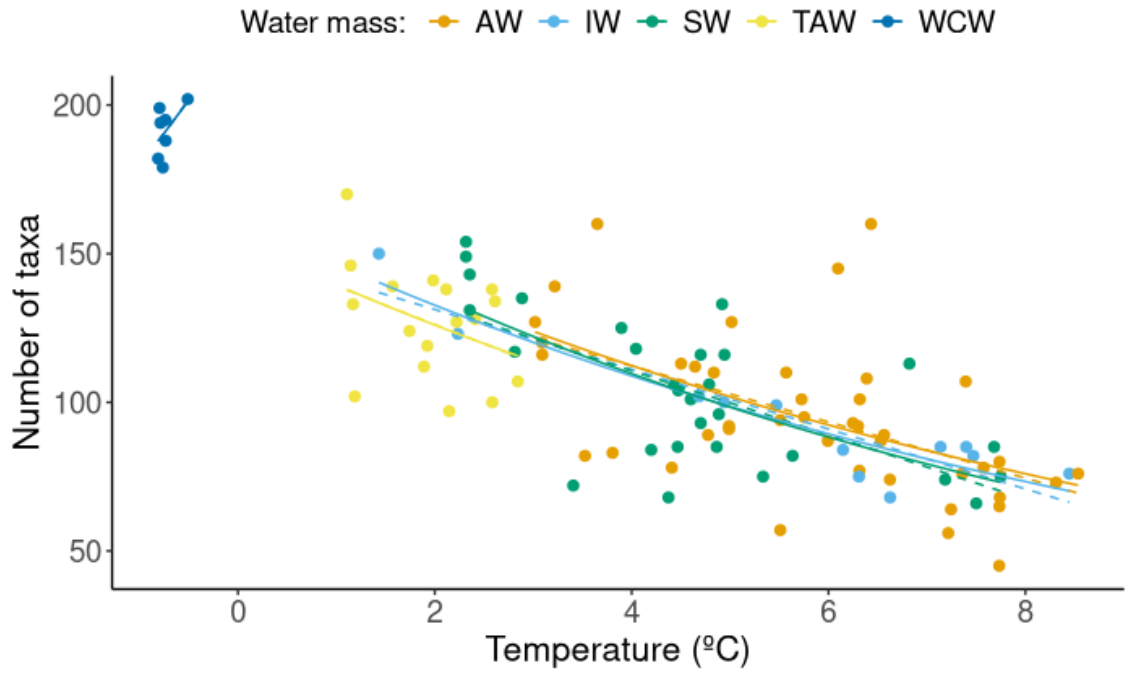
Supplementary Figure 10.40. Mean ± standard deviation (sd) of relative abundance (%) in Log10 scale of class across year and pelagic layers. The five most abundant classes across the entire dataset were illustrated with qualitative coloring and the remaining phyla were grouped into “Other”, in light gray, or “Unknown”, in dark gray.



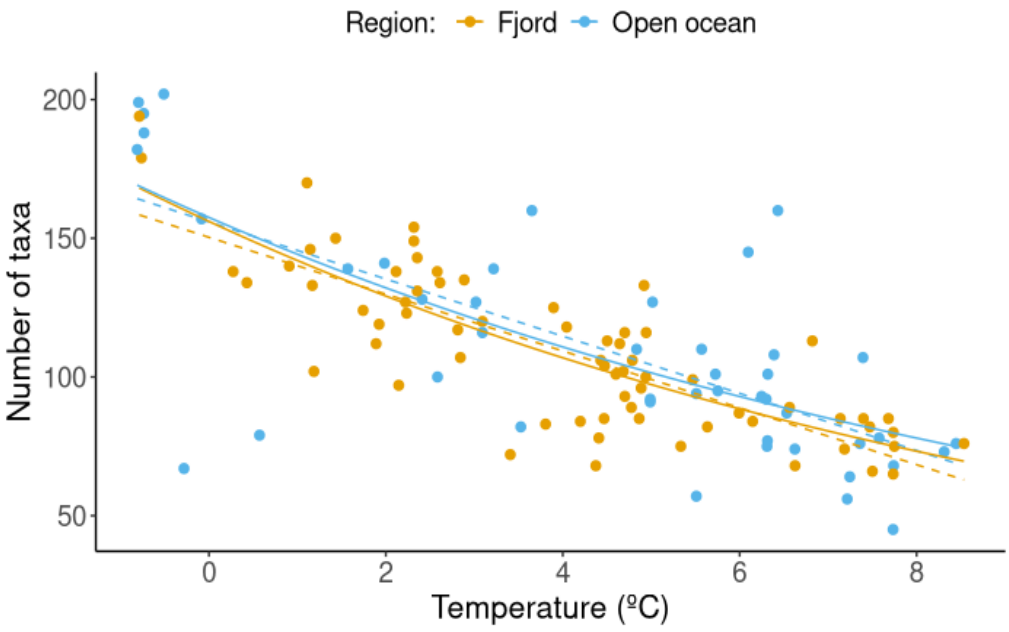
Supplementary Figure 10.41. Number of taxa present in a given sample (alpha diversity estimate) as a function of temperature (ºC), on the left side, and depth (m) in Log10 scale, on the right side. Samples are colored by pelagic layers (epipelagic, mesopelagic and bathypelagic). All samples with temperature (ºC) available were used (n = 114). Dashed regression lines represent the linear least squares model output, and the continuous lines represent the Poisson model.



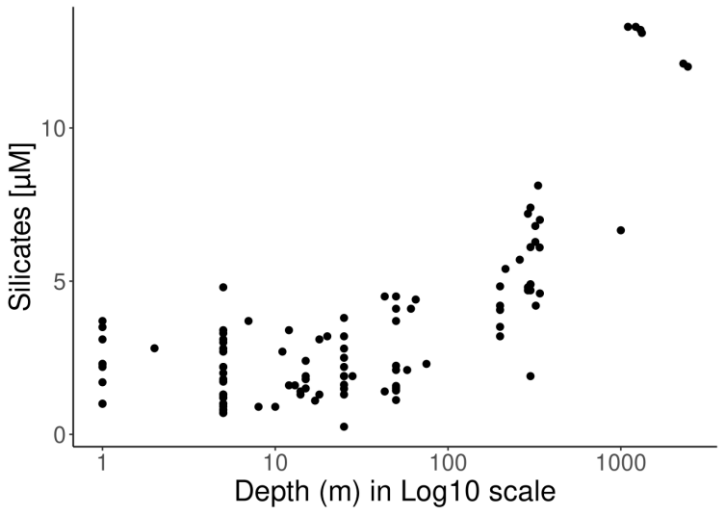
Supplementary Figure 10.42. Temperature (ºC) as a function of year (from 2016 to 2020), grouped by pelagic layer. The water masses selected are: Atlantic Water (AW); Intermediate Water (IW); Surface Water (SW); Transformed Atlantic Water (TAW); and Winter Cooled Water (WCW). All samples with available temperature and that belonged to a water mass with n > 10 samples were used (n = 107). Points are semitransparent and jittered, for better visibility in areas with overlapped points. Lines were added to improve readability.



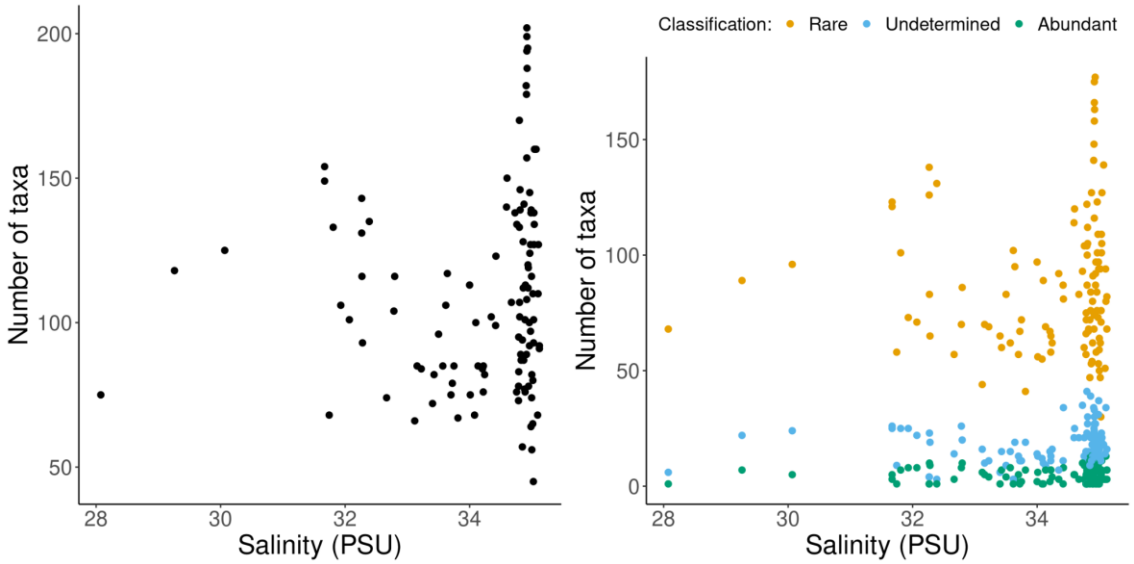
Supplementary Figure 10.43. Number of taxa as a function of temperature (ºC). Samples are colored based on water mass. The water masses selected are: Atlantic Water (AW); Intermediate Water (IW); Surface Water (SW); Transformed Atlantic Water (TAW); and Winter Cooled Water (WCW). Dashed regression lines represent the linear least squares model, and continuous lines represent the Poisson model output. All samples with available temperature (ºC) and that belonged to a water mass with n > 10 samples were used (n = 107).



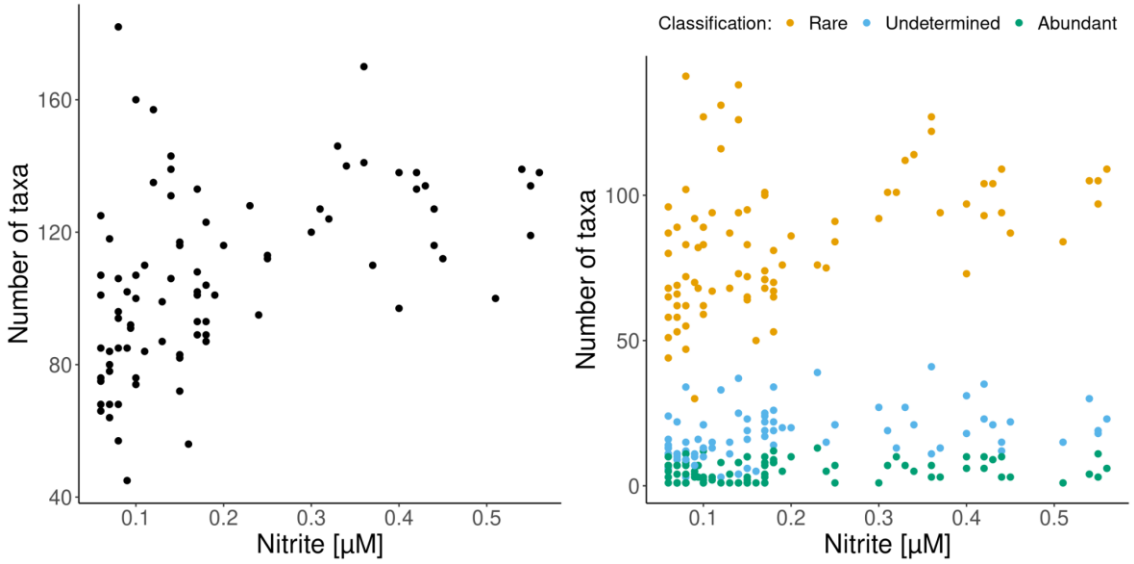
Supplementary Figure 10.44. Number of taxa as a function of temperature (ºC). Samples are colored based on region. Dashed regression lines represent the linear least squares model, and continuous regression lines represent the Poisson model. The regions are divided into the Fjord area (corresponding to samples taken at the stations within the Kongsfjorden) and Open Ocean (corresponding to samples taken at the stations outside the fjord, in the Fram Strait). All samples with available temperature (ºC) were used.



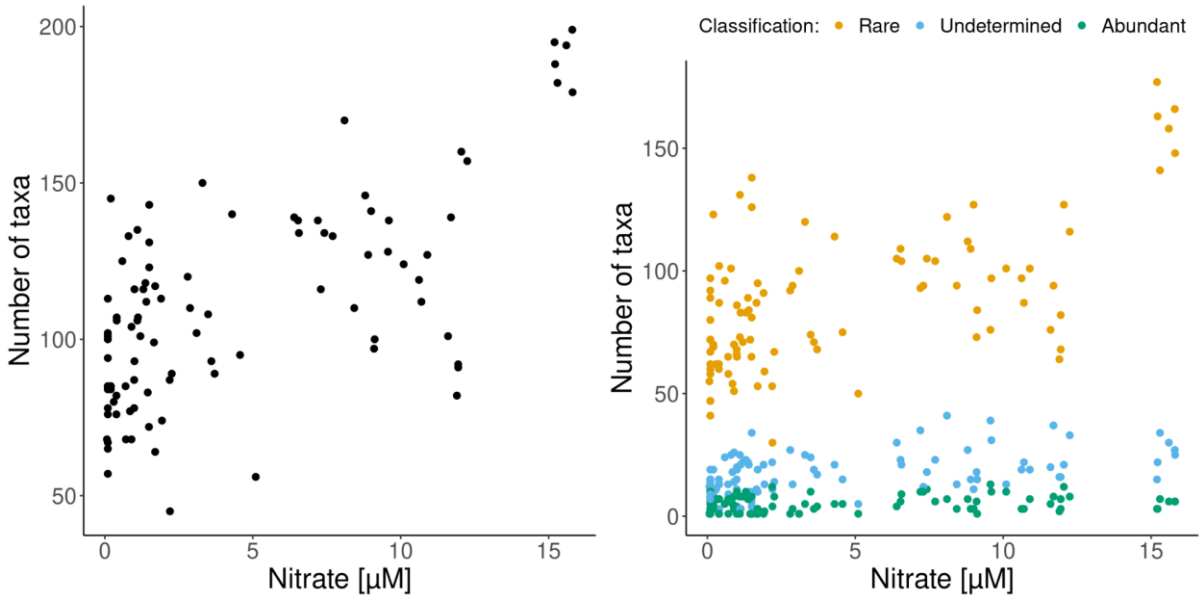
Supplementary Figure 10.45. Silicate concentration (µM) as a function of depth (m) in Log10 scale. All samples with silicate concentration (µM) available were used (n = 108).



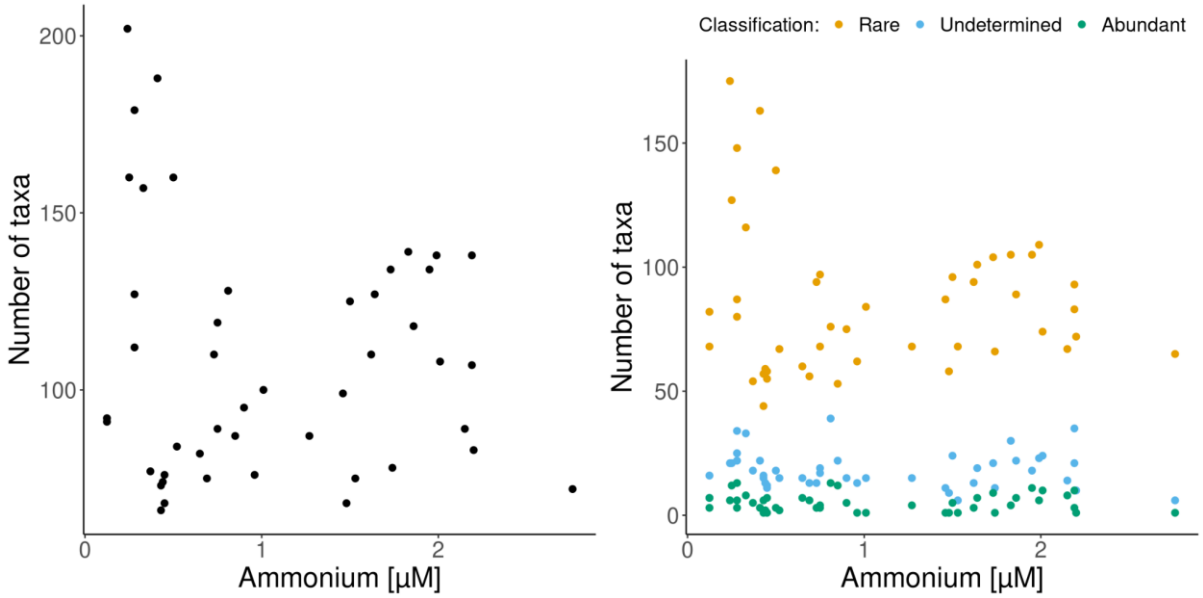
Supplementary Figure 10.46. Number of taxa present in a given sample (alpha diversity estimate) as a function of salinity concentration (PSU). The left panel shows alpha diversity for the overall dataset. The right panel shows the number of taxa per sample, in each classification, with qualitative coloring. All samples with salinity concentration (PSU) available were used (n = 114).



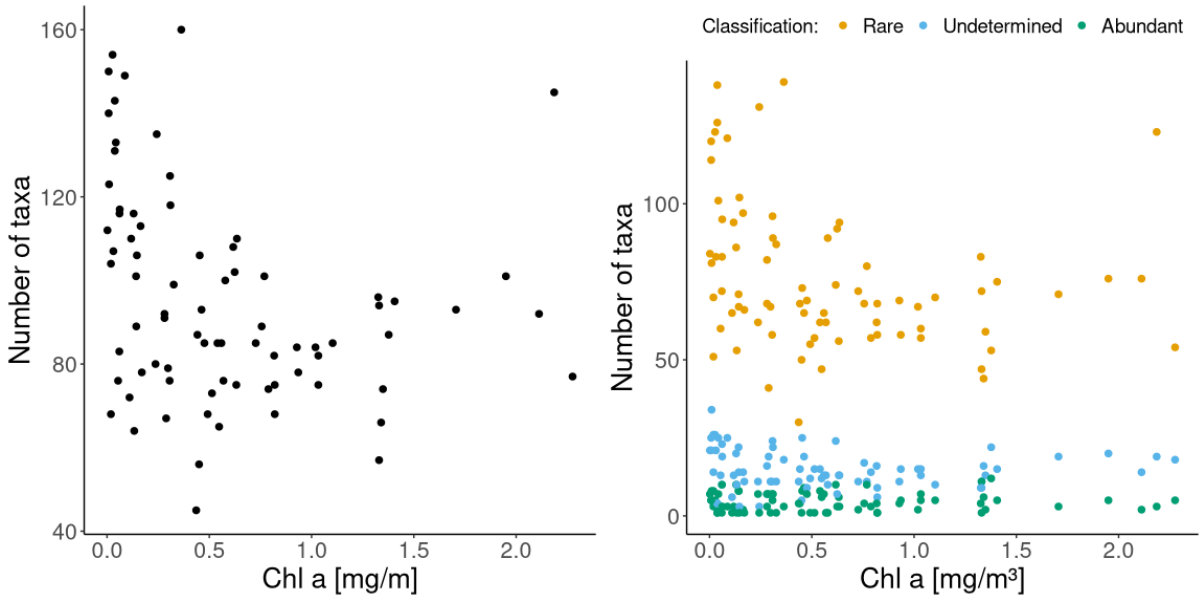
Supplementary Figure 10.47. Number of taxa present in a given sample (alpha diversity estimate) as a function of NO_2^- concentration (µM). The left panel shows alpha diversity for the overall dataset. The right panel shows the number of taxa per sample, in each classification, with qualitative coloring. All samples with NO_2^- concentration (µM) available were used (n = 117).



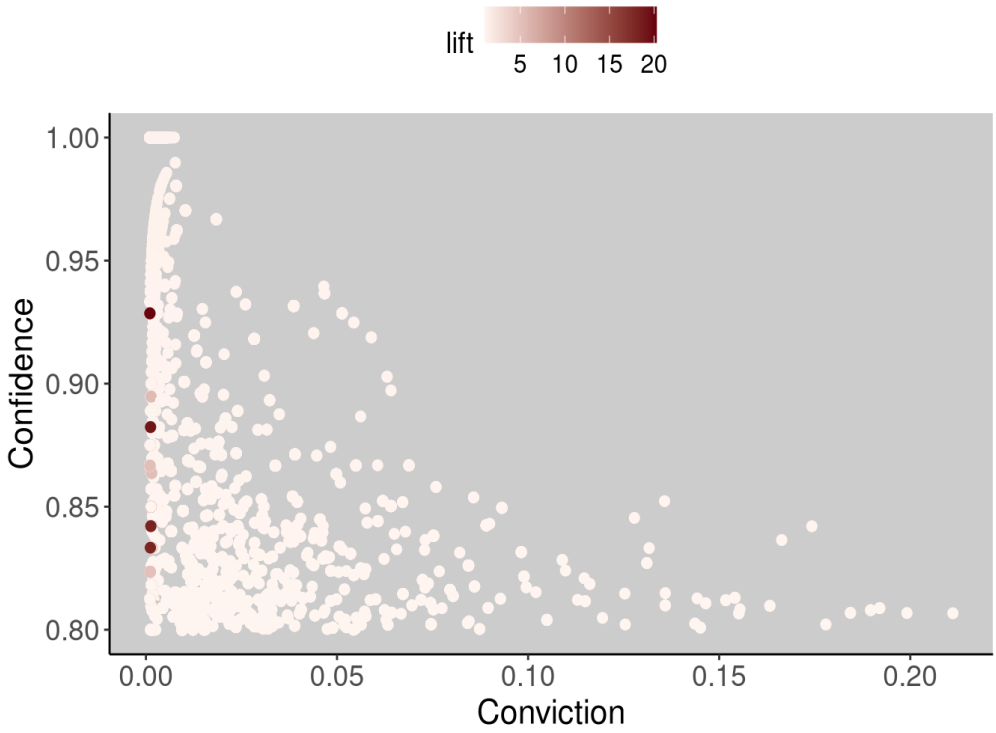
Supplementary Figure 10.48. Number of taxa present in a given sample (alpha diversity estimate) as a function of NO_3^- concentration (µM). The left panel shows alpha diversity for the overall dataset. The right panel shows the number of taxa per sample, in each classification, with qualitative coloring. All samples with NO_3^- concentration (µM) available were used (n = 108).



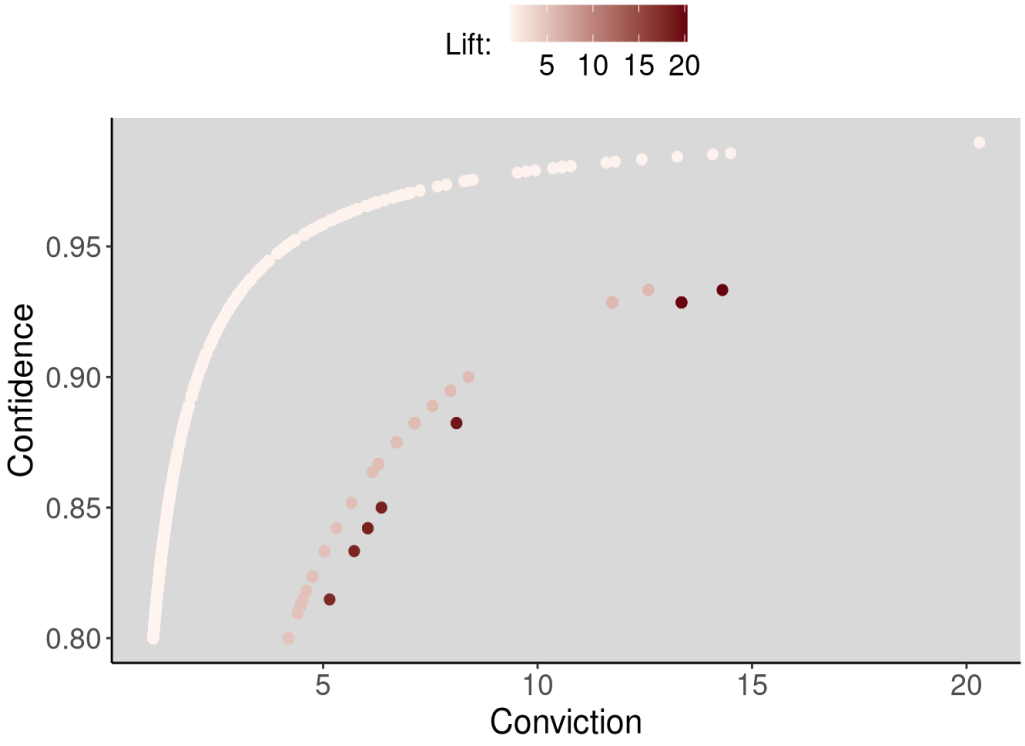
Supplementary Figure 10.49. Number of taxa present in a given sample (alpha diversity estimate) as a function of NH_4^+ concentration (µM). The left panel shows alpha diversity for the overall dataset. The right panel shows the number of taxa per sample, in each classification, with qualitative coloring. All samples with NH_4^+ concentration (µM) available were used (n = 46).



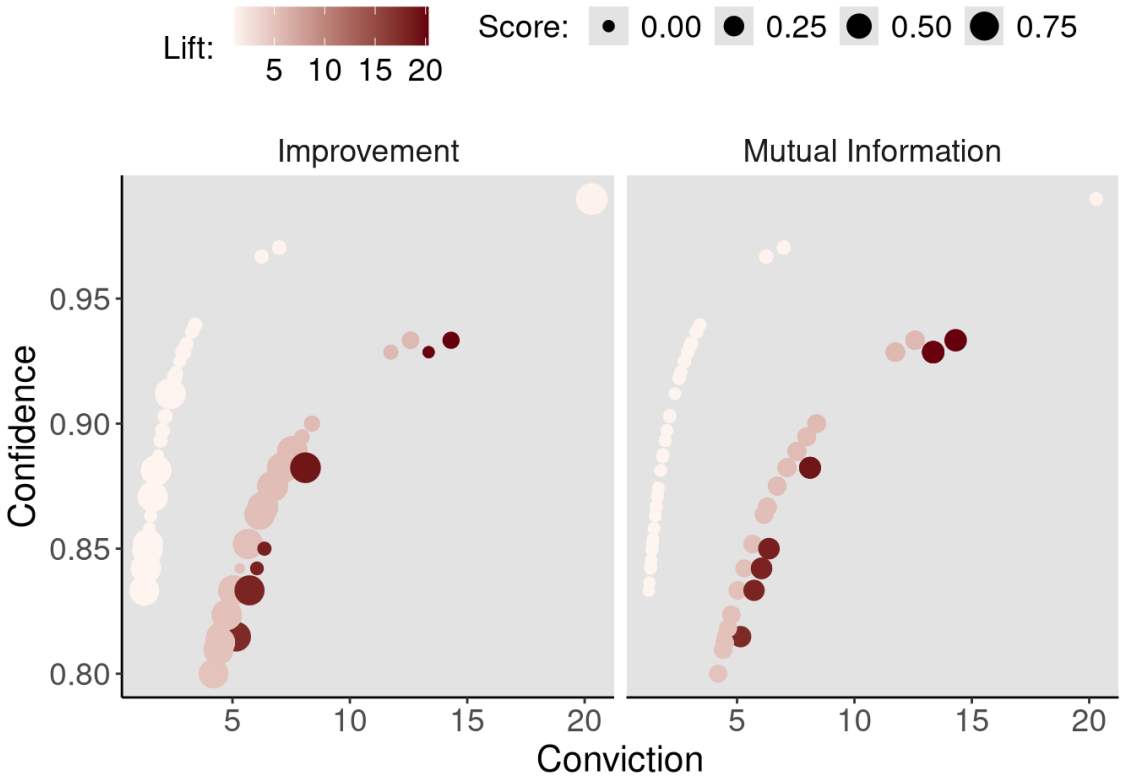
Supplementary Figure 10.50. Number of taxa present in a given sample (alpha diversity estimate) as a function of chlorophyll *a* concentration (mg/m³). The left panel shows alpha diversity for the overall dataset. The right panel shows the number of taxa per sample, in each classification, with qualitative coloring. All samples with Chl *a* concentration (µM) available were used (n = 46).



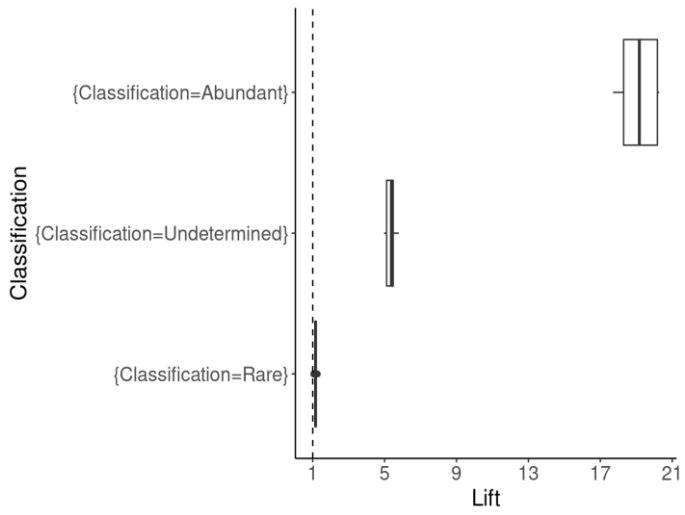
Supplementary Figure 10.51. Confidence as a function of Conviction, for all rules obtained (n = 18 550 rules). Rules were colored by Lift (red shading). Gray background was used to allow visibility of low Lift values.



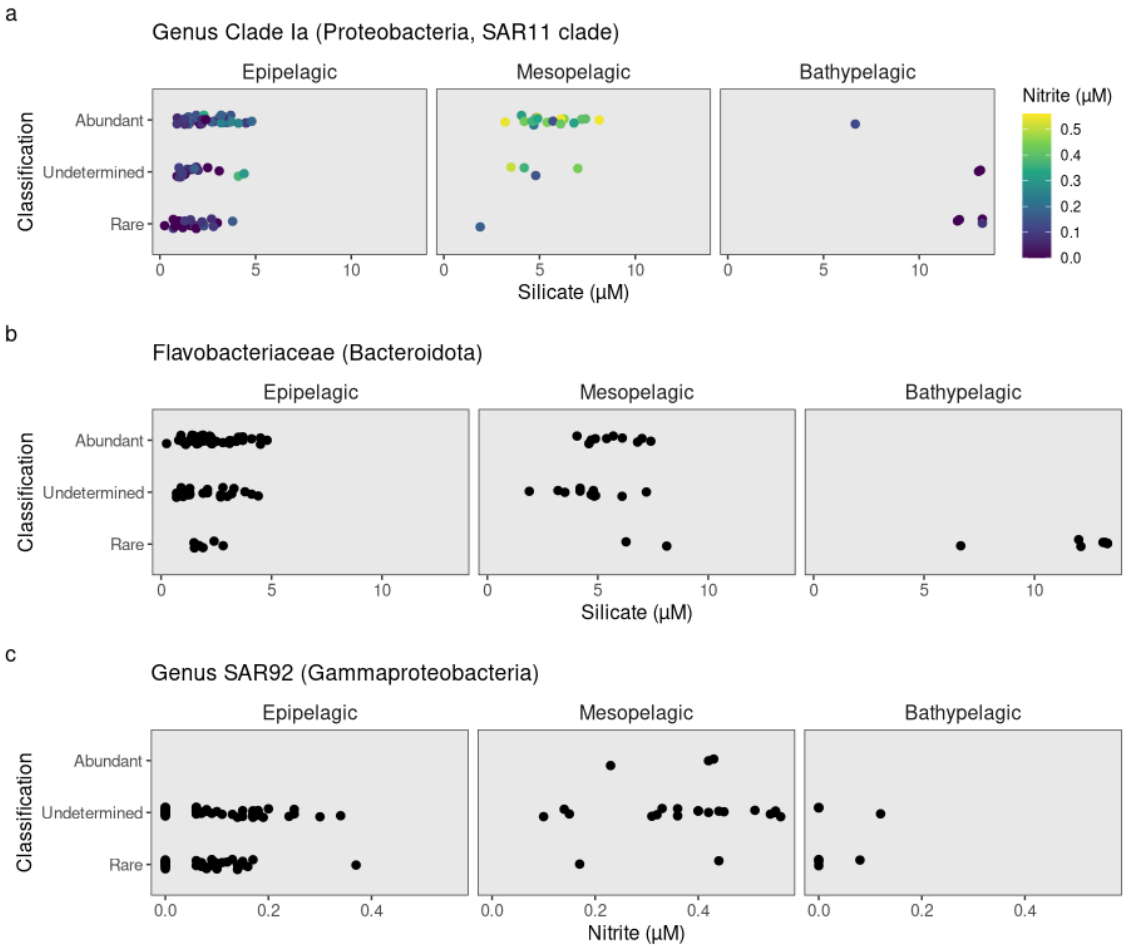
Supplementary Figure 10.52. Confidence as a function of Conviction for all rules. Rules were colored by Lift (red shading). Rules with Conviction equal to infinity were removed. n = 18 550 rules. Gray background was used to allow visibility of low Lift values. Note that most rules were overlapped in the low Lift values (white dots).



Supplementary Figure 10.53. Confidence as a function of Conviction of all rules that had antecedents statistically dependent on their consequences (n = 593 rules). The left grid shows the Improvement metric as the size of the rules, while the right grid shows Mutual Information as the size of the rules. Rules were colored by Lift (red shading). Rules with Conviction equal to infinity were removed. The gray background was used for visibility of lower Lift values.



Supplementary Figure 10.54. Boxplots with Lift values obtained for rules with each possible classification in the consequent: rare, undetermined or abundant.



Supplementary Figure 10.55. Abundance classifications of target taxa. **a**. Abundance classifications of *Clade Ia* (*Proteobacteria*, *SAR11 clade*) as a function of silicate concentration, in x axis, and colored by nitrite concentration (n = 108 samples); **b**. Abundance classifications of *Flavobacteriaceae* (*Bacteroidota*) as a function of silicate concentration (n = 108 samples); **c**. Abundance classification of genus *SAR91* (*Gammaproteobacteria*) as a function of nitrite concentrations (n = 108). All figures were divided into grids for the epipelagic, mesopelagic and bathypelagic layers.

11. Supplementary Tables

11.1. Supplementary Tables of Chapter 3

Supplementary Table 11.1. Details of sampling scheme of EMOSE dataset. Each line represents a sample, with information on sequencing strategy used (MetaB16SV4V5, MetaB18V9 and metagenome), the filter type (membrane or cartridge filter), the filtered volume (L), the filtering method (whole water or size fraction), filtration sequence (the sequence of filters used), pre-filtration step (yes or no), the pore size of the filter (either > 0.22 µm, 0.22-3 µm, 3-20 µm and > 20 µm), the planned volume (schematic representation of pooling of seawater samples for obtaining larger volumes) and the number of replicates.

Sample	Species richness	Number of paired reads	Size fraction	Sequencing strategy	Planned volume	Effected volume (L)	Filter type	Replicate	Sequencing platform
ERR2098412	354	100 000	> 0.22 µm	MetaB18SV9	4 x 2.5 L combined	10.0	Cartridge filter	R01	Hiseq 2500R
ERR2098413	352	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.1	Hiseq 2500R
ERR2098414	337	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.2	Hiseq 2500R
ERR2098415	345	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.3	Hiseq 2500R
ERR2098416	341	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.4	Hiseq 2500R
ERR2098419	372	100 000	> 0.22 µm	MetaB18SV9	10 L	10.0	Flat membrane	R01	Hiseq 2500R
ERR2098420	380	100 000	> 0.22 µm	MetaB18SV9	10 L	10.0	Flat membrane	R02	Hiseq 2500R
ERR2098421	290	100 000	> 0.22 µm	MetaB18SV9	10 L	10.0	Flat membrane	R03	Hiseq 2500R
ERR2098422	334	100 000	> 20 µm	MetaB18SV9	100 L	30.0	Flat membrane	R123	Hiseq 2500R
ERR2098423	289	100 000	0.22-3 µm	MetaB18SV9	10 L	10.0	Flat membrane	R01	Hiseq 2500R
ERR2098424	289	100 000	0.22-3 µm	MetaB18SV9	10 L	10.0	Flat membrane	R02	Hiseq 2500R
ERR2098425	338	100 000	0.22-3 µm	MetaB18SV9	10 L	10.0	Flat membrane	R03	Hiseq 2500R
ERR2098426	357	100 000	3-20 µm	MetaB18SV9	10 L	10.0	Flat membrane	R01	Hiseq 2500R
ERR2098427	341	100 000	3-20 µm	MetaB18SV9	10 L	10.0	Flat membrane	R02	Hiseq 2500R
ERR2098428	338	100 000	3-20 µm	MetaB18SV9	10 L	10.0	Flat membrane	R03	Hiseq 2500R
ERR2098429	368	100 000	> 20 µm	MetaB18SV9	100 L	120.0	Flat membrane	R01	Hiseq 2500R

ERR2098430	353	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098432	293	100 000	0.22-3 µm	MetaB18SV9	100 L	60.0	Flat membrane	R01	Hiseq 2500R
ERR2098433	311	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098435	369	100 000	3-20 µm	MetaB18SV9	100 L	60.0	Flat membrane	R01	Hiseq 2500R
ERR2098436	372	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098437	377	100 000	3-20 µm	MetaB18SV9	100 L	60.0	Flat membrane	R03	Hiseq 2500R
ERR2098438	383	100 000	> 20 µm	MetaB18SV9	5x100 L	496.0	Flat membrane	R00	Hiseq 2500R
ERR2098439	351	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R01	Hiseq 2500R
ERR2098440	366	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098441	364	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R03	Hiseq 2500R
ERR2098443	347	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R01	Hiseq 2500R
ERR2098444	331	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098445	316	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R03	Hiseq 2500R
ERR2098446	358	100 000	3-20 µm	MetaB18SV9	5x100 L	496.0	Flat membrane	R00	Hiseq 2500R
ERR2098447	381	100 000	3-20 µm	MetaB18SV9	2x50L	100.0	Flat membrane	R01	Hiseq 2500R
ERR2098448	359	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098449	364	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R03	Hiseq 2500R
ERR2098450	354	100 000	> 20 µm	MetaB18SV9	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500R
ERR2098452	312	100 000	0.22-3 µm	MetaB18SV9	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500R
ERR2098453	287	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R11	Hiseq 2500R
ERR2098454	349	100 000	3-20 µm	MetaB18SV9	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500R
ERR2098455	331	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R11	Hiseq 2500R
ERR2196988	383	100 000	3-20 µm	MetaB18SV9	8x100 L	716.0	Flat membrane	R00	MiSeq
ERR2098417	373	100 000	> 0.22 µm	MetaB18SV9	4 x 2.5 L combined	10.0	Cartridge filter	R02	Hiseq 2500R
ERR2098418	357	100 000	> 0.22 µm	MetaB18SV9	4 x 2.5 L combined	10.0	Cartridge filter	R03	Hiseq 2500R
ERR2098365	127	1 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R01	Hiseq 4000
ERR2098366	99	1 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R02	Hiseq 4000
ERR2098367	105	1 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R03	Hiseq 4000
ERR2098368	110	1 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R01	Hiseq 4000

ERR2098369	118	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.1	Hiseq 4000
ERR2098370	117	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.2	Hiseq 4000
ERR2098371	112	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.3	Hiseq 4000
ERR2098372	120	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.4	Hiseq 4000
ERR2098375	125	1 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098376	129	1 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R02	Hiseq 4000
ERR2098377	88	1 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098378	109	1 000	> 20 µm	Metagenome	100 L	30.0	Flat membrane	R123	Hiseq 4000
ERR2098379	91	1 000	0.22-3 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098381	97	1 000	0.22-3 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098382	125	1 000	3-20 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098383	128	1 000	3-20 µm	Metagenome	10 L	10.0	Flat membrane	R02	Hiseq 4000
ERR2098384	105	1 000	3-20 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098385	117	1 000	> 20 µm	Metagenome	100 L	120.0	Flat membrane	R01	Hiseq 4000
ERR2098386	113	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098387	101	1 000	> 20 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098388	98	1 000	0.22-3 µm	Metagenome	100 L	60.0	Flat membrane	R01	Hiseq 4000
ERR2098389	96	1 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098390	82	1 000	0.22-3 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098391	118	1 000	3-20 µm	Metagenome	100 L	60.0	Flat membrane	R01	Hiseq 4000
ERR2098392	119	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098393	114	1 000	3-20 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098394	125	1 000	> 20 µm	Metagenome	5x100 L	496.0	Flat membrane	R00	Hiseq 4000
ERR2098395	114	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098396	118	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098397	128	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R03	Hiseq 4000
ERR2098399	103	1 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098402	109	1 000	3-20 µm	Metagenome	5x100 L	496.0	Flat membrane	R00	Hiseq 4000
ERR2098403	108	1 000	3-20 µm	Metagenome	2x50 L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098404	119	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098405	120	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R03	Hiseq 4000

ERR2098406	116	1 000	> 20 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098407	113	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000
ERR2098408	92	1 000	0.22-3 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098409	105	1 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000
ERR2098410	129	1 000	3-20 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098411	126	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000
ERR2196983	118	1 000	> 20 µm	Metagenome	8x100 L	776.0	Flat membrane	R00	RHS4000
ERR2196984	90	1 000	0.22-3 µm	Metagenome	8x100 L	716.0	Flat membrane	R00	RHS4000
ERR2196985	120	1 000	3-20 µm	Metagenome	8x100 L	716.0	Flat membrane	R00	RHS4000
ERR2098373	98	1 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R02	Hiseq 4000
ERR2098374	111	1 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R03	Hiseq 4000

End of Supplementary Table 11.1.

Supplementary Table 11.2. List of samples from the EMOSE dataset that were discarded in the rarefaction step. For each one, the unique Sample ID is indicated, the sequencing strategy and taxonomic group.

Sample ID	Sequencing strategy	Taxonomic group
ERR2098479	MetaB16SV4V5	Prokaryotic
ERR2098489	MetaB16SV4V5	Prokaryotic
ERR2098490	MetaB16SV4V5	Prokaryotic
ERR2098431	MetaB18SV9	Protist
ERR2098434	MetaB18SV9	Protist
ERR2098442	MetaB18SV9	Protist
ERR2098451	MetaB18SV9	Protist
ERR2196986	MetaB18SV9	Protist
ERR2196987	MetaB18SV9	Protist
ERR2098365	Metagenome	Prokaryotic
ERR2098366	Metagenome	Prokaryotic
ERR2098367	Metagenome	Prokaryotic
ERR2098368	Metagenome	Prokaryotic
ERR2098369	Metagenome	Prokaryotic
ERR2098370	Metagenome	Prokaryotic
ERR2098371	Metagenome	Prokaryotic
ERR2098372	Metagenome	Prokaryotic
ERR2098375	Metagenome	Prokaryotic
ERR2098376	Metagenome	Prokaryotic
ERR2098377	Metagenome	Prokaryotic
ERR2098378	Metagenome	Prokaryotic
ERR2098379	Metagenome	Prokaryotic
ERR2098380	Metagenome	Prokaryotic
ERR2098381	Metagenome	Prokaryotic
ERR2098382	Metagenome	Prokaryotic
ERR2098383	Metagenome	Prokaryotic
ERR2098384	Metagenome	Prokaryotic
ERR2098385	Metagenome	Prokaryotic
ERR2098386	Metagenome	Prokaryotic
ERR2098387	Metagenome	Prokaryotic
ERR2098388	Metagenome	Prokaryotic
ERR2098389	Metagenome	Prokaryotic
ERR2098390	Metagenome	Prokaryotic
ERR2098391	Metagenome	Prokaryotic
ERR2098392	Metagenome	Prokaryotic
ERR2098393	Metagenome	Prokaryotic
ERR2098394	Metagenome	Prokaryotic
ERR2098395	Metagenome	Prokaryotic
ERR2098396	Metagenome	Prokaryotic
ERR2098397	Metagenome	Prokaryotic
ERR2098398	Metagenome	Prokaryotic
ERR2098399	Metagenome	Prokaryotic
ERR2098400	Metagenome	Prokaryotic
ERR2098401	Metagenome	Prokaryotic
ERR2098402	Metagenome	Prokaryotic
ERR2098403	Metagenome	Prokaryotic
ERR2098404	Metagenome	Prokaryotic
ERR2098405	Metagenome	Prokaryotic
ERR2098406	Metagenome	Prokaryotic

ERR2098407	Metagenome	Prokaryotic
ERR2098408	Metagenome	Prokaryotic
ERR2098409	Metagenome	Prokaryotic
ERR2098410	Metagenome	Prokaryotic
ERR2098411	Metagenome	Prokaryotic
ERR2196983	Metagenome	Prokaryotic
ERR2196984	Metagenome	Prokaryotic
ERR2196985	Metagenome	Prokaryotic
ERR2098373	Metagenome	Prokaryotic
ERR2098374	Metagenome	Prokaryotic
ERR2098380	Metagenome	Protist
ERR2098398	Metagenome	Protist
ERR2098400	Metagenome	Protist
ERR2098401	Metagenome	Protist

End of Supplementary Table 11.2.

Supplementary Table 11.3. Summary of the sequencing results of the EMOSE dataset. Centrality metrics are provided for the number of valid raw reads, the final number of high quality paired-reads and their proportions, values were divided by sequencing strategy and taxonomic groups, where applicable.

Reads	Sequencing strategy	Taxonomic group	Median	IQR	Minimum	Maximum	Mean	Standard deviation	n
Raw paired end	MetaB16SV4V5	Prokaryotic	1 462 584.50	267 471.75	714 103.00	2 841 890.00	1 459 043.70	353 350.19	60
Raw paired end	MetaB18SV9	Protist	1 322 612.00	319 425.00	670 823.00	3 876 463.00	1 498 954.40	559 604.10	47
Raw paired end	Metagenome	Prokaryotic/Protist	58 122 461.00	12 863 822.75	36 573 050.00	123 310 150.00	61 368 376.94	17 259 833.06	50
High quality paired	MetaB16SV4V5	Prokaryotic	660 878.00	438 932.50	169 945.00	1 517 860.00	680 517.83	286 603.87	60
High quality paired	Metagenome	Prokaryotic	23 376.50	43 212.75	628.00	98 749.00	28 359.70	23 219.32	50
High quality paired	MetaB18SV9	Protist	173 379.00	348 105.00	66 912.00	1 889 838.00	361 405.11	406 019.89	47
High quality paired	Metagenome	Protist	244 105.00	2 807.25	396.00	9 160.00	3 276.88	2 097.24	50
Percentage high quality/raw	MetaB16SV4V5	Prokaryotic	50.69%	30.48%	13.38%	72.47%	46.92%	16.49%	60
Percentage high quality/raw	Metagenome	Prokaryotic	0.04%	0.07%	<0.01%	0.10%	0.05%	0.03%	50
Percentage high quality/raw	MetaB18SV9	Protist	13.05%	26.13%	6.21%	60.34%	21.40%	16.65%	47
Percentage high quality/raw	Metagenome	Protist	0.01%	0.00%	0.00%	0.01%	0.01%	0.00%	50

Supplementary Table 11.4. Number of prokaryotic taxonomic lineages obtained in each sample (species richness) of the EMOSE dataset. Relevant sample metadata was added.

Sample	Species richness	Number of paired reads	Size fraction	Strategy	Planned volume	Effected volume (L)	Filter type	Replicate	Sequencing platform
ERR2098456	376	250 000	> 0.22 µm	MetaB16SV4V5	4 x 2.5 L combined	10.0	Cartridge filter	R01	Hiseq 2500 Rapid
ERR2098457	374	250 000	> 0.22 µm	MetaB16SV4V5	2.5 L	2.5	Cartridge filter	R01.1	Hiseq 2500 Rapid
ERR2098458	373	250 000	> 0.22 µm	MetaB16SV4V5	2.5 L	2.5	Cartridge filter	R01.2	Hiseq 2500 Rapid
ERR2098459	356	250 000	> 0.22 µm	MetaB16SV4V5	2.5 L	2.5	Cartridge filter	R01.3	Hiseq 2500 Rapid
ERR2098460	316	250 000	> 0.22 µm	MetaB16SV4V5	2.5 L	2.5	Cartridge filter	R01.4	Hiseq 2500 Rapid
ERR2098461	319	250 000	> 0.22 µm	MetaB16SV4V5	2.5 L	2.5	Cartridge filter	R01.4	Hiseq 2500 Rapid
ERR2098465	330	250 000	> 0.22 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098466	314	250 000	> 0.22 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098467	372	250 000	> 0.22 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R02	Hiseq 2500 Rapid
ERR2098468	297	250 000	> 0.22 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098469	300	250 000	> 0.22 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098470	629	250 000	> 20 µm	MetaB16SV4V5	100 L	30.0	Flat membrane	R123	Hiseq 2500 Rapid
ERR2098471	361	250 000	0.22-3 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098472	375	250 000	0.22-3 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R02	Hiseq 2500 Rapid
ERR2098473	341	250 000	0.22-3 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098474	521	250 000	3-20 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098475	493	250 000	3-20 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R02	Hiseq 2500 Rapid

ERR2098476	525	250 000	3-20 µm	MetaB16SV4V5	10 L	10.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098477	552	250 000	> 20 µm	MetaB16SV4V5	100 L	120.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098478	588	250 000	> 20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R02	Hiseq 2500 Rapid
ERR2098480	335	250 000	0.22-3 µm	MetaB16SV4V5	100 L	60.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098481	325	250 000	0.22-3 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R02	Hiseq 2500 Rapid
ERR2098482	315	250 000	0.22-3 µm	MetaB16SV4V5	100 L	60.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098483	453	250 000	3-20 µm	MetaB16SV4V5	100 L	60.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098484	418	250 000	3-20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R02	Hiseq 2500 Rapid
ERR2098485	455	250 000	3-20 µm	MetaB16SV4V5	100 L	60.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098486	581	250 000	> 20 µm	MetaB16SV4V5	5x100 L	496.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098487	487	250 000	> 20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098488	483	250 000	> 20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098491	547	250 000	> 20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098492	319	250 000	0.22-3 µm	MetaB16SV4V5	5x100 L	496.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098493	339	250 000	0.22-3 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098494	335	250 000	0.22-3 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R02	Hiseq 2500 Rapid
ERR2098495	382	250 000	0.22-3 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098496	423	250 000	3-20 µm	MetaB16SV4V5	5x100 L	496.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098497	440	250 000	3-20 µm	MetaB16SV4V5	2x50L	100.0	Flat membrane	R01	Hiseq 2500 Rapid

ERR2098498	467	250 000	3-20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R02	Hiseq 2500 Rapid
ERR2098499	370	250 000	3-20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098500	360	250 000	3-20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098501	654	250 000	> 20 µm	MetaB16SV4V5	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098502	622	250 000	> 20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R11	Hiseq 2500 Rapid
ERR2098503	297	250 000	0.22-3 µm	MetaB16SV4V5	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098504	318	250 000	0.22-3 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R11	Hiseq 2500 Rapid
ERR2098505	447	250 000	3-20 µm	MetaB16SV4V5	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098506	468	250 000	3-20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R11	Hiseq 2500 Rapid
ERR2098509	656	250 000	> 20 µm	MetaB16SV4V5	100 L	30.0	Flat membrane	R123	Hiseq 2500 Rapid
ERR2098510	592	250 000	> 20 µm	MetaB16SV4V5	100 L	120.0	Flat membrane	R01	Hiseq 2500 Rapid
ERR2098511	518	250 000	> 20 µm	MetaB16SV4V5	5x100 L	496.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098512	513	250 000	> 20 µm	MetaB16SV4V5	5x100 L	496.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2098513	529	250 000	> 20 µm	MetaB16SV4V5	100 L	100.0	Flat membrane	R03	Hiseq 2500 Rapid
ERR2098514	670	250 000	> 20 µm	MetaB16SV4V5	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500 Rapid
ERR2196989	656	250 000	> 20 µm	MetaB16SV4V5	8x100 L	776.0	Flat membrane	R00	MiSeq
ERR2196990	307	250 000	0.22-3 µm	MetaB16SV4V5	8x100 L	716.0	Flat membrane	R00	MiSeq
ERR2196991	443	250 000	3-20 µm	MetaB16SV4V5	8x100 L	716.0	Flat membrane	R00	MiSeq
ERR2098463	338	250 000	> 0.22 µm	MetaB16SV4V5	4 x 2.5 L combined	10.0	Cartridge filter	R02	Hiseq 2500 Rapid
ERR2098464	382	250 000	> 0.22 µm	MetaB16SV4V5	4 x 2.5 L combined	10.0	Cartridge filter	R03	Hiseq 2500 Rapid

ERR2098462	338	250 000	> 0.22 µm	MetaB16SV4V5	4 x 2.5 L combined	10.0	Cartridge filter	R02	Hiseq 2500 Rapid
ERR2098365	187	10 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R01	Hiseq 4000
ERR2098366	163	10 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R02	Hiseq 4000
ERR2098367	182	10 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R03	Hiseq 4000
ERR2098368	169	10 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R01	Hiseq 4000
ERR2098369	174	10 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.1	Hiseq 4000
ERR2098370	165	10 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.2	Hiseq 4000
ERR2098371	172	10 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.3	Hiseq 4000
ERR2098372	173	10 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.4	Hiseq 4000
ERR2098375	180	10 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098376	167	10 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R02	Hiseq 4000
ERR2098377	153	10 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098379	171	10 000	0.22-3 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098380	150	10 000	0.22-3 µm	Metagenome	10 L	10.0	Flat membrane	R02	Hiseq 4000
ERR2098381	157	10 000	0.22-3 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098388	159	10 000	0.22-3 µm	Metagenome	100 L	60.0	Flat membrane	R01	Hiseq 4000
ERR2098389	157	10 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098390	162	10 000	0.22-3 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098391	208	10 000	3-20 µm	Metagenome	100 L	60.0	Flat membrane	R01	Hiseq 4000
ERR2098392	181	10 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098393	217	10 000	3-20 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098398	151	10 000	0.22-3 µm	Metagenome	5x100 L	496.0	Flat membrane	R00	Hiseq 4000
ERR2098399	155	10 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098400	155	10 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098401	149	10 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R03	Hiseq 4000
ERR2098402	187	10 000	3-20 µm	Metagenome	5x100 L	496.0	Flat membrane	R00	Hiseq 4000
ERR2098403	195	10 000	3-20 µm	Metagenome	2x50L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098405	195	10 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R03	Hiseq 4000
ERR2098408	149	10 000	0.22-3 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098409	137	10 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000

ERR2098410	241	10 000	3-20 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098411	216	10 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000
ERR2196984	145	10 000	0.22-3 µm	Metagenome	8x100 L	716.0	Flat membrane	R00	RHS4000
ERR2196985	204	10 000	3-20 µm	Metagenome	8x100 L	716.0	Flat membrane	R00	RHS4000
ERR2098373	156	10 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R02	Hiseq 4000
ERR2098374	171	10 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R03	Hiseq 4000

End of Supplementary Table 11.4.

Supplementary Table 11.5. Details on the statistical tests for the prokaryotic species richness. The response variable was always the prokaryotic species richness, and significance was attested for with an alpha of 0.05.

Statistical test	Sequencing strategy	Volume (L)	Variable 1	Variable 1 (n)	Variable 2	Variable 2 (n)	Statistic	p-value	Adjusted p-value	Significance of adjusted p-value	n	df
Mann-Whitney U test	Metagenome	10	> 0.22 µm	3	0.22-3 µm	3	6.0	0.700	0.700	ns	NA	
Kruskal-Wallis	MetaB16SV4V5	10			NA		7.5	0.024	0.024	Significant	11	2
Dunn test	MetaB16SV4V5	10	> 0.22 µm	5	0.22-3 µm	3	1.2	0.226	0.678	ns	NA	
			> 0.22 µm	5	3-20 µm	3	2.7	0.006	0.019	Significant		
			0.22-3 µm	3	3-20 µm	3	1.4	0.176	0.527	ns		
Kruskal-Wallis	MetaB16SV4V5	10			NA		13.4	0.001	0.001	Significant	17	2
Dunn test	MetaB16SV4V5	10	0.22-3 µm	5	3-20 µm	6	1.6	0.119	0.357	ns	NA	
			0.22-3 µm	5	> 20 µm	6	3.6	0.000	0.001	Significant		
			3-20 µm	6	> 20 µm	6	2.2	0.030	0.089	ns		
Mann-Whitney U test	Metagenome	100	0.22-3 µm	5	3-20 µm	4	0.0	0.019	0.019	Significant		
Mann-Whitney U test	MetaB16SV4V5	100	Flat membrane	5	Cartridge filter	4	2.0	0.065	0.130	ns		
Mann-Whitney U test	Metagenome	100	Flat membrane	3	Cartridge filter	3	4.0	1.000	1.000	ns	NA	
Mann-Whitney U test	MetaB16SV4V5	100	2.5 L whole	5	4 x 2.5 L combined	4	6.0	0.389	0.778	ns		
Mann-Whitney U test	Metagenome	100	2.5 L whole	4	4 x 2.5 L combined	3	10.0	0.229	0.458	ns		

ns – not significant; NA – not applicable; df – degrees of freedom.

Supplementary Table 11.6 – Details on the PERMANOVA analysis of prokaryotic communities and betadisper permutations for verification of homogeneity of variance prerequisite. Alpha was 0.05 for all tests, and 999 permutations were used for all tests.

Statistical test	Taxonomic group	Sequencing strategy	Model formula	df	Sum of squares	R squared	Mean squared	F statistic	p-value	Significance
PERMANOVA	Prokaryotic	MetaB16SV4V5	Volume	1	0.104	0.034	NA	6.2362	**0.0020**	Significant
			Size fraction	3	2.081	0.689		41.5764	**0.0010**	Significant
			Volume:Size fraction	3	0.018	0.006		0.3560	**0.9090**	ns
			Residual	49	0.818	0.271				
			Total	56	3.020	1.000				
Betadisper	Prokaryotic	MetaB16SV4V5	Size fraction	3	0.098	NA	0.0326	10.7587	**0.0010**	Significant
			Residual	53	0.161		0.0030			
			Volume	9	0.232	NA	0.0258	3.0068	**0.0220**	Significant
			Residual	47	0.403		0.0086			
PERMANOVA	Prokaryotic	Metagenome	volume	1	0.012	0.073	NA	3.7518	**0.0070**	Significant
			Size fraction	2	0.052	0.311		8.0073	**0.0010**	Significant
			Volume:size fraction	2	0.009	0.052		1.3505	**0.2050**	ns
			Residual	29	0.094	0.563				
			Total	34	0.167	1.000				
Betadisper	Prokaryotic	Metagenome	Size fraction	2	0.001	NA	0.0005	1.2619	**0.2940**	ns
			Residual	32	0.014		0.0004			
			Volume	7	0.007	NA	0.0010	3.0012	**0.0200**	Significant
			Residual	27	0.009		0.0003			

ns – not significant; df – degrees of freedom; NA – not applicable.

Supplementary Table 11.7. Details of the Tukey test for significance between differences in prokaryotic communities' distance to centroids of size fractions, calculated by betadisper. Significance was attested by an alpha of 0.05.

Sequencing strategy	Taxonomy	Comparison	Difference to centroid	Lower	Upper	p-value adjusted	Significance
MetaB16SV4V5	Prokaryotes	0.22-3 µm-> 0.22 µm	-0.00206	-0.05830	0.05417	0.99967	ns
		3-20 µm-> 0.22 µm	0.09827	0.04309	0.15346	0.00010	Significant
		> 20 µm-> 0.22 µm	0.05741	0.00398	0.11084	0.03066	Significant
		3-20 µm-0.22-3 µm	0.10034	0.04410	0.15657	0.00010	Significant
		> 20 µm-0.22-3 µm	0.05948	0.00496	0.11399	0.02742	Significant
		> 20 µm-3-20 µm	-0.04086	-0.09430	0.01257	0.19060	ns
Metagenome		0.22-3 µm-> 0.22 µm	0.00815	-0.01177	0.02807	0.57907	ns
		3-20 µm-> 0.22 µm	0.01391	-0.00812	0.03593	0.28097	ns
		3-20 µm-0.22-3 µm	0.00576	-0.01627	0.02779	0.79780	ns

ns – not significant.

Supplementary Table 11.8. Number of protist taxonomic lineages obtained in each sample (Species richness). Relevant sample metadata was added.

Sample	Species richness	Number of paired reads	Size fraction	Sequencing strategy	Planned volume	Effected volume (L)	Filter type	Replicate	Sequencing platform
ERR2098412	354	100 000	> 0.22 µm	MetaB18SV9	4 x 2.5 L combined	10.0	Cartridge filter	R01	Hiseq 2500R
ERR2098413	352	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.1	Hiseq 2500R
ERR2098414	337	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.2	Hiseq 2500R
ERR2098415	345	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.3	Hiseq 2500R
ERR2098416	341	100 000	> 0.22 µm	MetaB18SV9	2.5 L	2.5	Cartridge filter	R01.4	Hiseq 2500R
ERR2098419	372	100 000	> 0.22 µm	MetaB18SV9	10 L	10.0	Flat membrane	R01	Hiseq 2500R
ERR2098420	380	100 000	> 0.22 µm	MetaB18SV9	10 L	10.0	Flat membrane	R02	Hiseq 2500R
ERR2098421	290	100 000	> 0.22 µm	MetaB18SV9	10 L	10.0	Flat membrane	R03	Hiseq 2500R
ERR2098422	334	100 000	> 20 µm	MetaB18SV9	100 L	30.0	Flat membrane	R123	Hiseq 2500R
ERR2098423	289	100 000	0.22-3 µm	MetaB18SV9	10 L	10.0	Flat membrane	R01	Hiseq 2500R

ERR2098424	289	100 000	0.22-3 µm	MetaB18SV9	10 L	10.0	Flat membrane	R02	Hiseq 2500R
ERR2098425	338	100 000	0.22-3 µm	MetaB18SV9	10 L	10.0	Flat membrane	R03	Hiseq 2500R
ERR2098426	357	100 000	3-20 µm	MetaB18SV9	10 L	10.0	Flat membrane	R01	Hiseq 2500R
ERR2098427	341	100 000	3-20 µm	MetaB18SV9	10 L	10.0	Flat membrane	R02	Hiseq 2500R
ERR2098428	338	100 000	3-20 µm	MetaB18SV9	10 L	10.0	Flat membrane	R03	Hiseq 2500R
ERR2098429	368	100 000	> 20 µm	MetaB18SV9	100 L	120.0	Flat membrane	R01	Hiseq 2500R
ERR2098430	353	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098432	293	100 000	0.22-3 µm	MetaB18SV9	100 L	60.0	Flat membrane	R01	Hiseq 2500R
ERR2098433	311	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098435	369	100 000	3-20 µm	MetaB18SV9	100 L	60.0	Flat membrane	R01	Hiseq 2500R
ERR2098436	372	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098437	377	100 000	3-20 µm	MetaB18SV9	100 L	60.0	Flat membrane	R03	Hiseq 2500R
ERR2098438	383	100 000	> 20 µm	MetaB18SV9	5x100 L	496.0	Flat membrane	R00	Hiseq 2500R
ERR2098439	351	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R01	Hiseq 2500R
ERR2098440	366	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098441	364	100 000	> 20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R03	Hiseq 2500R
ERR2098443	347	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R01	Hiseq 2500R
ERR2098444	331	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098445	316	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R03	Hiseq 2500R
ERR2098446	358	100 000	3-20 µm	MetaB18SV9	5x100 L	496.0	Flat membrane	R00	Hiseq 2500R
ERR2098447	381	100 000	3-20 µm	MetaB18SV9	2x50L	100.0	Flat membrane	R01	Hiseq 2500R
ERR2098448	359	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R02	Hiseq 2500R
ERR2098449	364	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R03	Hiseq 2500R
ERR2098450	354	100 000	> 20 µm	MetaB18SV9	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500R
ERR2098452	312	100 000	0.22-3 µm	MetaB18SV9	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500R
ERR2098453	287	100 000	0.22-3 µm	MetaB18SV9	100 L	100.0	Flat membrane	R11	Hiseq 2500R
ERR2098454	349	100 000	3-20 µm	MetaB18SV9	10x100 L	1 000.0	Flat membrane	R00	Hiseq 2500R
ERR2098455	331	100 000	3-20 µm	MetaB18SV9	100 L	100.0	Flat membrane	R11	Hiseq 2500R
ERR2196988	383	100 000	3-20 µm	MetaB18SV9	8x100 L	716.0	Flat membrane	R00	MiSeq
ERR2098417	373	100 000	> 0.22 µm	MetaB18SV9	4 x 2.5 L combined	10.0	Cartridge filter	R02	Hiseq 2500R

ERR2098418	357	100 000	> 0.22 µm	MetaB18SV9	4 x 2.5 L combined	10.0	Cartridge filter	R03	Hiseq 2500R
ERR2098365	127	1 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R01	Hiseq 4000
ERR2098366	99	1 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R02	Hiseq 4000
ERR2098367	105	1 000	> 0.22 µm	Metagenome	1L	1.0	Cartridge filter	R03	Hiseq 4000
ERR2098368	110	1 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R01	Hiseq 4000
ERR2098369	118	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.1	Hiseq 4000
ERR2098370	117	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.2	Hiseq 4000
ERR2098371	112	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.3	Hiseq 4000
ERR2098372	120	1 000	> 0.22 µm	Metagenome	2.5 L	2.5	Cartridge filter	R01.4	Hiseq 4000
ERR2098375	125	1 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098376	129	1 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R02	Hiseq 4000
ERR2098377	88	1 000	> 0.22 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098378	109	1 000	> 20 µm	Metagenome	100 L	30.0	Flat membrane	R123	Hiseq 4000
ERR2098379	91	1 000	0.22-3 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098381	97	1 000	0.22-3 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098382	125	1 000	3-20 µm	Metagenome	10 L	10.0	Flat membrane	R01	Hiseq 4000
ERR2098383	128	1 000	3-20 µm	Metagenome	10 L	10.0	Flat membrane	R02	Hiseq 4000
ERR2098384	105	1 000	3-20 µm	Metagenome	10 L	10.0	Flat membrane	R03	Hiseq 4000
ERR2098385	117	1 000	> 20 µm	Metagenome	100 L	120.0	Flat membrane	R01	Hiseq 4000
ERR2098386	113	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098387	101	1 000	> 20 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098388	98	1 000	0.22-3 µm	Metagenome	100 L	60.0	Flat membrane	R01	Hiseq 4000
ERR2098389	96	1 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098390	82	1 000	0.22-3 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098391	118	1 000	3-20 µm	Metagenome	100 L	60.0	Flat membrane	R01	Hiseq 4000
ERR2098392	119	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098393	114	1 000	3-20 µm	Metagenome	100 L	60.0	Flat membrane	R03	Hiseq 4000
ERR2098394	125	1 000	> 20 µm	Metagenome	5x100 L	496.0	Flat membrane	R00	Hiseq 4000
ERR2098395	114	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098396	118	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000

ERR2098397	128	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R03	Hiseq 4000
ERR2098399	103	1 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098402	109	1 000	3-20 µm	Metagenome	5x100 L	496.0	Flat membrane	R00	Hiseq 4000
ERR2098403	108	1 000	3-20 µm	Metagenome	2x50L	100.0	Flat membrane	R01	Hiseq 4000
ERR2098404	119	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R02	Hiseq 4000
ERR2098405	120	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R03	Hiseq 4000
ERR2098406	116	1 000	> 20 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098407	113	1 000	> 20 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000
ERR2098408	92	1 000	0.22-3 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098409	105	1 000	0.22-3 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000
ERR2098410	129	1 000	3-20 µm	Metagenome	10x100 L	1 000.0	Flat membrane	R00	Hiseq 4000
ERR2098411	126	1 000	3-20 µm	Metagenome	100 L	100.0	Flat membrane	R11	Hiseq 4000
ERR2196983	118	1 000	> 20 µm	Metagenome	8x100 L	776.0	Flat membrane	R00	RHS4000
ERR2196984	90	1 000	0.22-3 µm	Metagenome	8x100 L	716.0	Flat membrane	R00	RHS4000
ERR2196985	120	1 000	3-20 µm	Metagenome	8x100 L	716.0	Flat membrane	R00	RHS4000
ERR2098373	98	1 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R02	Hiseq 4000
ERR2098374	111	1 000	> 0.22 µm	Metagenome	4 x 2.5 L combined	10.0	Cartridge filter	R03	Hiseq 4000

End of Supplementary Table 11.8.

Supplementary Table 11.9. Details on the statistical tests for the protists species richness. The response variable was always the protists species richness, and significance was attested for with an alpha of 0.05.

Test	Sequencing strategy	Taxonomic group	Volume (L)	Variable 1	Variable 1 (n)	Variable 2	Variable 2 (n)	Statistic	p-value	Adjusted p-value	Significance	n	df
Kruskal-Wallis	MetaB18SV9	Protist	10	NA	NA	NA	NA	3.95	0.13	0.28	ns	9	2
Kruskal-Wallis	Metagenome	Protist	10	NA	NA	NA	NA	1.91	0.38	0.77	ns	8	2
Kruskal-Wallis	MetaB18SV9	Protist	10	NA	NA	NA	NA	7.96	0.02	0.04	Significant	14	2
Dunn	MetaB18SV9	Protist	10	0.22-3 m	5	3-20 µm	5	2.61	0.01	0.05	ns	NA	
				0.22-3 µm	5	> 20 µm	4	2.17	0.03	0.18	ns		
				3-20 µm	5	> 20 µm	4	-0.29	0.79	1.00	ns		
Kruskal-Wallis	Metagenome	Protist	100	NA	NA	NA	NA	6.79	0.03	0.07	ns	13	2
Mann-Whitney U test	MetaB18SV9	Protist	100	Flat membrane	3	Cartridge filter	3	5.00	1.00	1.00	ns	NA	
Mann-Whitney U test	Metagenome	Protist	100	Flat membrane	3	Cartridge filter	3	6.00	0.70	1.00	ns	NA	
Mann-Whitney U test	MetaB18SV9	Protist	100	2.5 L whole	4	4 x 2.5 L combined	3	0.00	0.06	0.11	ns	NA	
Mann-Whitney U test	Metagenome	Protist	100	2.5 L whole	4	4 x 2.5 L combined	3	12.00	0.06	0.11	ns	NA	

ns – not significant; NA – Not applicable; df – degrees of freedom.

Supplementary Table 11.10. Details on the PERMANOVA analysis of protist communities and betadisper permutations for verification of variance prerequisite. Significance was attested using an alpha of 0.05.

Test	Taxonomic group	Sequencing strategy	Model formula	df	Sum of squares	r-squared	Mean squared	F statistic	p-value	Significance
PERMANOVA	Protist	MetaB18SV9	Volume	1	0.06	0.0184	NA	3.456	0.028	Significant
			Size fraction	3	2.50	0.7925		49.493	0.001	Significant
			Volume:Size fraction	3	0.04	0.0129		0.808	0.538	ns
			Residual	33	0.56	0.1761	NA			
			Total	40	3.16	1.0000				
Betadisper	Protist	MetaB18SV9	Size fraction	3	0.01	NA	0.002	0.583	0.652	ns
			Residual	37	0.11		0.003			
			Volume	8	0.37	NA	0.046	7.135	0.001	Significant
			Residual	32	0.21		0.006			
PERMANOVA	Protist	Metagenome	Volume	1	0.08	0.0195	NA	2.368	0.063	ns
			Size fraction	3	2.67	0.6531		26.475	0.001	Significant
			Volume:Size fraction	3	0.06	0.0149		0.605	0.829	ns
			Residual	38	1.28	0.3125	NA			
			Total	45	4.08	1.0000				
Betadisper	Protist	Metagenome	Size fraction	3	0.01	NA	0.005	2.054	0.083	ns
			Residual	42	0.10		0.002			
			Volume	10	0.42	NA	0.042	10.432	0.001	Significant
			Residual	35	0.14		0.004			

ns – not significant; NA – Not applicable; df – degrees of freedom.

Supplementary Table 11.11. Details of the Tukey test for significance between differences in protist communities' distance to centroids of size fractions, calculated by betadisper.

Sequencing strategy	Taxonomy	Comparison	Difference to centroid	Lower	Upper	p-value adjusted	Significance	Alpha
MetaB18SV9	Protist	0.22-3 µm > 0.22 µm	-0.00967	-0.076	0.06	0.98	ns	0.05
		3-20 µm > 0.22 µm	-0.00555	-0.068	0.06	0.99	ns	0.05
		> 20 µm > 0.22 µm	0.02231	-0.048	0.09	0.83	ns	0.05
		3-20 µm-0.22-3 µm	0.00412	-0.058	0.07	0.99	ns	0.05
		> 20 µm-0.22-3 µm	0.03198	-0.038	0.10	0.62	ns	0.05
		> 20 µm-3-20 µm	0.02786	-0.039	0.09	0.68	ns	0.05
Metagenome		0.22-3 µm-> 0.22 µm	-0.00303	-0.059	0.05	0.99	ns	0.05
		3-20 µm-> 0.22 µm	-0.02199	-0.073	0.03	0.66	ns	0.05
		> 20 µm-> 0.22 µm	0.02716	-0.026	0.08	0.53	ns	0.05
		3-20 µm-0.22-3 µm	-0.01896	-0.075	0.04	0.80	ns	0.05
		> 20 µm-0.22-3 µm	0.03020	-0.028	0.09	0.52	ns	0.05
		> 20 µm-3-20 µm	0.04915	-0.004	0.10	0.08	ns	0.05

ns – not significant.

11.2. Supplementary Tables of Chapter 4

Supplementary Table 11.12. List of samples removed (< 10 000 reads) of MOSJ 2019.

Sample
M19_105
M19_5
M19_56
M19_60
M19_70

Supplementary Table 11.13. Summary statistics of sequencing results of MOSJ 2019.

Strategy	Step	Maximum	Minimum	Mean	Standard deviation	Median	IQR	n
PacBio	CCS	53 770	20 119	33 919	11 039	34 182	18 610	13
PacBio	High quality	29 839	10 221	16 987	6 799	16 779	12 998	13
Illumina	Input	176 856	14 874	72 774	42 138	67 997	34 824	13
Illumina	High quality	134 744	12 190	55 032	32 437	53 782	29 845	13

Supplementary Table 11.14. Details of Mann-Whitney U test (also known as Wilcoxon test), used to compare the distribution of two independent variables. Alpha of 0.05 for all tests.

Taxonomic level	Database	Response variable	Group1	Group2	n1	n2	Statistic (Mann-Whitney U test)	*p* value	Significance	Effect size (r)	Magnitude
Not rarefied											
Kingdom	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	54.5	1.30E-01	ns	0.30	Moderate
Kingdom	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	104.0	7.92E-02	ns	0.35	Moderate
Phylum	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Phylum	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	140.0	2.98E-03	Significant	0.59	Large
Class	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Class	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	153.0	4.67E-04	Significant	0.69	Large
Order	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Order	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	145.0	2.02E-03	Significant	0.61	Large
Family	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	166.0	1.35E-06	Significant	0.82	Large
Family	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	140.0	3.49E-03	Significant	0.56	Large
Genus	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Genus	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Species	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Species	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	58.0	1.86E-01	ns	0.27	Small
Rarefied											
Kingdom	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	49.0	7.17E-02	ns	0.36	Moderate

Kingdom	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	97.5	1.66E-01	ns	0.28	Small
Phylum	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	167.0	7.69E-07	Significant	0.83	Large
Phylum	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	116.0	7.71E-02	Ns	0.35	Moderate
Class	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.64E-05	Significant	0.85	Large
Class	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	145.0	2.02E-03	Significant	0.61	Large
Order	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Order	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	139.0	5.46E-03	Significant	0.55	Large
Family	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	166.0	1.35E-06	Significant	0.82	Large
Family	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	137.0	6.08E-03	Significant	0.53	Large
Genus	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Genus	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Species	GTDB	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	169.0	1.92E-07	Significant	0.85	Large
Species	Silva	Percentage of classified ASVs	Full-length 16S rRNA gene	V4V5 16S rRNA gene	13	13	55.5	0.144	ns	0.29	Small

End of Supplementary Table 11.14.

Supplementary Table 11.15. Details on linear regressions, including p-values and R-squared values.

Not rarefied									
Sequencing approach	**Database**	**Taxonomic level**	**Score**	**Linear regression**	**p-value**	**alpha**	**Significance**	**R squared**	**n**
V4V5 16S rRNA gene	GTDB	Phylum	ASV richness	y = 0.1062x + 240.2281	0.055	0.05	ns	0.2945	13
V4V5 16S rRNA gene	GTDB	Phylum	Shannon	y = 0.0003x + 3.7291	0.611	0.05	ns	0.0244	13
V4V5 16S rRNA gene	GTDB	Phylum	Simpson	y = 0x + 0.881	0.962	0.05	ns	0.0002	13
V4V5 16S rRNA gene	Silva	Phylum	ASV richness	y = 0.1087x + 259.7195	0.066	0.05	ns	0.2749	13
V4V5 16S rRNA gene	Silva	Phylum	Shannon	y = 0.0003x + 3.7785	0.629	0.05	ns	0.0219	13
V4V5 16S rRNA gene	Silva	Phylum	Simpson	y = 0x + 0.8828	0.954	0.05	ns	0.0003	13
full-length 16S rRNA gene	Silva	Phylum	ASV richness	y = 0.1571x + 219.6865	0.006	0.05	Significant	0.5143	13
full-length 16S rRNA gene	Silva	Phylum	Shannon	y = 0.0006x + 3.2461	0.323	0.05	ns	0.0889	13
full-length 16S rRNA gene	Silva	Phylum	Simpson	y = 0x + 0.7781	0.863	0.05	ns	0.0028	13
full-length 16S rRNA gene	GTDB	Phylum	ASV richness	y = 0.1581x + 214.5926	0.005	0.05	Significant	0.5271	13
full-length 16S rRNA gene	GTDB	Phylum	Shannon	y = 0.0005x + 4.0856	0.135	0.05	ns	0.1911	13
full-length 16S rRNA gene	GTDB	Phylum	Simpson	y = 0x + 0.9407	0.763	0.05	ns	0.0086	13
V4V5 16S rRNA gene	GTDB	Order	ASV richness	y = 0.0997x + 270.8249	0.127	0.05	ns	0.1984	13
V4V5 16S rRNA gene	GTDB	Order	Shannon	y = 0.0003x + 3.704	0.611	0.05	ns	0.0243	13
V4V5 16S rRNA gene	GTDB	Order	Simpson	y = 0x + 0.8796	0.966	0.05	ns	0.0002	13
V4V5 16S rRNA gene	Silva	Order	ASV richness	y = 0.0837x + 269.5693	0.196	0.05	ns	0.1467	13
V4V5 16S rRNA gene	Silva	Order	Shannon	y = 0.0002x + 4.1074	0.316	0.05	ns	0.0911	13
V4V5 16S rRNA gene	Silva	Order	Simpson	y = 0x + 0.9625	0.583	0.05	ns	0.0283	13
full-length 16S rRNA gene	Silva	Order	ASV richness	y = 0.1062x + 193.6143	0.007	0.05	Significant	0.4948	13
full-length 16S rRNA gene	Silva	Order	Shannon	y = 0.0004x + 3.9666	0.132	0.05	ns	0.1939	13
full-length 16S rRNA gene	Silva	Order	Simpson	y = 0x + 0.9356	0.701	0.05	ns	0.0139	13
full-length 16S rRNA gene	GTDB	Order	ASV richness	y = 0.1337x + 201.4051	0.005	0.05	Significant	0.5338	13
full-length 16S rRNA gene	GTDB	Order	Shannon	y = 0.0005x + 4.0127	0.126	0.05	ns	0.1993	13
full-length 16S rRNA gene	GTDB	Order	Simpson	y = 0x + 0.9377	0.722	0.05	ns	0.0119	13
V4V5 16S rRNA gene	GTDB	Genus	ASV richness	y = 0.0051x + 35.8347	0.378	0.05	ns	0.1984	13
V4V5 16S rRNA gene	GTDB	Genus	Shannon	y = 0.0003x + 2.1967	0.024	0.05	Significant	0.0243	13
V4V5 16S rRNA gene	GTDB	Genus	Simpson	y = 0x + 0.7997	0.085	0.05	ns	0.0002	13
V4V5 16S rRNA gene	Silva	Genus	ASV richness	y = 0.0056x + 11.1931	0.046	0.05	Significant	0.1467	13

V4V5 16S rRNA gene	Silva	Genus	Shannon	y = 0.0006x + 1.1444	0.020	0.05	Significant	0.0911	13
V4V5 16S rRNA gene	Silva	Genus	Simpson	y = 0.0001x + 0.5798	0.062	0.05	ns	0.0283	13
full-length 16S rRNA gene	Silva	Genus	ASV richness	y = 0.0132x + 131.449	0.370	0.05	ns	0.4948	13
full-length 16S rRNA gene	Silva	Genus	Shannon	y = 0.0002x + 3.523	0.441	0.05	ns	0.1939	13
full-length 16S rRNA gene	Silva	Genus	Simpson	y = 0x + 0.9099	0.727	0.05	ns	0.0139	13
full-length 16S rRNA gene	GTDB	Genus	ASV richness	y = 0.1226x + 166.1957	0.004	0.05	Significant	0.5338	13
full-length 16S rRNA gene	GTDB	Genus	Shannon	y = 0.0005x + 3.7813	0.150	0.05	ns	0.1993	13
full-length 16S rRNA gene	GTDB	Genus	Simpson	y = 0x + 0.9169	0.745	0.05	ns	0.0119	13
V4V5 16S rRNA gene	GTDB	Species	ASV richness	y = 0.0029x + 20.3744	0.382	0.05	ns	0.0702	13
V4V5 16S rRNA gene	GTDB	Species	Shannon	y = 0.0002x + 1.9082	0.203	0.05	ns	0.1425	13
V4V5 16S rRNA gene	GTDB	Species	Simpson	y = 0x + 0.7649	0.319	0.05	ns	0.0902	13
V4V5 16S rRNA gene	Silva	Species	ASV richness	y = 0.0013x + 4.4372	0.144	0.05	ns	0.1833	13
V4V5 16S rRNA gene	Silva	Species	Shannon	y = 0.0002x + 0.7955	0.176	0.05	ns	0.1600	13
V4V5 16S rRNA gene	Silva	Species	Simpson	y = 0x + 0.4923	0.737	0.05	ns	0.0107	13
full-length 16S rRNA gene	Silva	Species	ASV richness	y = 0.0023x + 2.974	0.005	0.05	Significant	0.5611	12
full-length 16S rRNA gene	Silva	Species	Shannon	y = 0.0004x + 0.8303	0.032	0.05	Significant	0.3812	12
full-length 16S rRNA gene	Silva	Species	Simpson	y = 0.0001x + 0.4886	0.159	0.05	ns	0.1878	12
full-length 16S rRNA gene	GTDB	Species	ASV richness	y = 0.1073x + 148.7017	0.006	0.05	Significant	0.5085	13
full-length 16S rRNA gene	GTDB	Species	Shannon	y = 0.0004x + 3.7656	0.175	0.05	ns	0.1602	13
full-length 16S rRNA gene	GTDB	Species	Simpson	y = 0x + 0.9149	0.888	0.05	ns	0.0019	13

Rarefied

Sequencing approach	Database	Taxonomic level	Score	Linear regression	p-value	alpha	Significance	r.squared	n
V4V5 16S rRNA gene	GTDB	Phylum	ASV richness	y = 0.1145x + 289.6469	0.112	0.05	ns	0.3123	13
V4V5 16S rRNA gene	GTDB	Phylum	Shannon	y = 0.0003x + 3.7499	0.600	0.05	ns	0.0244	13
V4V5 16S rRNA gene	GTDB	Phylum	Simpson	y = 0x + 0.8813	0.964	0.05	ns	0.0003	13
V4V5 16S rRNA gene	Silva	Phylum	ASV richness	y = 0.1169x + 316.5625	0.129	0.05	ns	0.2931	13
V4V5 16S rRNA gene	Silva	Phylum	Shannon	y = 0.0003x + 3.8027	0.619	0.05	ns	0.0240	13
V4V5 16S rRNA gene	Silva	Phylum	Simpson	y = 0x + 0.8832	0.956	0.05	ns	0.0002	13
full-length 16S rRNA gene	Silva	Phylum	ASV richness	y = 0.1492x + 213.296	0.004	0.05	Significant	0.5368	13
full-length 16S rRNA gene	Silva	Phylum	Shannon	y = 0.0006x + 3.2171	0.325	0.05	ns	0.0883	13
full-length 16S rRNA gene	Silva	Phylum	Simpson	y = 0x + 0.7768	0.863	0.05	ns	0.0027	13

full-length 16S rRNA gene	GTDB	Phylum	ASV richness	y = 0.1481x + 208.028	0.004	0.05	Significant	0.5346	13
full-length 16S rRNA gene	GTDB	Phylum	Shannon	y = 0.0005x + 4.0558	0.132	0.05	ns	0.1915	13
full-length 16S rRNA gene	GTDB	Phylum	Simpson	y = 0x + 0.9399	0.753	0.05	ns	0.0102	13
V4V5 16S rRNA gene	GTDB	Order	ASV richness	y = 0.0988x + 228.2612	0.057	0.05	ns	0.2919	13
V4V5 16S rRNA gene	GTDB	Order	Shannon	y = 0.0003x + 3.6876	0.616	0.05	ns	0.0237	13
V4V5 16S rRNA gene	GTDB	Order	Simpson	y = 0x + 0.8794	0.957	0.05	ns	0.0003	13
V4V5 16S rRNA gene	Silva	Order	ASV richness	y = 0.0847x + 230.6347	0.082	0.05	ns	0.2497	13
V4V5 16S rRNA gene	Silva	Order	Shannon	y = 0.0002x + 4.0949	0.325	0.05	ns	0.0881	13
V4V5 16S rRNA gene	Silva	Order	Simpson	y = 0x + 0.9621	0.580	0.05	ns	0.0287	13
full-length 16S rRNA gene	Silva	Order	ASV richness	y = 0.1062x + 192.8262	0.007	0.05	Significant	0.4988	13
full-length 16S rRNA gene	Silva	Order	Shannon	y = 0.0004x + 3.9642	0.133	0.05	ns	0.1932	13
full-length 16S rRNA gene	Silva	Order	Simpson	y = 0x + 0.9356	0.702	0.05	ns	0.0138	13
full-length 16S rRNA gene	GTDB	Order	ASV richness	y = 0.1339x + 197.3396	0.005	0.05	Significant	0.5200	13
full-length 16S rRNA gene	GTDB	Order	Shannon	y = 0.0005x + 3.9967	0.126	0.05	ns	0.2000	13
full-length 16S rRNA gene	GTDB	Order	Simpson	y = 0x + 0.9372	0.710	0.05	ns	0.0131	13
V4V5 16S rRNA gene	GTDB	Genus	ASV richness	y = 0.0045x + 31.6404	0.322	0.05	ns	0.0892	13
V4V5 16S rRNA gene	GTDB	Genus	Shannon	y = 0.0003x + 2.177	0.017	0.05	Significant	0.4161	13
V4V5 16S rRNA gene	GTDB	Genus	Simpson	y = 0.0001x + 0.7955	0.066	0.05	ns	0.2745	13
V4V5 16S rRNA gene	Silva	Genus	ASV richness	y = 0.0054x + 8.1211	0.013	0.05	Significant	0.4448	13
V4V5 16S rRNA gene	Silva	Genus	Shannon	y = 0.0006x + 1.1064	0.014	0.05	Significant	0.4384	13
V4V5 16S rRNA gene	Silva	Genus	Simpson	y = 0.0001x + 0.5762	0.063	0.05	ns	0.2795	13
full-length 16S rRNA gene	Silva	Genus	ASV richness	y = 0.0136x + 128.7295	0.378	0.05	ns	0.0713	13
full-length 16S rRNA gene	Silva	Genus	Shannon	y = 0.0002x + 3.522	0.437	0.05	ns	0.0558	13
full-length 16S rRNA gene	Silva	Genus	Simpson	y = 0x + 0.9107	0.734	0.05	ns	0.0109	13
full-length 16S rRNA gene	GTDB	Genus	ASV richness	y = 0.1216x + 163.521	0.005	0.05	Significant	0.5276	13
full-length 16S rRNA gene	GTDB	Genus	Shannon	y = 0.0005x + 3.7686	0.143	0.05	ns	0.1845	13

full-length 16S rRNA gene	GTDB	Genus	Simpson	y = 0x + 0.9165	0.724	0.05	ns	0.0118	13
V4V5 16S rRNA gene	GTDB	Species	ASV richness	y = 0.0038x + 23.0114	0.328	0.05	ns	0.1624	13
V4V5 16S rRNA gene	GTDB	Species	Shannon	y = 0.0003x + 1.921	0.147	0.05	ns	0.1642	13
V4V5 16S rRNA gene	GTDB	Species	Simpson	y = 0x + 0.766	0.286	0.05	ns	0.0919	13
V4V5 16S rRNA gene	Silva	Species	ASV richness	y = 0.0013x + 5.8899	0.379	0.05	ns	0.1440	13
V4V5 16S rRNA gene	Silva	Species	Shannon	y = 0.0002x + 0.8338	0.257	0.05	ns	0.2858	13
V4V5 16S rRNA gene	Silva	Species	Simpson	y = 0x + 0.4984	0.671	0.05	ns	0.1170	13
full-length 16S rRNA gene	Silva	Species	ASV richness	y = 0.0023x + 3.175	0.016	0.05	Significant	0.4215	13
full-length 16S rRNA gene	Silva	Species	Shannon	y = 0.0005x + 0.8829	0.032	0.05	Significant	0.2768	13
full-length 16S rRNA gene	Silva	Species	Simpson	y = 0.0001x + 0.5941	0.227	0.05	ns	0.1470	13
full-length 16S rRNA gene	GTDB	Species	ASV richness	y = 0.1039x + 141.407	0.004	0.05	Significant	0.5368	13
full-length 16S rRNA gene	GTDB	Species	Shannon	y = 0.0004x + 3.7085	0.151	0.05	ns	0.1749	13
full-length 16S rRNA gene	GTDB	Species	Simpson	y = 0x + 0.9126	0.841	0.05	ns	0.0025	13

ns – not significant.
End of Supplementary Table 11.15.

Supplementary Table 11.16. Species with binomial name with more than 1 ASV, identified with GTDB and full-length 16S rRNA gene sequencing.

Binomial name	GTDB reference	ASVs	16S gene copy numer	Genome information (accessed 25/10/2023)
Pelagibacter ubique	RS_GCF_000012345_1	32	1	https://gtdb.ecogenomic.org/genome?gid=GCF_000012345.1
Haliea alexandrii	RS_GCF_003719295_1	12	1	https://gtdb.ecogenomic.org/genome?gid=GCF_003719295.1
Varunaivibrio sulfuroxidans	RS_GCF_004341725_1	11	1	https://gtdb.ecogenomic.org/genome?gid=GCF_004341725.1
Woeseia oceani	RS_GCF_001677435_1	8	1	https://gtdb.ecogenomic.org/genome?gid=GCF_001677435.1
Vicingus serpentipes	RS_GCF_007993035_1	7	2	https://gtdb.ecogenomic.org/genome?gid=GCF_007993035.1
Luteitalea pratensis	RS_GCF_001618865_1	6	1	https://gtdb.ecogenomic.org/genome?gid=GCF_001618865.1
Fuerstia marisgermanicae	RS_GCF_001983935_1	5	2	https://gtdb.ecogenomic.org/genome?gid=GCF_001983935.1
Thioglobus_B perditus	GB_GCA_002909145_1	5	1	https://gtdb.ecogenomic.org/genome?gid=GCA_002909145.1
Chryseobacterium sediminis	RS_GCF_008386505_1	3	2	https://gtdb.ecogenomic.org/genome?gid=GCF_008386505.1
Ilumatobacter fluminis	RS_GCF_004364865_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_004364865.1
Leucothrix pacifica	RS_GCF_003172875_1	3	2	https://gtdb.ecogenomic.org/genome?gid=GCF_003172875.1
Octadecabacter ascidiaceicola	RS_GCF_900185015_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_900185015.1
Pelagibius litoralis	RS_GCF_011683915_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_011683915.1
Polaromonas glacialis	RS_GCF_000709345_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_000709345.1
Psychroserpens damuponensis	RS_GCF_000826655_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_000826655.1
Reichenbachiella versicolor	RS_GCF_003171675_1	3	2	https://gtdb.ecogenomic.org/genome?gid=GCF_003171675.1
Rubritalea profundi	RS_GCF_002954445_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_002954445.1
Teredinibacter waterburyi	RS_GCF_001922985_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_001922985.1
Ulvibacter antarcticus	RS_GCF_003688405_1	3	1	https://gtdb.ecogenomic.org/genome?gid=GCF_003688405.1
Amylibacter cionae	GB_GCA_014643735_1	2	1	https://gtdb.ecogenomic.org/genome?gid=GCA_014643735.1
Colwellia psychrerythraea_A	RS_GCF_000012325_1	2	9	https://gtdb.ecogenomic.org/genome?gid=GCF_000012325.1
Coxiella burnetii	RS_GCF_005280755_1	2	1	https://gtdb.ecogenomic.org/genome?gid=GCF_005280755.1
Fluviicola taffensis	RS_GCF_000194605_1	2	2	https://gtdb.ecogenomic.org/genome?gid=GCF_000194605.1
Limimaricola cinnabarinus	RS_GCF_000466965_1	2	1	https://gtdb.ecogenomic.org/genome?gid=GCF_000466965.1
Loktanella atrilutea	RS_GCF_900128995_1	2	1	https://gtdb.ecogenomic.org/genome?gid=GCF_900128995.1

Thiobacillus thioparus	RS_GCF_000373385_1	2	1	https://gtdb.ecogenomic.org/genome?gid=GCF_000373385.1
Yoonia sediminilitoris	RS_GCF_003058085_1	2	1	https://gtdb.ecogenomic.org/genome?gid=GCF_003058085.1

End of Supplementary Table 11.16.

Supplementary Table 11.17. Details of PERMANOVA tests. Alpha to ascertain significant was always 0.05.

Sequencing approach	Taxonomic Level	Response variable	Permutations	DF	F statistic	R squared	*p* value	Significance
full-length 16S rRNA gene	Phylum	Database	999	35	0.69	0.02	0.801	ns
full-length 16S rRNA gene	Order	Database	999	25	0.845	0.03	0.66	ns
full-length 16S rRNA gene	Genus	Database	999	25	0.43	0.02	0.995	ns
full-length 16S rRNA gene	Species	Database	999	32	5.28	0.15	0.001	Significant
V4V5 16S rRNA gene	Phylum	Database	999	35	-0.17	0.00	1.000	ns
V4V5 16S rRNA gene	Order	Database	999	25	-0.11	0.00	1.000	ns
V4V5 16S rRNA gene	Genus	Database	999	25	4.73	0.16	0.001	Significant
V4V5 16S rRNA gene	Species	Database	999	35	3.78	0.10	0.004	Significant

ns – not significant.

11.3. Supplementary Tables of Chapter 5

Supplementary Table 11.18. Summary of alpha diversity values for the ASVs, OTUs and mOTUs datasets, for the total and rare communities, with and without rarefaction, for the rarity definitions derived from the MultiCoLA algorithm and from the 0.1%, per sample, definition.

Sample	Total community			Rare Biosphere (MultiCoLA)			Rare Biosphere (0.1%, per sample)		
	Total reads	Total ASVs	Shannon equitability	Total reads	Total ASVs	Shannon equitability	Total reads	Total ASVs	Shannon equitability
ASVs dataset									
Not rarefied									
NB_5	65985	391	0.66	26211	381	0.88	5999	270	0.95
NB_50	90297	466	0.69	42000	456	0.87	9327	325	0.94
NB_250	72055	416	0.39	17876	413	0.85	7243	360	0.93
TR_5	54595	495	0.74	35725	489	0.84	6168	343	0.95
TR_50	66220	509	0.74	32708	495	0.88	8330	365	0.95
TR_250	75684	775	0.74	43329	763	0.85	13258	640	0.95
YP_5	176670	457	0.67	48666	430	0.87	15819	350	0.92
YP_20	165800	570	0.67	50951	543	0.86	16619	451	0.93
YP_250	32662	174	0.69	15691	168	0.80	1291	100	0.95
Rarefied to sample with lowest number of total reads									
NB_5	32629	378	0.66	15852	375	0.84	5543	312	0.93
NB_50	32629	454	0.70	17795	446	0.85	7825	398	0.92
NB_250	32629	395	0.40	7998	398	0.84	4723	372	0.92
TR_5	32629	481	0.74	22069	476	0.83	5673	377	0.94
TR_50	32629	496	0.74	22050	492	0.83	6904	413	0.94
TR_250	32629	747	0.74	22934	739	0.81	8787	675	0.93
YP_5	32629	423	0.68	19456	411	0.77	8395	393	0.88

YP_20	32629	519	0.67	17806	521	0.80	9008	487	0.88
YP_250	32629	173	0.69	14632	166	0.81	1291	100	0.95
OTUs dataset									
Not rarefied									
NB_5	62862	189	0.47	9088	178	0.73	1191	145	0.86
NB_50	81051	222	0.49	10964	207	0.72	1835	178	0.86
NB_250	64798	198	0.41	9054	191	0.75	1419	157	0.85
TR_5	52679	324	0.55	11173	308	0.76	1819	259	0.88
TR_50	65901	267	0.51	12550	254	0.77	1958	204	0.87
TR_250	72115	250	0.55	12008	233	0.76	1844	190	0.88
YP_5	213069	273	0.48	13472	249	0.76	5423	230	0.82
YP_20	182211	294	0.51	13454	267	0.78	5863	246	0.84
YP_250	77033	178	0.40	4511	164	0.67	1204	150	0.84
Rarefied to sample with lowest number of total reads									
NB_5	52679	177	0.47	7549	166	0.74	997	133	0.87
NB_50	52679	203	0.50	7184	188	0.74	1285	160	0.87
NB_250	52679	190	0.41	7334	183	0.76	1229	150	0.86
TR_5	52679	324	0.55	11173	308	0.76	1819	259	0.88
TR_50	52679	256	0.51	9926	243	0.78	1576	193	0.88
TR_250	52679	233	0.55	10374	219	0.76	1393	174	0.89
YP_5	52679	205	0.51	5796	190	0.76	1357	162	0.86
YP_20	52679	232	0.53	10095	220	0.74	1583	182	0.88
YP_250	52679	158	0.41	4554	147	0.66	880	131	0.84
mOTUs dataset									
No rarefaction									
NB_5	1458	135	0.67	696	131	0.84	59	59	1.00
NB_50	1525	140	0.67	733	136	0.84	65	65	1.00
NB_250	1380	126	0.65	633	122	0.81	60	60	1.00

TR_5	1564	167	0.73	912	163	0.85	67	67	1.00
TR_50	1385	168	0.72	854	165	0.85	71	71	1.00
TR_250	1216	146	0.75	697	142	0.87	58	58	1.00
YP_5	2301	160	0.62	807	153	0.81	119	96	0.99
YP_20	2518	184	0.65	1282	180	0.80	128	102	0.99
YP_250	3083	153	0.66	784	142	0.82	127	96	0.98
Rarefied to sample with lowest number of total reads									
NB_5	1216	125	0.68	584	117	0.85	57	57	1
NB_50	1216	126	0.68	585	123	0.85	63	63	1
NB_250	1216	119	0.66	555	116	0.81	56	56	1
TR_5	1216	151	0.73	696	147	0.86	56	56	1
TR_50	1216	155	0.72	676	155	0.87	64	64	1
TR_250	1216	146	0.75	697	142	0.87	58	58	1
YP_5	1216	124	0.65	624	113	0.79	65	65	1
YP_20	1216	138	0.68	737	131	0.81	71	71	1
YP_250	1216	101	0.72	579	96	0.79	48	48	1

End of Supplementary Table 11.18.

11.4. Supplementary Tables of Chapter 6

Supplementary Table 11.19. Summary statistics of sequencing results from V4V5 16S rRNA gene amplicon (ASV and OTU) and shotgun metagenomic (mOTU) sequencing of the N-ICE 2015 samples. Detailed information on the number of reads and ASVs after each key bioinformatic processing step.

Phylogenetic unit	High quality reads (mean ± sd)	High quality reads after abundance table pre-processing (mean ± sd)	Rarefied reads (mean ± sd)	Initial ASVs (mean ± sd)	ASVs after pre-processing (mean ± sd)	ASVs after rarefaction (mean ± sd)	Number of initial samples	Number of samples after rarefaction
ASV	54 763 ± 29 893	46 422 ± 31 179	10 000 ± 0	312 ± 96	225 ± 72	212 ± 63	9	9
OTU	96 874 ± 58 259	96 874 ± 58 259	10 000 ± 0	247 ± 50	247 ± 50	135 ± 34	9	9
mOTU	2 051 ± 806	2 051 ± 806	1 000 ± 0	201 ± 36	106 ± 18	96 ± 14	9	9

Supplementary Table 11.20. Summary statistics of sequencing results from V4V5 and full-length 16S rRNA gene amplicon sequencing of the MOSJ 2019 samples. Detailed information on the number of reads and ASVs after each key bioinformatic processing step.

Marker gene	High quality reads (mean ± sd)	High quality reads after pre-processing (mean ± sd)	Rarefied reads (mean ± sd)	Initial ASVs (mean ± sd)	ASVs after rarefaction (mean ± sd)	Number of initial samples	Number of samples after quality filtering
V4V5 16S rRNA	55 066 ± 32 475	41 979 ± 31 629	10 000 ± 0	313 ± 155	205 ± 83	18	6
full-length 16S rRNA	13 177 ± 8 448	7 596 ± 4 640	5000 ± 0	229 ± 149	193 ± 73	18	6

Supplementary Table 11.21. Sampling metadata of the MOSJ 2016-2020 samples that were used for analysis.

Sample ID	Station	Depth (m)	Year	Latitude	Longitude	Pelagic layer	Filtration volume (mL)	Rarefaction
M16_HG-IV_B	HG-IV	50	2016	79.1	4.2	Epipelagic	3 400	Kept
M16_HG-IV_M	HG-IV	43	2016	79.1	4.2	Epipelagic	3 000	Kept
M16_HG-IV_S	HG-IV	5	2016	79.1	4.2	Epipelagic	2 550	Removed
M16_KB0_B	KB0	320	2016	79.0	11.1	Mesopelagic	1 600	Kept
M16_KB0_S	KB0	5	2016	79.0	11.1	Epipelagic	1 300	Kept
M16_KB3_B	KB3	332	2016	79.0	12.0	Mesopelagic	1 400	Kept
M16_KB3_M	KB3	50	2016	79.0	12.0	Epipelagic	1 600	Kept
M16_KB3_S	KB3	5	2016	79.0	12.0	Epipelagic	1 600	Kept
M16_KB6_B	KB6	50	2016	78.9	12.4	Epipelagic	1 530	Kept
M16_KB6_M	KB6	25	2016	78.9	12.4	Epipelagic	1 670	Kept
M16_KB6_S	KB6	2	2016	78.9	12.4	Epipelagic	1 000	Kept
M16_V12_B	V12	200	2016	79.0	9.5	Epipelagic	2 100	Kept
M17_HG-IV_1000m	HG-IV	1000	2017	79.1	4.2	Mesopelagic	4 550	Kept
M17_HG-IV_200m	HG-IV	200	2017	79.1	4.2	Epipelagic	2 750	Kept
M17_HG-IV_25m	HG-IV	25	2017	79.1	4.2	Epipelagic	2 500	Kept
M17_HG-IV_50m	HG-IV	50	2017	79.1	4.2	Epipelagic	2 500	Kept
M17_KB0_B	KB0	300	2017	79.0	11.1	Mesopelagic	2 500	Kept
M17_KB0_M	KB0	50	2017	79.0	11.1	Epipelagic	2 200	Kept
M17_KB3_B	KB3	300	2017	79.0	12.0	Mesopelagic	2 000	Kept
M17_KB3_M	KB3	50	2017	79.0	12.0	Epipelagic	2 325	Kept
M17_KB3_S	KB3	10	2017	79.0	12.0	Epipelagic	1 500	Kept
M17_KB6_B	KB6	46	2017	78.9	12.4	Epipelagic	1 400	Kept
M17_KB6_M	KB6	25	2017	78.9	12.4	Epipelagic	1 450	Kept
M17_KB6_S	KB6	1	2017	78.9	12.4	Epipelagic	1 200	Kept
M17_V12_25m	V12	25	2017	79.0	9.5	Epipelagic	2 366	Kept
M17_V12_B	V12	200	2017	79.0	9.5	Epipelagic	2 300	Kept
M17_V12_M	V12	50	2017	79.0	9.5	Epipelagic	3 400	Kept

M17_V6_1000M	V6	1000	2017	78.9	7.8	Mesopelagic	2 150	Kept
M17_V6_B	V6	200	2017	78.9	7.8	Epipelagic	4 600	Kept
M17_V6_M	V6	50	2017	78.9	7.8	Epipelagic	2 800	Kept
m18_11	KB7	67	2018	79.0	12.4	Epipelagic	3 410	Kept
m18_12	KB7	67	2018	79.0	12.4	Epipelagic	3 480	Kept
m18_13	KB7	67	2018	79.0	12.4	Epipelagic	4 226	Kept
m18_14	KB6	50	2018	78.9	12.4	Epipelagic	1 966	Kept
m18_15	KB6	50	2018	78.9	12.4	Epipelagic	1 977	Kept
m18_16	KB6	50	2018	78.9	12.4	Epipelagic	2 270	Kept
m18_17	KB6	50	2018	78.9	12.4	Epipelagic	2 332	Kept
m18_18	KB5	83	2018	78.9	12.4	Epipelagic	1 380	Kept
m18_19	KB5	83	2018	78.9	12.4	Epipelagic	1 000	Kept
m18_20	KB5	83	2018	78.9	12.4	Epipelagic	2 920	Kept
m18_21	KB3	345	2018	79.0	12.0	Mesopelagic	2 250	Kept
m18_22	KB3	345	2018	79.0	12.0	Mesopelagic	4 250	Kept
m18_23	KB2	308	2018	79.0	11.7	Mesopelagic	2 870	Kept
m18_24	KB2	308	2018	79.0	11.7	Mesopelagic	4 774	Kept
m18_25	KB1	344	2018	79.0	11.4	Mesopelagic	3 590	Kept
m18_26	KB1	344	2018	79.0	11.4	Mesopelagic	4 930	Kept
m18_27	KB1	344	2018	79.0	11.4	Mesopelagic	5 000	Kept
m18_28	KB0	327	2018	79.0	11.1	Mesopelagic	3 540	Kept
m18_29	KB0	327	2018	79.0	11.1	Mesopelagic	4 210	Kept
m18_30	KB0	327	2018	79.0	11.1	Mesopelagic	5 000	Kept
m18_31	V12	222	2018	79.0	9.5	Mesopelagic	1 000	Kept
m18_32	V12	222	2018	79.0	9.5	Mesopelagic	1 100	Kept
m18_33	V12	222	2018	79.0	9.5	Mesopelagic	4 750	Kept
m18_34	V6	1130	2018	78.9	7.8	Bathypelagic	1 500	Kept
m18_35	V6	1130	2018	78.9	7.8	Bathypelagic	1 500	Kept
m18_36	V6	1130	2018	78.9	7.8	Bathypelagic	5 000	Kept
m18_37	V10	293	2018	78.9	8.5	Mesopelagic	1 400	Kept
m18_38	V10	293	2018	78.9	8.5	Mesopelagic	1 000	Kept
m18_39	V10	293	2018	78.9	8.5	Mesopelagic	5 000	Kept

m18_41	V12	222	2018	79.0	9.5	Mesopelagic	1 000	Kept
M19_1	KB3	5	2019	79.0	12.0	Epipelagic	1 800	Kept
M19_100	HG-I	5	2019	79.1	6.1	Epipelagic	4 000	Removed
M19_101	HG-I	25	2019	79.1	6.1	Epipelagic	4 000	Kept
M19_103	HG-I	1220	2019	79.1	6.1	Bathypelagic	7 000	Kept
M19_105	HG-IV	5	2019	79.1	4.2	Epipelagic	4 500	Kept
M19_106	HG-IV	18	2019	79.1	4.2	Epipelagic	4 200	Kept
M19_108	HG-IV	2300	2019	79.1	4.2	Bathypelagic	5 000	Kept
M19_11	KB7	1	2019	79.0	12.4	Epipelagic	2 000	Kept
M19_14	KB7	5	2019	79.0	12.4	Epipelagic	2 000	Kept
M19_17	KB7	61	2019	79.0	12.4	Epipelagic	4 000	Kept
M19_20	KB6	1	2019	78.9	12.4	Epipelagic	1 750	Kept
M19_24	KB6	7	2019	78.9	12.4	Epipelagic	1 000	Kept
M19_28	KB6	50	2019	78.9	12.4	Epipelagic	1 000	Kept
M19_33	KB5	1	2019	78.9	12.4	Epipelagic	500	Kept
M19_37	KB5	12	2019	78.9	12.4	Epipelagic	500	Kept
M19_40	KB5	65	2019	78.9	12.4	Epipelagic	1 000	Kept
M19_44	KB2	1	2019	79.0	11.7	Epipelagic	4 500	Kept
M19_46	KB2	13	2019	79.0	11.7	Epipelagic	3 500	Kept
M19_48	KB2	290	2019	79.0	11.7	Mesopelagic	2 200	Kept
M19_5	KB3	15	2019	79.0	12.0	Epipelagic	3 000	Kept
M19_50	KB1	1	2019	79.0	11.4	Epipelagic	3 800	Kept
M19_52	KB1	14	2019	79.0	11.4	Epipelagic	4 000	Kept
M19_54	KB1	340	2019	79.0	11.4	Mesopelagic	4 800	Kept
M19_56	KB0	5	2019	79.0	11.1	Epipelagic	1 500	Kept
M19_60	KB0	14	2019	79.0	11.1	Epipelagic	3 500	Kept
M19_62	KB0	320	2019	79.0	11.1	Mesopelagic	3 500	Kept
M19_64	V12	5	2019	79.0	9.5	Epipelagic	3 000	Kept
M19_66	V12	28	2019	79.0	9.5	Epipelagic	3 200	Kept
M19_70	V12	215	2019	79.0	9.5	Mesopelagic	3 100	Kept
M19_72	V10	5	2019	78.9	8.5	Epipelagic	1 800	Kept
M19_76	V10	25	2019	78.9	8.5	Epipelagic	2 000	Kept

M19_79	V10	260	2019	78.9	8.5	Mesopelagic	5 000	Kept
M19_81	V6	5	2019	78.9	7.8	Epipelagic	4 000	Kept
M19_84	V6	25	2019	78.9	7.8	Epipelagic	5 000	Kept
M19_88	V6	1100	2019	78.9	7.8	Bathypelagic	4 000	Kept
M19_9	KB3	300	2019	79.0	12.0	Mesopelagic	5 000	Kept
M19_91	KH	5	2019	79.1	7.0	Epipelagic	4 500	Kept
M19_93	KH	25	2019	79.1	7.0	Epipelagic	4 000	Kept
M19_96	KH	1300	2019	79.1	7.0	Bathypelagic	5 000	Kept
M20-14	KB6	50	2020	78.93	12.39	Epipelagic	1 340	Kept
M20-18	KB6	5	2020	78.93	12.39	Epipelagic	1 440	Kept
M20-19	KB3	339	2020	78.95	11.96	Mesopelagic	200	Kept
M20-21	KB3	25	2020	78.95	11.96	Epipelagic	1 300	Kept
M20-24	KB3	5	2020	78.95	11.96	Epipelagic	3 770	Kept
M20-25	KB0	322	2020	79.05	11.14	Mesopelagic	4 400	Kept
M20-28	KB0	17	2020	79.05	11.14	Epipelagic	2 600	Kept
M20-30	KB0	5	2020	79.05	11.14	Epipelagic	2 900	Kept
M20-31	V12	200	2020	79.0	9.5	Epipelagic	5 000	Kept
M20-33	V12	20	2020	79.0	9.5	Epipelagic	3 500	Kept
M20-35	V12	5	2020	79.0	9.5	Epipelagic	2 000	Kept
M20-37	V10	290	2020	78.9	8.5	Mesopelagic	4 100	Kept
M20-39	V10	25	2020	78.9	8.5	Epipelagic	4 400	Kept
M20-41	V10	5	2020	78.9	8.5	Epipelagic	3 400	Kept
M20-43	KH	1324	2020	79.1	7.0	Bathypelagic	NA	Kept
M20-45	KH	15	2020	79.1	7.0	Epipelagic	NA	Kept
M20-47	KH	5	2020	79.1	7.0	Epipelagic	1 100	Kept
M20-49	HG-IV	2443	2020	79.1	4.2	Bathypelagic	4 700	Kept
M20-51	HG-IV	25	2020	79.1	4.2	Epipelagic	3 750	Kept
M20-53	HG-IV	5	2020	79.1	4.2	Epipelagic	3 800	Kept

End of Supplementary Table 11.21.

Supplementary Table 11.22. Sequencing statistics from V4V5 16S rRNA gene amplicon sequencing of the MOSJ samples from 2016 to 2020.

Year	High quality reads (mean ± sd)	High quality reads after pre-processing (mean ± sd)	Rarefied reads (mean ± sd)	Initial ASVs (mean ± sd)	ASVs after rarefaction (mean ± sd)	Number of initial samples	Number of samples after quality filtering
2016	16 529 ± 5632	16 462 ± 5635	10 000	157 ± 33	157 ± 33	12	11
2017	42 179 ± 11 082	39 904 ± 10 340	10 000	359 ± 163	330 ± 148	18	18
2018	53 728 ± 14 146	52 568 ± 14 213	10 000	317 ± 102	276 ± 89	30	30
2019	59 530 ± 36 841	58 638 ± 36 924	10 000	299 ± 163	274 ± 128	39	38
2020	44 295 ± 23 911	43 563 ± 24 221	10 000	279 ± 104	263 ± 93	20	20
Total			NA			119	117

Supplementary Table 11.23. List of approaches used to define the rare biosphere in literature, from 2006 to 2024 (July). The survey focused on threshold-based approaches vs other alternatives and only accounted for definitions that were explicitly explained in the methods, or in reference to another study. We remain neutral regarding the methodological options and conclusions of the studies mentioned in this table, including those labeled as “undefined or unclear”.

Year	**Type of definition**	**Number of thresholds**	**Rarity threshold**	**Application to**	**Reference**
2006	Undefined or unclear	Na	Na	Na	(Sogin *et al*., 2006)
2008	Absolute abundance threshold	1	several	Na	(Huse *et al*., 2008)
2008	Absolute abundance threshold	1	several	Na	(Dethlefsen *et al*., 2008)
2008	Absolute abundance threshold	2	Na	Na	(Elshahed *et al*., 2008)
2008	Absolute abundance threshold	1	Na	Na	(Neufeld, Li and Mohn, 2008)
2009	Undefined or unclear	Na	Na	Na	(Webster *et al*., 2009)
2009	Relative Abundance threshold	2	0.0100%	Sample	(Galand *et al*., 2009)
2010	Undefined or unclear	Na	Na	Na	(Kunin *et al*., 2010)
2009	Absolute abundance threshold	1	Na	Sample	(Hamp, Jones and Fodor, 2009)
2011	Absolute abundance threshold	1	Na	Sample	(Gaidos, Rusch and Ilardo, 2011)

2010	Other	5	0.0040%	Sample	(Youssef, Couger and Elshahed, 2010)
2010	Relative Abundance threshold	1	0.1000%	Whole dataset	(Pester *et al*., 2010)
2010	Absolute abundance threshold	1	Na	Na	(Huse *et al*., 2010)
2010	Undefined or unclear	Na	Na	Na	(Sunagawa, Woodley and Medina, 2010)
2012	Relative Abundance threshold	2	0.1000%	Na	(Sjöstedt *et al*., 2012)
2012	Undefined or unclear	Na	Na	Na	(Youssef, Steidley and Elshahed, 2012)
2012	Undefined or unclear	Na	Na	Na	(Gobet *et al*., 2012)
2011	Undefined or unclear	Na	Na	Na	(Kim *et al*., 2011)
2013	Undefined or unclear	Na	Na	Na	(Kamono *et al*., 2013)
2013	Relative Abundance threshold	1	1.0000%	Whole dataset	(Bachy *et al*., 2013)
2012	Undefined or unclear	Na	Na	Na	(Taylor *et al*., 2013)
2012	Undefined or unclear	Na	Na	Na	(Lee *et al*., 2012)
2012	Undefined or unclear	Na	Na	Na	(Lynch, Bartram and Neufeld, 2012)
2012	Undefined or unclear	Na	Na	Na	(Shade *et al*., 2012)
2012	Absolute abundance threshold	Na	Na	Sample	(Bowen *et al*., 2012)
2012	Undefined or unclear	Na	Na	Na	(Gonzalez *et al*., 2012)
2012	Undefined or unclear	Na	Na	Na	(Youssef, Ashlock-Savage and Elshahed, 2012)
2012	Undefined or unclear	Na	Na	Na	(Pepe-Ranney *et al*., 2012)
2014	Relative Abundance threshold	Na	Na	Na	(Erwin *et al*., 2014)
2013	Undefined or unclear	Na	Na	Na	(Rodriguez-Lanetty *et al*., 2013)
2013	Undefined or unclear	Na	Na	Na	(Bowman *et al*., 2013)
2014	Relative Abundance threshold	1	0.1000%	Na	(Stoeck *et al*., 2014)
2013	Relative Abundance threshold	2	0.2000%	Sample	(Hugoni *et al*., 2013)
2014	Relative Abundance threshold	1	1.0000%	Whole dataset	(Auld *et al*., 2013)
2013	Undefined or unclear	Na	Na	Na	(Denonfoux *et al*., 2013)
2014	Relative Abundance threshold	Na	0.0100%	Whole dataset	(Reveillaud *et al*., 2014)
2013	Absolute abundance threshold	2	0.0060%	Whole dataset	(Logares *et al*., 2013)
2014	Relative Abundance threshold	4	several	Other	(Logares *et al*., 2014)
2014	Undefined or unclear	Na	Na	Na	(Quigley *et al*., 2014)

2014	Relative Abundance threshold	1	0.1000%	Sample	(Quero and Luna, 2014)
2014	Undefined or unclear	Na	Na	Na	(Tytgat *et al*., 2014)
2014	Undefined or unclear	Na	Na	Na	(Shade and Gilbert, 2015)
2014	Relative Abundance threshold	Na	1.0000%	Na	(Alonso-Sáez *et al*., 2014)
2014	Relative Abundance threshold	Na	0.1000%	Whole dataset	(Sauret *et al*., 2014)
2015	Undefined or unclear	Na	Na	Na	(Gagic *et al*., 2015)
2015	Relative Abundance threshold	2	0.1000%	Sample	(Anderson, Sogin and Baross, 2015)
2015	Undefined or unclear	Na	Na	Na	(Połka *et al*., 2015)
2015	Undefined or unclear	Na	Na	Na	(Dini-Andreote *et al*., 2015)
2015	Relative Abundance threshold	1	0.1000%	Whole dataset	(Debroas, Hugoni and Domaizon, 2015)
2015	Relative Abundance threshold	4	several	Other	(Liu *et al*., 2015)
2015	Undefined or unclear	Na	Na	Na	(Alonso-Sáez, Díaz-Pérez and Morán, 2015)
2015	Absolute abundance threshold	1	Na	Sample	(İnceoğlu *et al*., 2015)
2015	Relative Abundance threshold	1	0.0001%	Na	(Delmont *et al*., 2015)
2015	Undefined or unclear	Na	Na	Na	(Mylnikov *et al*., 2015)
2015	Relative Abundance threshold	3	0.1000%	Na	(Baltar *et al*., 2015)
2015	Undefined or unclear	Na	Na	Na	(Gittel *et al*., 2015)
2016	Relative Abundance threshold	1	0.1000%	Sample	(Fuentes *et al*., 2016)
2016	Relative Abundance threshold	1	0.0150%	Sample	(Hamasaki *et al*., 2016)
2015	Undefined or unclear	Na	Na	Na	(Miranda-Sánchez, Rivera and Vinuesa, 2016)
2015	Absolute abundance threshold	1	Na	Sample	(Coveley, Elshahed and Youssef, 2015)
2015	Undefined or unclear	Na	Na	Na	(Salazar *et al*., 2016)
2015	Undefined or unclear	Na	Na	Na	(Hol *et al*., 2015)
2016	Relative Abundance threshold	1	0.1000%	Sample	(Niño-García, Ruiz-González and del Giorgio, 2016)
2016	Undefined or unclear	Na	Na	Na	(Kang *et al*., 2016)
2016	Relative Abundance threshold	3	0.1000%	Sample	(Molina *et al*., 2016)
2016	Relative Abundance threshold	3	0.1000%	Sample	(Schiaffino *et al*., 2016)
2016	Relative Abundance threshold	1	1.0000%	Sample	(Fernandez-Gonzalez, Huber and Vallino, 2016)
2016	Relative Abundance threshold	1	5.0000%	Sample	(Grattepanche *et al*., 2016)
2016	Relative Abundance threshold	Na	0.1000%	Sample	(Crespo *et al*., 2016)

2016	Relative Abundance threshold	Na	0.4800%	Sample	(Mora-Ruiz *et al*., 2016)
2016	Undefined or unclear	Na	Na	Na	(López-Pérez *et al*., 2016)
2016	Undefined or unclear	Na	Na	Na	(Locey and Lennon, 2016)
2016	Relative Abundance threshold	Na	0.1000%	Sample	(Hausmann *et al*., 2016)
2016	Undefined or unclear	Na	Na	Na	(De Mandal *et al*., 2016)
2018	Relative Abundance threshold	2	0.0010%	Whole dataset	(W. Zhang *et al*., 2018)
2017	Absolute abundance threshold	1	0.0200%	Whole dataset	(Klein Jan *et al*., 2017)
2018	Undefined or unclear	Na	Na	Na	(Ziegler *et al*., 2018)
2017	Relative Abundance threshold	Na	1.0000%	Whole dataset	(Yang *et al*., 2017)
2018	Relative Abundance threshold	1	0.1000%	Sample	(Peter *et al*., 2018)
2018	Relative Abundance threshold	Na	1.0000%	Sample	(Hardge *et al*., 2018)
2017	Relative Abundance threshold	1	0.1000%	Sample	(Idris *et al*., 2017)
2017	Relative Abundance threshold	1	0.1000%	Sample	(Richa *et al*., 2017)
2017	Undefined or unclear	Na	Na	Na	(Royo-Llonch *et al*., 2017)
2018	Undefined or unclear	Na	Na	Na	(Chiriac *et al*., 2018)
2018	Undefined or unclear	Na	Na	Na	(Zhu *et al*., 2018)
2017	Relative Abundance threshold	1	0.1000%	Sample	(Dawson *et al*., 2017)
2017	Relative Abundance threshold	1	0.1500%	Whole dataset	(Ganesan *et al*., 2017)
2017	Relative Abundance threshold	1	0.1000%	Sample	(Niño-García, Ruiz-González and del Giorgio, 2017)
2017	Relative Abundance threshold	1	0.1000%	Sample	(Wang *et al*., 2017)
2018	Relative Abundance threshold	1	0.1000%	Sample	(De Anda *et al*., 2018)
2019	Undefined or unclear	Na	Na	Na	(Villegas-Plazas *et al*., 2019)
2018	Relative Abundance threshold	1	0.1000%	Sample	(Wei *et al*., 2018)
2018	Relative Abundance threshold	3	0.0100%	Sample	(Y. Zhang *et al*., 2018)
2019	Relative Abundance threshold	1	0.0100%	Whole dataset	(Fernández-Gómez *et al*., 2019)
2018	Undefined or unclear	Na	Na	Na	(Zehavi, Probst and Mizrahi, 2018)
2018	Relative Abundance threshold	1	0.1000%	Na	(Lavergne *et al*., 2018)
2018	Relative Abundance threshold	1	0.0100%	Na	(Kroeger *et al*., 2018)
2018	Relative Abundance threshold	1	0.1000%	Whole dataset	(Vigneron *et al*., 2018)
2018	Undefined or unclear	Na	Na	Na	(Yin *et al*., 2018)

2018	Relative Abundance threshold	1	0.0100%	Sample	(Pessi *et al*., 2018)
2018	Relative Abundance threshold	1	0.0100%	Sample	(Kaminsky and Morales, 2018)
2018	Relative Abundance threshold	1	0.0100%	Sample	(Meisner *et al*., 2018)
2018	Relative Abundance threshold	Na	Na	Whole dataset	(Kalenitchenko *et al*., 2018)
2019	Undefined or unclear	Na	Na	Na	(Haro-Moreno, Rodriguez-Valera and López-Pérez, 2019)
2019	Relative Abundance threshold	1	0.1000%	Sample	(Gokul *et al*., 2019)
2020	Undefined or unclear	Na	Na	Na	(Doud *et al*., 2020)
2019	Relative Abundance threshold	1	0.1000%	Whole dataset	(Yamamoto *et al*., 2019)
2019	Undefined or unclear	Na	Na	Na	(Hur and Park, 2019)
2019	Undefined or unclear	Na	Na	Na	(Ramírez *et al*., 2019)
2019	Relative Abundance threshold	Na	0.2000%	Na	(Schreiber *et al*., 2019)
2019	Relative Abundance threshold	3	0.1000%	Sample	(Shaw *et al*., 2019)
2019	Relative Abundance threshold	Na	0.1000%	Sample	(Kirstein *et al*., 2019)
2019	Undefined or unclear	Na	Na	Na	(Cerro-Gálvez *et al*., 2019)
2019	Relative Abundance threshold	1	0.0100%	Sample	(Liu, Xue and Yang, 2019)
2019	Relative Abundance threshold	1	0.1000%	Sample	(Hausmann *et al*., 2019)
2020	Undefined or unclear	Na	Na	Na	(Varela *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Zhang *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Wang *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Dam *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Karthikeyan *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Mars Brisbin, Conover and Mitarai, 2020)
2020	Undefined or unclear	Na	Na	Na	(Bay *et al*., 2020)
2020	Relative Abundance threshold	Na	0.1000%	Na	(Richter-Heitmann *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Vázquez-Rosas-Landa *et al*., 2020)
2020	Relative Abundance threshold	3	0.0100%	Whole dataset	(Du *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Belheouane *et al*., 2020)
2020	Relative Abundance threshold	1	0.1000%	Sample	(Pinseel *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Hou *et al*., 2020)
2020	Relative Abundance threshold	1	1.0000%	Whole dataset	(Choi and Park, 2020)

2020	Undefined or unclear	Na	Na	Na	(Pan *et al*., 2020)
2020	Relative Abundance threshold	2	0.1000%	Sample	(Ji *et al*., 2020)
2020	Relative Abundance threshold	6	0.0100%	Mixed	(Liang *et al*., 2020)
2020	Undefined or unclear	Na	Na	Na	(Rocca *et al*., 2020)
2020	Relative Abundance threshold	1	0.0100%	Sample	(Hu *et al*., 2020)
2021	Undefined or unclear	Na	Na	Na	(Omae *et al*., 2021)
2021	Undefined or unclear	Na	Na	Na	(Rabelo-Ruiz *et al*., 2021)
2021	Relative Abundance threshold	3	several	Sample	(Catão C. P. *et al*., 2021)
2021	Relative Abundance threshold	2	0.0100%	Na	(Y. Zhang *et al*., 2021)
2021	Relative Abundance threshold	2	0.01%	Sample	(N. Zhang *et al*., 2021)
2021	Relative Abundance threshold	2	0.1000%	Sample	(X. Zhang *et al*., 2021)
2021	Relative Abundance threshold	3	0.0500%	Whole dataset	(Shu *et al*., 2021)
2021	Relative Abundance threshold	1	0.1000%	Sample	(Scales *et al*., 2021)
2021	Relative Abundance threshold	1	1.0000%	Unclear	(Esquivel-Hernández *et al*., 2021)
2021	Undefined or unclear	Na	Na	Na	(Umbach, Stegelmeier and Neufeld, 2021)
2021	Other	Na	Na	Na	(Nicolas *et al*., 2021)
2021	Relative Abundance threshold	1	0.0100%	Na	(Gschwend *et al*., 2021)
2021	Undefined or unclear	Na	Na	Na	(Papadakis *et al*., 2021)
2021	Relative Abundance threshold	1	0.001	Sample	(Yadav *et al*., 2021)
2021	Relative Abundance threshold	1	0.1000%	Sample	(Campos *et al*., 2021)
2022	Other	Na	Na	Na	(Pust and Tümmler, 2022)
2022	Undefined or unclear	Na	Na	Na	(Schultz *et al*., 2022)
2022	Relative Abundance threshold	3	0.1000%	Sample	(Wu, Hua and Gu, 2022)
2022	Other	Na	Na	Na	(Espinosa-Asuar *et al*., 2022)
2022	Relative Abundance threshold	1	0.1000%	Sample	(Silva *et al*., 2022)
2022	Relative Abundance threshold	1	0.1000%	Sample	(Pascoal *et al*., 2022a)
2022	Relative Abundance threshold	1	0.1000%	Sample	(Labrie *et al*., 2022)
2022	Relative Abundance threshold	1	1.0000%	Whole dataset	(Honeyman *et al*., 2022)
2022	Relative Abundance threshold	2	0.1000%	Sample	(Jia *et al*., 2022)
2022	Relative Abundance threshold	1	0.1000%	Sample	(Crevecoeur, Prairie and del Giorgio, 2022)

2022	Other	several	several	Sample	(Jia, Dini-Andreote and Salles, 2022)
2022	Relative Abundance threshold	2	0.1000%	Whole dataset	(Siles, Hendrickson and Terry, 2022)
2022	Relative Abundance threshold	2	0.1000%	Sample	(Qin *et al*., 2022)
2023	Relative Abundance threshold	2	several	Sample	(Gattoni *et al*., 2023)
2023	Relative Abundance threshold	1	0.1000%	Sample	(Tang *et al*., 2023)
2023	Relative Abundance threshold	several	several	Whole dataset	(Luan *et al*., 2023)
2023	Relative Abundance threshold	1	0.3000%	Sample	(Araujo *et al*., 2023)
2023	Relative Abundance threshold	2	0.0100%	Sample	(Li *et al*., 2023)
2023	Relative Abundance threshold	2	0.0500%	Whole dataset	(Cui *et al*., 2023)
2023	Relative Abundance threshold	several	0.0100%	Whole dataset	(Zhu *et al*., 2023)
2023	Undefined or unclear	Na	Na	Na	(Jia *et al*., 2023)
2023	Other	Na	Na	Na	(Ramond *et al*., 2023)
2023	Undefined or unclear	Na	Na	Na	(Riley *et al*., 2023)
2023	Relative Abundance threshold	2	0.0100%	Sample	(Sanz-Sáez *et al*., 2023)
2024	Relative Abundance threshold	2	0.1000%	Sample	(Sebastián *et al*., 2024)
2024	Relative Abundance threshold	1	1.0000%	Sample	(Skoog and Bosak, 2024)
2024	Relative Abundance threshold	several	several	Whole dataset	(Bontemps, Moënne-Loccoz and Hugoni, 2024)
2024	Undefined or unclear	Na	Na	Na	(Mallmin, Traulsen and De Monte, 2024)
2024	Relative Abundance threshold	2	0.1000%	Sample	(Yang *et al*., 2024)

End of Supplementary Table 11.23. Na – not applicable/available.

11.5. Supplementary Tables of Chapter 8

Supplementary Table 11.24. Statistics on the number of reads before and after major processing steps, and taxonomic groups obtained (MOSJ 2016-2020).

Year	High quality reads (mean ± sd)	High quality reads after pre-processing (mean ± sd)	taxonomic groups (mean ± sd)	Number of initial samples	Number of samples after quality filtering
2016	16 529 ± 5632	15 363 ± 4 498	87 ± 18	12	11
2017	42 179 ± 11 082	39 782 ± 11 055	120 ± 34	18	18
2018	53 728 ± 14 146	48 815 ± 14 731	118 ± 30	30	30
2019	59 530 ± 36 841	53 104 ± 36 347	103 ± 37	38	37
2020	44 295 ± 23 911	41 532 ± 23 657	109 ± 34	20	20
Total		NA		119	116

Supplementary Table 11.25. Summary of two-way ANOVA results and Levene test, comparing the number of taxa across year and pelagic layers. Significance was attested for using an alpha of 0.05.

		two-way ANOVA		Levene test		
Response variable	Effect	p-value	Significance	p-value	Significance	Homogeneity of variance
Number of taxa	year	7.70E-04	significant			
Number of taxa	pelagic layer	1.49E-20	significant	0.07	ns	confirmed
Number of taxa	year and pelagic layer	5.20E-01	ns			

Supplementary Table 11.26. *Post-hoc* test for two-way ANOVA, using pairwise t-test, comparing the number of taxa between each combination of year and of pelagic layers. The alpha used to attest significance was 0.05. *p* values were adjusted by Bonferroni correction.

Response variable	Effect 1	Effect 2	n1	n2	p-value	p.signif	p.adj	p.adj.signif
Number of taxa	2016	2017	11	18	0.010	Significant	0.099	ns
Number of taxa	2016	2018	11	30	0.008	Significant	0.085	ns
Number of taxa	2017	2018	18	30	0.840	ns	1.000	ns
Number of taxa	2016	2019	11	37	0.146	ns	1.000	ns
Number of taxa	2017	2019	18	37	0.083	ns	0.832	ns
Number of taxa	2018	2019	30	37	0.075	ns	0.746	ns
Number of taxa	2016	2020	11	20	0.072	ns	0.722	ns
Number of taxa	2017	2020	18	20	0.321	ns	1.000	ns
Number of taxa	2018	2020	30	20	0.363	ns	1.000	ns
Number of taxa	2019	2020	37	20	0.521	ns	1.000	ns
Number of taxa	Epipelagic	Mesopelagic	84	24	4.03E-08	Significant	1.21E-07	Significant
Number of taxa	Epipelagic	Bathypelagic	84	8	1.30E-18	Significant	3.91E-18	Significant
Number of taxa	Mesopelagic	Bathypelagic	24	8	7.32E-09	Significant	2.20E-08	Significant

End of Supplementary Table 8.3.

Supplementary Table 11.27. Shapiro test results for number of taxa within each combination of year and pelagic layers. The alpha used to attest significance was 0.05.

Response variable	Explaining variable 1	Explaining variable 2	Statistic	p	Significance	Normal distribution
Number of taxa	2016	Epipelagic	0.928	0.500	ns	Yes
Number of taxa	2016	Mesopelagic	0.977	0.712	ns	Yes
Number of taxa	2017	Epipelagic	0.872	0.083	ns	Yes
Number of taxa	2017	Mesopelagic	0.818	0.113	ns	Yes
Number of taxa	2018	Epipelagic	0.934	0.134	ns	Yes
Number of taxa	2018	Mesopelagic	0.957	0.796	ns	Yes
Number of taxa	2019	Epipelagic	0.967	0.505	ns	Yes
Number of taxa	2019	Mesopelagic	0.916	0.475	ns	Yes
Number of taxa	2020	Epipelagic	0.938	0.388	ns	Yes
Number of taxa	2020	Mesopelagic	0.905	0.457	ns	Yes
Number of taxa	Interaction between year and pelagic layer		0.979	0.072	ns	Yes

Supplementary Table 11.28. Linear least squares model results for number of taxa as a function of years, accounting for pelagic layers. Significance was attested for using an alpha of 0.05.

Response variable	Explanatory variable	Subset	Regression equation	p-value	Significance	R2	Note
Number of taxa	Years	NA	y = 107.98 + 0.52x	0.84	ns	-0.008	model does not fit data
Number of abundant taxa	Years	Epipelagic layer	y = 3.37 + 0.45x	0.09	ns	0.021	
	Years	Mesopelagic layer	y = 4.93 + 0.92x	0.13	ns	0.058	
	Years	Bathypelagic layer	y = 8.05 - 0.92x	0.1	ns	0.28	
Number of undetermined taxa	Years	Epipelagic layer	y = 12.99 + 1.02x	0.08	ns	0.023	model does not fit data
	Years	Mesopelagic layer	y = 18.06 + 2.11x	0.19	ns	0.033	model does not fit data
	Years	Bathypelagic layer	y = 30.49 - 1.76x	0.44	ns	-0.044	model does not fit data
Number of rare taxa	Years	Epipelagic layer	y = 82.12 - 2.35x	0.26	ns	0.003	model does not fit data
	Years	Mesopelagic layer	y = 100.17 - 0.77x	0.77	ns	-0.041	model does not fit data

	Years	Bathypelagic layer	y = 141.01 + 5.52x	0.43	ns	-0.043	model does not fit data

End of Supplementary Table 11.28.

Supplementary Table 11.29. Poison model results for number of taxa as a function of years, accounting for pelagic layers. Significance was attested for using an alpha of 0.05.

Response variable	Explanatory variable	Subset	Regression equation	p-value	Significance	AIC
Number of taxa	Years	NA	log(y) = 4.68 + 0.00x	0.51	ns	1912.3
Number of abundant taxa	Years	Epipelagic layer	log(y) = 1.23 + 0.11x	0.019	Significant	432
	Years	Mesopelagic layer	log(y) = 1.62 + 0.14x	0.03	Significant	138.6
	Years	Bathypelagic layer	log(y) = 2.13 - 0.16x	0.22	ns	35.1
Number of undetermined taxa	Years	Epipelagic layer	log(y) = 2.57 + 0.07x	0.0048	Significant	609.3
	Years	Mesopelagic layer	log(y) = 2.91 + 0.09x	0.0058	Significant	224.4
	Years	Bathypelagic layer	log(y) = 3.42 - 0.07x	0.277	ns	55.2
Number of rare taxa	Years	Epipelagic layer	log(y) = 4.41 - 0.03x	0.0036	Significant	1042.7
	Years	Mesopelagic layer	log(y) = 4.61 - 0.01x	0.631	ns	218.4
	Years	Bathypelagic layer	log(y) = 4.95 + 0.04x	0.164	ns	76.1

Supplementary Table 11.30. Pearson and Kendal correlations between the number of taxa and years variable, accounting for pelagic layers. Significance was attested for using an alpha of 0.05.

			Pearson correlation			Kendall correlation		
Response variable	Explanatory variable	Subset	correlation	p-value	significance	tau	p-value	significance
Number of taxa	Years	NA	0.018	0.840	ns	-0.024	0.732	ns
Number of abundant taxa	Years	Epipelagic layer	0.181	0.099	ns	0.129	0.136	ns

	Years	Mesopelagic layer	0.315	0.134	ns	0.241	0.136	ns
	Years	Bathypelagic layer	-0.621	0.100	ns	-0.591	0.069	ns
Number of undetermined taxa	Years	Epipelagic layer	0.186	0.089	ns	0.146	0.080	ns
	Years	Mesopelagic layer	0.274	0.195	ns	0.161	0.307	ns
	Years	Bathypelagic layer	-0.323	0.435	ns	-0.276	0.364	ns
Number of rare taxa	Years	Epipelagic layer	-0.123	0.264	ns	-0.121	0.141	ns
	Years	Mesopelagic layer	-0.062	0.772	ns	-0.069	0.664	ns
	Years	Bathypelagic layer	0.325	0.432	ns	0.118	0.698	ns

End of Supplementary Table 11.30.

Supplementary Table 11.31. Relative abundance of phylum and class level, across the whole MOSJ 2016-2020 dataset (n = 117).

Phylum	Relative abundance (%)
Bacteroidota	41.78458
Proteobacteria	41.44114
Verrucomicrobiota	4.90280
Crenarchaeota	2.69667
Thermoplasmatota	1.51847
Actinobacteriota	1.47179
Cyanobacteria	1.43058
Planctomycetota	1.40446
Marinimicrobia	0.75986
SAR324	0.69162
Nitrospinota	0.62136
Bdellovibrionota	0.27096
Chloroflexi	0.17951
Desulfobacterota	0.13088
Gemmatimonadota	0.12709
Firmicutes	0.11537
Acidobacteriota	0.09806
NB1	0.09357
Nanoarchaeota	0.07819
Dadabacteria	0.04334
PAUC34f	0.03236
Myxococcota	0.01895
Caldatribacteriota	0.01491
Patescibacteria	0.01294
Campylobacterota	0.01242
Hydrogenedentes	0.01170
AncK6	0.00875
Halobacterota	0.00463
Fibrobacterota	0.00387
Margulisbacteria	0.00379
Nitrospirota	0.00297
RCP2	0.00290
Schekmanbacteria	0.00260
Spirochaetota	0.00200
Dependentiae	0.00119
Calditrichota	0.00101
WPS	0.00058
WS1	0.00054
Euryarchaeota	0.00043
GAL15	0.00025
Asgardarchaeota	0.00023
Deinococcota	0.00016
MBNT15	0.00016
Synergistota	0.00016
Fusobacteriota	0.00012
Deferrisomatota	0.00006
Sva0485	0.00004

Class level	Relative abundance (%)
Bacteroidia	41.77585
Alphaproteobacteria	23.35552
Gammaproteobacteria	17.96000
Verrucomicrobiae	4.84278
Nitrososphaeria	2.69667
Thermoplasmata	1.51847
Cyanobacteriia	1.43058
Acidimicrobiia	0.92250
SAR406	0.75986
clade	0.69162
Planctomycetes	0.64223
Nitrospinia	0.61961
Actinobacteria	0.53500
OM190	0.43203
Phycisphaerae	0.31595
Bdellovibrionia	0.26330
Dehalococcoidia	0.15465
BD2	0.10904
Desulfobulbia	0.10389
Bacilli	0.07881
Nanoarchaeia	0.07819
Vicinamibacteria	0.07353
Dadabacteriia	0.04334
Lentisphaeria	0.04282
Clostridia	0.02465
Gemmatimonadetes	0.01806
Desulfuromonadia	0.01802
Kiritimatiellae	0.01629
JS1	0.01491
WCHB1	0.01378
JG30	0.01337
Campylobacteria	0.01242
Subgroup	0.01224
Hydrogenedentia	0.01170
Microgenomatia	0.01153
Myxococcia	0.01036
Rhodothermia	0.00842
Polyangia	0.00795
Oligoflexia	0.00766
BD7	0.00700
Thermoanaerobacteria	0.00694
Aminicenantia	0.00649
Anaerolineae	0.00575
Halobacteria	0.00463
Fibrobacteria	0.00387
Syntrophia	0.00362
028H05	0.00315
Pla3	0.00293
Thermoanaerobaculia	0.00260
Moorellia	0.00200
Spirochaetia	0.00190
Nitrospiria	0.00187
Acidobacteriae	0.00179

P9X2b3D02	0.00175
Dethiobacteria	0.00130
P2	0.00124
Babeliae	0.00119
Calditrichia	0.00101
SHA	0.00099
N9D0	0.00097
Blastocatellia	0.00076
Gracilibacteria	0.00074
Omnitrophia	0.00072
Holophagae	0.00064
bacteriap25	0.00064
Desulfobacteria	0.00060
Desulfitobacteriia	0.00056
SGST604	0.00056
Thermoleophilia	0.00051
Negativicutes	0.00049
TK10	0.00047
Pla4	0.00045
TK17	0.00043
Thermococci	0.00043
Desulfarculia	0.00041
Chloroflexia	0.00033
Parcubacteria	0.00029
KD4	0.00023
Lokiarchaeia	0.00023
Chlamydiae	0.00019
Berkelbacteria	0.00017
Deinococci	0.00016
Ignavibacteria	0.00016
Kryptonia	0.00016
Synergistia	0.00016
Fusobacteriia	0.00012
Saccharimonadia	0.00012
ABY1	0.00010
Leptospirae	0.00010
Defferrisomatia	0.00006
NA	0.27846

End of Supplementary Table 11.31.

Supplementary Table 11.32. PERMANOVA test results, comparing community composition between years and pelagic layers. 999 permutations were used. Significance was attested for using an alpha of 0.05.

Variable	Degrees of freedom	Sum of squares	R2	F statistic	p-value	Significance
Year	1	1.1799	0.0469	7.009	0.001	Significant
PelagicLayer	2	5.0429	0.2003	14.978	0.001	Significant
year:PelagicLayer	2	0.4361	0.0173	1.295	0.183	ns
Residual	110	18.5175	0.7355			
Total	115	25.1764	1			

ns - not significant

Supplementary Table 11.33. Linear least squared model results for number of taxa as a function of temperature and depth, accounting for pelagic layers, water masses and region. Tests were carried out for the whole community (all), as well as the subset of abundant, undetermined and rare taxa, except for tests accounting for pelagic layers, water mass and region. Significance was attested for using an alpha of 0.05.

Response variable	Explanatory variable	Subset	Regression equation	p-value	Significance	R2	Notes
Number of taxa (all)	Temperature	NA	y = 152.211 - 10.187x	2.00E-16	Significant	0.559	
Number of abundant taxa	Temperature		y = 5.746 - 0.170x	1.59E-01	ns	0.008	
Number of undetermined taxa	Temperature		y = 23.438 - 1.433x	9.00E-07	Significant	0.189	
Number of rare taxa	Temperature		y = 123.027 - 8.583x	2.20E-16	Significant	0.515	
Number of taxa (all)	Temperature	Epipelagic	y = 131.77 - 6.75x	4.75E-07	Significant	0.267	
Number of taxa (all)	Temperature	Mesopelagic	y = 135.74 - 3.55x	0.363	ns	0.363	
Number of taxa (all)	Temperature	Bathypelagic	y = 161.9 - 38.45x	0.075	ns	0.341	
Number of taxa (all)	Temperature	AW	y = 149.99 - 9.44x	1.80E-03	Significant	0.281	
Number of taxa (all)	Temperature	IW	y = 151.33 - 10.06x	2.45E-05	Significant	0.828	
Number of taxa (all)	Temperature	SW	y = 153.84 - 10.80x	6.67E-05	Significant	0.431	

Number of taxa (all)	Temperature	TAW	y = 151.09 - 12.96x	0.128	ns	0.091	
Number of taxa (all)	Temperature	WCW	y = 224.03 + 44.62x	0.217	ns	0.143	
Number of taxa (all)	Temperature	Kongsfjorden	y = 150.36 - 10.25x	2.00E-16	Significant	0.65	
Number of taxa (all)	Temperature	Fram Strait	y = 155.93 - 10.31x	3.94E-08	Significant	0.49	
Number of taxa (all)	Depth	Epipelagic	y = 91.433 + 0.26x	0.062	ns	0.03	model does not fit
Number of taxa (all)	Depth	Mesopelagic	y = 127.1 + 0.003x	0.969	ns	-0.045	model does not fit
Number of taxa (all)	Depth	Bathypelagic	y = 178.5 + 0.006x	0.528	ns	-0.1047	model does not fit

End of Supplementary Table 11.33.

Supplementary Table 11.34. Poisson model results for number of taxa as a function of temperature and depth, accounting for pelagic layers, water masses and region. Tests were carried out for the whole community (all), as well as the subset of abundant, undetermined and rare taxa, except for tests accounting for pelagic layers, water mass and region. Significance was attested for using an alpha of 0.05.

Response variable	Explanatory variable	Subset	Regression equation	p-value	Significance	AIC
Number of taxa (all)	Temperature	NA	log(y) = 5.05 - 0.09x	2.00E-16	Significant	1230.3
Number of abundant taxa	Temperature		log(y) = 7.75 - 0.03x	0.043	Significant	608.29
Number of undetermined taxa	Temperature		log(y) = 3.17 - 0.08x	2.00E-16	Significant	872.6
Number of rare taxa	Temperature		log(y) = 4.84 - 0.09x	2.00E-16	Significant	1220.4
Number of taxa (all)	Temperature	Epipelagic	log(y) = 4.91 - 0.07x	2.00E-16	Significant	883.9
Number of taxa (all)	Temperature	Mesopelagic	log(y) = 4.91 - 0.03x	0.104	ns	233.6
Number of taxa (all)	Temperature	Bathypelagic	log(y) = 5.09 - 0.21x	0.064	ns	61.1
Number of taxa (all)	Temperature	AW	log(y) = 5.11 - 0.09x	2.00E-16	Significant	477.3
Number of taxa (all)	Temperature	IW	log(y) = 5.08 - 0.09x	5.19E-13	Significant	88.4
Number of taxa (all)	Temperature	SW	log(y) = 5.13 - 0.11x	2.00E-16	Significant	284.9
Number of taxa (all)	Temperature	TAW	log(y) = 5.03 - 0.10x	9.00E-03	Significant	156.6
Number of taxa (all)	Temperature	WCW	log(y) = 5.42 - 0.23x	0.416	ns	55.32
Number of taxa (all)	Temperature	Kongsfjorden	log(y) = 5.05 - 0.09x	2.00E-16	Significant	610.6

Number of taxa (all)	Temperature	Fram Strait	log(y) = 5.05 - 0.08x	2.00E-16	Significant	619.4
Number of taxa (all)	Depth	Epipelagic	log(y) = 4.52 + 0.003x	3.37E-06	Significant	1026
Number of taxa (all)	Depth	Mesopelagic	log(y) = 4.84 + 0.00005x	0.944	ns	236.3
Number of taxa (all)	Depth	Bathypelagic	log(y) = 5.19 + 0.00003	0.518	ns	68.2

End of Supplementary Table 11.34.

Supplementary Table 11.35. Pearson and Kendal correlations between the number of taxa and temperature and depth, accounting for pelagic layers. Tests were carried out for the whole community (all), as well as the subset of abundant, undetermined and rare taxa, except for tests accounting for pelagic layers, water mass and region. Significance was attested for using an alpha of 0.05.

			Pearson correlation			**Kendall correlation**		
Response variable	**Explanatory variable**	**Subset**	**Pearson correlation**	**p-value**	**Significance**	**Tau**	**p-value**	**Significance**
Number of taxa (all)	Temperature	NA	-0.751	2.20E-16	Significant	-0.551	2.20E-16	Significant
Number of abundant taxa	Temperature		-0.13	0.16	ns	-0.087	0.193	ns
Number of undetermined taxa	Temperature		-0.44	9.00E-07	Significant	-0.292	6.72E-06	Significant
Number of rare taxa	Temperature		-0.72	2.20E-16	Significant	-0.539	2.20E-16	Significant
Number of taxa (all)	Temperature	Epipelagic	-0.52	4.75E-07	Significant	-0.415	4.89E-08	Significant
Number of taxa (all)	Temperature	Mesopelagic	-0.19	3.63E-01	ns	-0.169	2.52E-01	ns
Number of taxa (all)	Temperature	Bathypelagic	-0.66	0.075	ns	-0.071	0.905	ns
Number of taxa (all)	Temperature	AW	-0.546	0.000182	Significant	-0.422	9.04E-05	Significant
Number of taxa (all)	Temperature	IW	-0.918	2.45E-05	Significant	-0.687	1.00E-03	Significant
Number of taxa (all)	Temperature	SW	-0.671	6.67E-05	Significant	-0.463	4.70E-04	Significant
Number of taxa (all)	Temperature	TAW	-0.384	0.1278	ns	-0.273	0.127	ns
Number of taxa (all)	Temperature	WCW	0.534	0.2166	ns	0.238	0.562	ns
Number of taxa (all)	Temperature	Kongsfjorden	-0.807	2.20E-16	Significant	-0.59	1.43E-12	Significant
Number of taxa (all)	Temperature	Fram Strait	-0.713	3.94E-08	Significant	-0.5	1.27E-06	Significant

Number of taxa (all)	Depth	Epipelagic	0.204	0.062	ns	0.061	0.436	ns
Number of taxa (all)	Depth	Mesopelagic	0.008	0.969	ns	0.019	0.899	ns
Number of taxa (all)	Depth	Bathypelagic	0.231	0.583	ns	-0.036	0.901	ns

End of Supplementary Table 11.35.

Supplementary Table 11.36. Linear least squared model results for number of taxa as a function of silicate concentration (µM). Tests were carried out for the whole community (all), as well as the subset of abundant, undetermined and rare taxa. Significance was attested for using an alpha of 0.05.

Response variable	Explanatory variable	Regression equation	p-value	significance	R2
Number of taxa (all)	Silicates	y = 77.29 + 8.58x	2.20E-16	Significant	0.557
Number of abundant taxa	Silicates	y = 4.17 + 0.239x	0.0256	Significant	0.0375
Number of undetermined taxa	Silicates	y = 12.384 + 1.351x	1.87E-07	Significant	0.221
Number of rare taxa	Silicates	y = 60.729 + 6.989x	2.20E-16	Significant	0.486

Supplementary Table 11.37. Poisson model results for number of taxa as a function of silicate concentration (µM). Tests were carried out for the whole community (all), as well as the subset of abundant, undetermined and rare taxa. Significance was attested for using an alpha of 0.05.

Response variable	Explanatory variable	Regression equation	p-value	Significance	AIC
Number of taxa (all)	Silicates	log(y) = 4.42 + 0.06x	2.00E-16	Significant	1199.7
Number of abundant taxa	Silicates	log(y) = 1.46 + 0.04x	0.0016	Significant	569.7
Number of undetermined taxa	Silicates	log(y) = 2.59 + 0.06x	2.00E-16	Significant	822
Number of rare taxa	Silicates	log(y) = 4.19 + 0.06x	2.00E-16	Significant	1204

Supplementary Table 11.38. Pearson and Kendall correlation between the number of taxa and silicate concentration (μM). Tests were carried out for the whole community (all), as well as the subset of abundant, undetermined and rare taxa. Significance was attested for using an alpha of 0.05.

		Pearson correlation			Kendall correlation		
Response variable	**Explanatory variable**	**Pearson correlation**	**p-value**	**significance**	**Tau**	**p-value**	**Significance**
Number of taxa (all)	Silicates	0.749	2.20E-16	Significant	0.471	1.09E-12	Significant
Number of abundant taxa	Silicates	-0.133	0.158	ns	-0.087	0.194	ns
Number of undetermined taxa	Silicates	-0.44	9.00E-07	Significant	-0.292	6.75E-06	Significant
Number of rare taxa	Silicates	-0.72	2.20E-16	Significant	-0.539	2.20E-16	Significant

Supplementary Table 11.39. Top 100 association rules, ordered by descending Lift. Additional statistics are provided (Support, Confidence, Coverage, and Count). Note that: "NO2" = NO_2^-; "NO3" = NO_3^-; "Chl" = Chl *a*; "PO4" = PO_4^-; "NH4" = NH_4^+.

Antecedent	**Consequent**	**Support**	**Confidence**	**Coverage**	**Lift**	**Count**
{taxon=*Bacteria*_Proteobacteria_*Alphaproteobacteria*_SAR11 clade_Clade I_Clade Ia, NO2 = very high, Si = very high}	{Classification=Abundant}	0.11%	93.33%	0.0012	20.28	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high,Si=very high}	{Classification=Abundant}	0.10%	92.86%	0.0011	20.18	13
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,NO2=very high,Si=very high,Chl=Unknown}	{Classification=Abundant}	0.10%	92.86%	0.0011	20.18	13
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high,Si=very high,Chl=Unknown}	{Classification=Abundant}	0.10%	92.86%	0.0011	20.18	13
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,Si=very high}	{Classification=Abundant}	0.12%	88.24%	0.0013	19.17	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,Si=very high,Chl=Unknown}	{Classification=Abundant}	0.12%	88.24%	0.0013	19.17	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,NO2=very high,Chl=Unknown}	{Classification=Abundant}	0.13%	85.00%	0.0016	18.47	17
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high}	{Classification=Abundant}	0.13%	84.21%	0.0015	18.30	16

{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Abundant}	0.13%	84.21%	0.0015	18.30	16
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_, PelagicLayer=Epipelagic,NH4=Unknown,Si=low}	{Classification=Abundant}	0.12%	83.33%	0.0014	18.11	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,NO2=very high}	{Classification=Abundant}	0.17%	81.48%	0.0021	17.70	22
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PO4=very high,NO2=very high}	{Classification=Undetermined}	0.11%	93.33%	0.0012	5.79	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NO2=very high,Si=very high}	{Classification=Undetermined}	0.11%	93.33%	0.0012	5.79	14
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,NO2=very high,NH4=Unknown}	{Classification=Undetermined}	0.11%	93.33%	0.0012	5.79	14
{taxon=Bacteria_Planctomycetota_OM190_,PO4=very high,Si=very high,Temperature=very low}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,PO4=very high,NO2=very high}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,NO2=very high,Si=very high}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,PO4=very high,Si=very high}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PO4=very high,NO2=very high,Chl=Unknown}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NO2=very high,Si=very high,Chl=Unknown}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Planctomycetota_OM190_,PO4=very high,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,PO4=very high,NO2=very high,Chl=Unknown}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae, PelagicLayer=Mesopelagic,NO2=very high,Si=very high,Chl=Unknown}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae, PelagicLayer=Mesopelagic,PO4=very high,Si=very high,Chl=Unknown}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NO2=very high,Chl=Unknown}	{Classification=Undetermined}	0.14%	90.00%	0.0016	5.59	18
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,NO2=very high}	{Classification=Undetermined}	0.13%	89.47%	0.0015	5.55	17
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Undetermined}	0.13%	89.47%	0.0015	5.55	17

{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PO4=very low}	{Classification=Undetermined}	0.19%	88.89%	0.0021	5.52	24
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Epipelagic,PO4=very low}	{Classification=Undetermined}	0.19%	88.89%	0.0021	5.52	24
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PO4=very low,Si=very low}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Epipelagic,NO2=Undetected}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,NH4=Unknown,Salinity=very high}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,PO4=very high}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,Si=very high}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,PelagicLayer=Epipelagic,NO2=Undetected}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,NH4=Unknown,Si=high}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Epipelagic,PO4=very low,Si=very low}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,PO4=very high,Chl=Unknown}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,Si=very high,Chl=Unknown}	{Classification=Undetermined}	0.12%	88.24%	0.0013	5.48	15
{taxon=Bacteria_Planctomycetota_OM190_,Si=very high,Temperature=very low}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Nitrospinota_Nitrospinia_Nitrospinales_Nitrospinaceae_LS-NOB,Si=very high,Temperature=very low}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_NS9 marine group_,NH4=Unknown,Salinity=low}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PO4=very low,NH4=Unknown}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,PelagicLayer=Epipelagic,Salinity=very high}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Planctomycetota_OM190_,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14

{taxon=Bacteria_Nitrospinota_Nitrospinia_Nitrospinales_Nitrospinaceae_LS-NOB,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Epipelagic,PO4=very low,NH4=Unknown}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Epipelagic,NH4=Unknown,Si=high}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,PelagicLayer=Epipelagic,NH4=Unknown,Si=high}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,PelagicLayer=Epipelagic,Salinity=very high,WaterMass=AW}	{Classification=Undetermined}	0.11%	87.50%	0.0013	5.43	14
{taxon=Bacteria_Planctomycetota_OM190_,PO4=very high,Temperature=very low}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_Nitrosomonas,PelagicLayer=Mesopelagic,Si=very high}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Verrucomicrobiales_Rubritaleaceae_Roseibacillus,PO4=very high,NO2=very high}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Verrucomicrobiales_Rubritaleaceae_Roseibacillus,NO2=very high,NH4=Unknown}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,NO2=very high,NH4=Unknown}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NO2=very high,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,NO2=very high,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,NO2=very high,NH4=Unknown}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Planctomycetota_OM190_,PO4=very high,Chl=Unknown,Temperature=very low}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_Nitrosomonas,PelagicLayer=Mesopelagic,Si=very high,Chl=Unknown}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,NO2=very high,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13

{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,PelagicLayer=Mesopelagic,NO2=very high,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Undetermined}	0.10%	86.67%	0.0012	5.38	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_NS9 marine group_,PelagicLayer=Epipelagic,Salinity=low}	{Classification=Undetermined}	0.15%	86.36%	0.0017	5.36	19
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,NH4=Unknown,WaterMass=AW}	{Classification=Undetermined}	0.15%	86.36%	0.0017	5.36	19
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,Si=high}	{Classification=Undetermined}	0.18%	85.19%	0.0021	5.29	23
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Epipelagic,NH4=Unknown,WaterMass=AW}	{Classification=Undetermined}	0.13%	84.21%	0.0015	5.23	16
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic}	{Classification=Undetermined}	0.16%	83.33%	0.0019	5.17	20
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_Nitrosomonas,PelagicLayer=Mesopelagic,NO2=very high}	{Classification=Undetermined}	0.12%	83.33%	0.0014	5.17	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,Chl=Unknown}	{Classification=Undetermined}	0.16%	83.33%	0.0019	5.17	20
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_Nitrosomonas,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Undetermined}	0.12%	83.33%	0.0014	5.17	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,WaterMass=TAW}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,WaterMass=TAW}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Verrucomicrobiales_Rubritaleaceae_Roseibacillus,NH4=Unknown,Si=high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,NH4=Unknown,Temperature=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14

{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Mesopelagic,PO4=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Ulvibacter,NH4=Unknown,Temperature=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NH4=Unknown,Si=high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NH4=Unknown,Salinity=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,PelagicLayer=Mesopelagic,PO4=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,PelagicLayer=Mesopelagic,Si=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS5 marine group,Si=very high,Temperature=very low}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Epipelagic,NH4=Unknown,Temperature=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PelagicLayer=Mesopelagic,PO4=very high,Chl=Unknown}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Ulvibacter,PelagicLayer=Epipelagic,NH4=Unknown,Temperature=very high}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,PelagicLayer=Mesopelagic,PO4=very high,Chl=Unknown}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,PelagicLayer=Mesopelagic,Si=very high,Chl=Unknown}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS5 marine group,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Undetermined}	0.11%	82.35%	0.0013	5.11	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,Salinity=very high,WaterMass=AW}	{Classification=Undetermined}	0.14%	81.82%	0.0017	5.08	18
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,PelagicLayer=Epipelagic,Si=high}	{Classification=Undetermined}	0.14%	81.82%	0.0017	5.08	18
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,NO2=very high}	{Classification=Undetermined}	0.17%	81.48%	0.0021	5.06	22
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,NO2=very high}	{Classification=Undetermined}	0.17%	81.48%	0.0021	5.06	22

{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,NO2=very high}	{Classification=Undetermined}	0.17%	81.48%	0.0021	5.06	22
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Ulvibacter,PO4=very low,NH4=Unknown}	{Classification=Undetermined}	0.10%	81.25%	0.0013	5.04	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic,WaterMass=TAW}	{Classification=Undetermined}	0.10%	81.25%	0.0013	5.04	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,Chl=Unknown,WaterMass=TAW}	{Classification=Undetermined}	0.10%	81.25%	0.0013	5.04	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,PelagicLayer=Mesopelagic,WaterMass=TAW}	{Classification=Undetermined}	0.10%	81.25%	0.0013	5.04	13

End of Supplementary Table 11.39.

Supplementary Table 11.40. Dependent rules, with Lift above 15, ordered by Conviction. Dependent rules were identified by Chi squared test, using an alpha of 0.05. Additional metrics are provided (Support, Confidence, Coverage and Count).

Antecedent	Consequent	Support	Confidence	Coverage	Lift	Count	Conviction	Chi-test p-value
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,NO2=very high,Si=very high}	{Classification=Abundant}	0.11%	93.33%	0.0012	20.28	14	14.31	3.00E-56
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high,Si=very high}	{Classification=Abundant}	0.10%	92.86%	0.0011	20.18	13	13.36	9.59E-52
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,NO2=very high,Si=very high,Chl=Unknown}	{Classification=Abundant}	0.10%	92.86%	0.0011	20.18	13	13.36	9.59E-52
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high,Si=very high,Chl=Unknown}	{Classification=Abundant}	0.10%	92.86%	0.0011	20.18	13	13.36	9.59E-52
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,Si=very high}	{Classification=Abundant}	0.12%	88.24%	0.0013	19.17	15	8.11	1.17E-56
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade	{Classification=Abundant}	0.12%	88.24%	0.0013	19.17	15	8.11	1.17E-56

Ia,PelagicLayer=Mesopelagic,Si=very high,Chl=Unknown}								
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,NO2=very high,Chl=Unknown}	{Classification=Abundant}	0.13%	85.00%	0.0016	18.47	17	6.36	7.91E-62
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high}	{Classification=Abundant}	0.13%	84.21%	0.0015	18.30	16	6.04	2.04E-57
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Abundant}	0.13%	84.21%	0.0015	18.30	16	6.04	2.04E-57
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_,PelagicLayer=Epipelagic,NH4=Unknown,Si=low}	{Classification=Abundant}	0.12%	83.33%	0.0014	18.11	15	5.72	5.09E-53
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,NO2=very high}	{Classification=Abundant}	0.17%	81.48%	0.0021	17.70	22	5.15	6.27E-77

End of Supplementary Table 11.40.

Supplementary Table 11.41. Non-redundant and dependent rules, independent of Lift. Additional metrics are provided (Support, Confidence, Coverage, Count, Conviction, Mutual Information and Improvement). Note that: "NO2" = NO_2^-; "NO3" = NO_3^-; "Chl" = Chl *a*; "PO4" = PO_4^-; "NH4" = NH_4^+.

Antecedent	Consequent	Support	Confidence	Lift	Count	Mutual Information	Improvement
{Chl=Unknown,WaterMass=WCW}	{Classification=Rare}	8.91%	84.24%	1.063	1128	0.003	0.000
{Chl=very low,Temperature=low,Salinity=very low}	{Classification=Rare}	3.90%	87.13%	1.099	494	0.005	0.022
{NH4=Unknown,Chl=low,WaterMass=SW}	{Classification=Rare}	3.17%	88.13%	1.112	401	0.007	0.881
{NO2=high,Chl=very low,Temperature=low}	{Classification=Rare}	4.47%	87.08%	1.098	566	0.005	0.871
{PelagicLayer=Bathypelagic,WaterMass=WCW}	{Classification=Rare}	8.91%	84.24%	1.063	1128	0.003	0.000
{PelagicLayer=Epipelagic,Temperature=low}	{Classification=Rare}	13.17%	83.32%	1.051	1668	0.002	0.833
{PO4=high,NO2=high,Salinity=low}	{Classification=Rare}	2.04%	91.20%	1.150	259	0.011	0.912
{PO4=low,Temperature=low}	{Classification=Rare}	6.72%	85.19%	1.074	851	0.004	0.852

{Salinity=high,WaterMass=WCW}	{Classification=Rare}	8.91%	84.24%	1.063	1128	0.003	0.000
{Si=low,Salinity=very low}	{Classification=Rare}	5.13%	92.86%	1.171	650	0.019	0.087
{Si=low}	{Classification=Rare}	17.42%	84.20%	1.062	2207	0.004	0.842
{Temperature=low,Salinity=very low}	{Classification=Rare}	9.32%	84.95%	1.071	1180	0.004	0.850
{Temperature=very low,WaterMass=WCW}	{Classification=Rare}	8.91%	84.24%	1.063	1128	0.003	0.000
{WaterMass=WCW}	{Classification=Rare}	8.91%	84.24%	1.063	1128	0.003	0.842
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_,PelagicLayer=Mesopelagic}	{Classification=Undetermined}	0.16%	83.33%	5.172	20	0.152	0.833
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_,PelagicLayer=Epipelagic,NH4=Unknown,Si=low}	{Classification=Abundant}	0.12%	83.33%	18.106	15	0.282	0.833
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS4 marine group,PelagicLayer=Epipelagic,NO2=Undetected}	{Classification=Undetermined}	0.12%	88.24%	5.476	15	0.167	0.882
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS5 marine group,Si=very high,Temperature=very low}	{Classification=Undetermined}	0.11%	82.35%	5.111	14	0.141	0.824
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Ulvibacter,NH4=Unknown,Temperature=very high}	{Classification=Undetermined}	0.11%	82.35%	5.111	14	0.141	0.824
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_NS9 marine group_,NH4=Unknown,Salinity=low}	{Classification=Undetermined}	0.11%	87.50%	5.430	14	0.162	0.875
{taxon=Bacteria_Nitrospinota_Nitrospinia_Nitrospinales_Nitrospinaceae_LS-NOB,Si=very high,Temperature=very low}	{Classification=Undetermined}	0.11%	87.50%	5.430	14	0.162	0.875
{taxon=Bacteria_Planctomycetota_OM190_,Si=very high,Temperature=very low}	{Classification=Undetermined}	0.11%	87.50%	5.430	14	0.162	0.875
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Ascidiaceihabitans,PelagicLayer=Epipelagic,NH4=Unknown,Si=high}	{Classification=Undetermined}	0.10%	81.25%	5.043	13	0.135	0.813
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Planktomarina,PO4=very low}	{Classification=Undetermined}	0.19%	88.89%	5.517	24	0.181	0.889
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,Si=very high}	{Classification=Abundant}	0.12%	88.24%	19.171	15	0.311	0.882

{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade III_}	{Classification=Rare}	0.68%	100.00%	1.261	86	NA	1.000
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade IV_}	{Classification=Rare}	0.77%	98.98%	1.248	97	0.032	0.990
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_Nitrosomonas,PelagicLayer=Mesopelagic,Si=very high}	{Classification=Undetermined}	0.10%	86.67%	5.379	13	0.157	0.867
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_KI89A clade_}	{Classification=Rare}	0.74%	100.00%	1.261	94	NA	1.000
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,WaterMass=TAW}	{Classification=Undetermined}	0.11%	82.35%	5.111	14	0.141	0.824
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Verrucomicrobiales_Rubritaleaceae_Roseibacillus,PO4=very high,NO2=very high}	{Classification=Undetermined}	0.10%	86.67%	5.379	13	0.157	0.867

End of Supplementary Table 11.41.

Supplementary Table 11.42. Top 100 rules with Conviction above 5, Confidence above 90%, and Lift above 1. Improvement was used to select non-redundant rules. Rules were ordered by Lift. Additional metrics were provided (Support and Count). Note that: "NO2" = NO_2^-; "NO3" = NO_3^-; "Chl" = Chl *a*; "PO4" = PO_4^-; "NH4" = NH_4^+.

Antecedent	Consequent	Support	Confidence	Coverage	Lift	Count
{taxon=Archaea_Nanoarchaeota_Nanoarchaeia_Woesearchaeales_SCGC AAA011-D5_}	{Classification=Rare}	0.13%	100%	0.0013	1.26	16
{taxon=Archaea_Thermoplasmatota_Thermoplasmata_Marine Group II,PO4=very low}	{Classification=Rare}	0.13%	100%	0.0013	1.26	17
{taxon=Bacteria_Acidobacteriota_Vicinamibacteria_Vicinamibacterales_Vicinamibacteraceae_}	{Classification=Rare}	0.15%	100%	0.0015	1.26	19
{taxon=Bacteria_Actinobacteriota_Acidimicrobiia_Microtrichales_Microtrichaceae_}	{Classification=Rare}	0.14%	100%	0.0014	1.26	18
{taxon=Bacteria_Actinobacteriota_Actinobacteria_}	{Classification=Rare}	0.17%	100%	0.0017	1.26	22
{taxon=Bacteria_Actinobacteriota_Actinobacteria_Corynebacteriales_Nocardiaceae_Rhodococcus}	{Classification=Rare}	0.12%	100%	0.0012	1.26	15
{taxon=Bacteria_Actinobacteriota_Actinobacteria_Micrococcales_Microbacteriaceae_Candidatus Aquiluna}	{Classification=Rare}	0.54%	100%	0.0054	1.26	68

{taxon=Bacteria_Actinobacteriota_Actinobacteria_Micrococcales_Microbacteriaceae_ML602J-51}	{Classification=Rare}	0.31%	100%	0.0031	1.26	39
{taxon=Bacteria_Actinobacteriota_Actinobacteria_Propionibacteriales_Propionibacteriaceae_Cutibacterium}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_}	{Classification=Rare}	0.12%	100%	0.0012	1.26	15
{taxon=Bacteria_Bacteroidota_Bacteroidia_Bacteroidales_Prolixibacteraceae_}	{Classification=Rare}	0.13%	100%	0.0013	1.26	17
{taxon=Bacteria_Bacteroidota_Bacteroidia_Chitinophagales}	{Classification=Rare}	0.35%	100%	0.0035	1.26	44
{taxon=Bacteria_Bacteroidota_Bacteroidia_Chitinophagales_Saprospiraceae_Aureispira}	{Classification=Rare}	0.45%	100%	0.0045	1.26	57
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Crocinitomicaceae_Crocinitomix}	{Classification=Rare}	0.58%	100%	0.0058	1.26	74
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Flavicella}	{Classification=Rare}	0.32%	100%	0.0032	1.26	40
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Flavobacterium}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Lutibacter}	{Classification=Rare}	0.14%	100%	0.0014	1.26	18
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Maritimimonas}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS3a marine group,Salinity=high}	{Classification=Rare}	0.13%	100%	0.0013	1.26	16
{taxon=Bacteria_Bacteroidota_Bacteroidia_Sphingobacteriales_Lentimicrobiaceae_}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Bacteroidota_Bacteroidia_Sphingobacteriales_Sphingobacteriaceae_Pedobacter}	{Classification=Rare}	0.13%	100%	0.0013	1.26	17
{taxon=Bacteria_Bdellovibrionota_Bdellovibrionia_Bacteriovoracales_Bacteriovoracaceae_Peredibacter}	{Classification=Rare}	0.30%	100%	0.0030	1.26	38
{taxon=Bacteria_Bdellovibrionota_Oligoflexia_Oligoflexales_Oligoflexaceae_}	{Classification=Rare}	0.13%	100%	0.0013	1.26	16
{taxon=Bacteria_Campylobacterota_Campylobacteria_Campylobacterales_Sulfurimonadaceae_Sulfurimonas}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Chloroflexi_Anaerolineae_Anaerolineales_Anaerolineaceae_}	{Classification=Rare}	0.13%	100%	0.0013	1.26	16
{taxon=Bacteria_Desulfobacterota__}	{Classification=Rare}	0.16%	100%	0.0016	1.26	20
{taxon=Bacteria_Desulfobacterota_Desulfuromonadia_Bradymonadales}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Desulfobacterota_Desulfuromonadia_PB19}	{Classification=Rare}	0.22%	100%	0.0022	1.26	28
{taxon=Bacteria_Firmicutes_Bacilli_Staphylococcales_Staphylococcaceae_Staphylococcus}	{Classification=Rare}	0.12%	100%	0.0012	1.26	15
{taxon=Bacteria_Marinimicrobia (SAR406 clade)__,PO4=very low}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14

{taxon=Bacteria_Myxococcota_Myxococcia_Myxococcales_Myxococcaceae_P3OB-42}	{Classification=Rare}	0.15%	100%	0.0015	1.26	19
{taxon=Bacteria_Myxococcota_Polyangia_}	{Classification=Rare}	0.14%	100%	0.0014	1.26	18
{taxon=Bacteria_Planctomycetota_OM190_,WaterMass=AW}	{Classification=Rare}	0.15%	100%	0.0015	1.26	19
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycisphaerales_AKAU3564 sediment group_}	{Classification=Rare}	0.13%	100%	0.0013	1.26	16
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycisphaerales_Phycisphaeraceae_}	{Classification=Rare}	0.15%	100%	0.0015	1.26	19
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycisphaerales_Phycisphaeraceae_FS140-16B-02 marine group}	{Classification=Rare}	0.28%	100%	0.0028	1.26	36
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycisphaerales_Phycisphaeraceae_JL-ETNP-F27,PelagicLayer=Mesopelagic}	{Classification=Rare}	0.13%	100%	0.0013	1.26	17
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycisphaerales_Phycisphaeraceae_Urania-1B-19 marine sediment group}	{Classification=Rare}	0.18%	100%	0.0018	1.26	23
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_Pir4 lineage}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_Rhodopirellula,PelagicLayer=Mesopelagic}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_Rubripirellula}	{Classification=Rare}	0.16%	100%	0.0016	1.26	20
{taxon=Bacteria_Planctomycetota_Planctomycetes_Planctomycetales_Gimesiaceae_}	{Classification=Rare}	0.26%	100%	0.0026	1.26	33
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_}	{Classification=Rare}	0.52%	100%	0.0052	1.26	66
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_AT-s3-44}	{Classification=Rare}	0.21%	100%	0.0021	1.26	27
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Caulobacterales_Caulobacteraceae_Brevundimonas}	{Classification=Rare}	0.12%	100%	0.0012	1.26	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Parvibaculales_Parvibaculaceae_}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Puniceispirillales}	{Classification=Rare}	0.36%	100%	0.0036	1.26	46
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Puniceispirillales_EF100-94H03_}	{Classification=Rare}	0.12%	100%	0.0012	1.26	15
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Puniceispirillales_SAR116 clade_Candidatus Puniceispirillum}	{Classification=Rare}	0.38%	100%	0.0038	1.26	48
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_}	{Classification=Rare}	0.54%	100%	0.0054	1.26	69
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Octadecabacter}	{Classification=Rare}	0.51%	100%	0.0051	1.26	65

{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Tateyamaria}	{Classification=Rare}	0.17%	100%	0.0017	1.26	21
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Yoonia-Loktanella,Temperature=very high}	{Classification=Rare}	0.16%	100%	0.0016	1.26	20
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Magnetospiraceae_Magnetospira,Chl=very high}	{Classification=Rare}	0.13%	100%	0.0013	1.26	17
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales}	{Classification=Rare}	0.18%	100%	0.0018	1.26	23
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales_Rickettsiaceae_Candidatus Megaira}	{Classification=Rare}	0.13%	100%	0.0013	1.26	16
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales_S25-593_}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ib,WaterMass=TAW}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade II_,Chl=high}	{Classification=Rare}	0.15%	100%	0.0015	1.26	19
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade III_}	{Classification=Rare}	0.68%	100%	0.0068	1.26	86
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Thalassobaculales_Nisaeaceae_OM75 clade}	{Classification=Rare}	0.63%	100%	0.0063	1.26	80
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Arenicellales_Arenicellaceae_Arenicella}	{Classification=Rare}	0.16%	100%	0.0016	1.26	20
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Comamonadaceae_Polaromonas}	{Classification=Rare}	0.24%	100%	0.0024	1.26	30
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Comamonadaceae_Rhodoferax}	{Classification=Rare}	0.21%	100%	0.0021	1.26	26
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Comamonadaceae_RS62 marine group}	{Classification=Rare}	0.37%	100%	0.0037	1.26	47
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Hydrogenophilaceae_Thiobacillus}	{Classification=Rare}	0.22%	100%	0.0022	1.26	28
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_IS-44}	{Classification=Rare}	0.36%	100%	0.0036	1.26	45
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_Nitrosomonas,Temperature=high}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Coxiellales_Coxiellaceae_Coxiella}	{Classification=Rare}	0.18%	100%	0.0018	1.26	23
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Ectothiorhodospirales_Ectothiorhodospiraceae_}	{Classification=Rare}	0.36%	100%	0.0036	1.26	46
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Alteromonadaceae_Glaciecola}	{Classification=Rare}	0.19%	100%	0.0019	1.26	24

{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Kangiellaceae_}	{Classification=Rare}	0.53%	100%	0.0053	1.26	67
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Moritellaceae_Moritella}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Psychromonadaceae_Psychromonas}	{Classification=Rare}	0.25%	100%	0.0025	1.26	32
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Vibrionaceae_Photobacterium}	{Classification=Rare}	0.15%	100%	0.0015	1.26	19
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Francisellales_Francisellaceae_Francisella}	{Classification=Rare}	0.16%	100%	0.0016	1.26	20
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Ga0077536}	{Classification=Rare}	0.51%	100%	0.0051	1.26	65
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_HOC36,WaterMass=TAW}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Legionellales_Legionellaceae_}	{Classification=Rare}	0.20%	100%	0.0020	1.26	25
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Nitrosococcales_Nitrosococcaceae_Cm1-21}	{Classification=Rare}	0.26%	100%	0.0026	1.26	33
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Halieaceae_Halioglobus}	{Classification=Rare}	0.19%	100%	0.0019	1.26	24
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Halieaceae_Luminiphilus}	{Classification=Rare}	0.17%	100%	0.0017	1.26	21
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Halomonadaceae_Halomonas}	{Classification=Rare}	0.13%	100%	0.0013	1.26	17
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_KI89A clade_}	{Classification=Rare}	0.74%	100%	0.0074	1.26	94
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Litoricolaceae_Litoricola}	{Classification=Rare}	0.19%	100%	0.0019	1.26	24
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_OM182 clade_,Chl=very high}	{Classification=Rare}	0.13%	100%	0.0013	1.26	17
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_Porticoccus}	{Classification=Rare}	0.49%	100%	0.0049	1.26	62
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Spongiibacteraceae_BD1-7 clade}	{Classification=Rare}	0.34%	100%	0.0034	1.26	43
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Spongiibacteraceae_Oceanicoccus}	{Classification=Rare}	0.33%	100%	0.0033	1.26	42
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Spongiibacteraceae_Sinobacterium}	{Classification=Rare}	0.37%	100%	0.0037	1.26	47
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Thioglobaceae_Candidatus Thioglobus,PelagicLayer=Mesopelagic}	{Classification=Rare}	0.10%	100%	0.0010	1.26	13
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_SS1-B-07-19}	{Classification=Rare}	0.39%	100%	0.0039	1.26	50

{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Tenderiales_Tenderiaceae_Candidatus Tenderia}	{Classification=Rare}	0.30%	100%	0.0030	1.26	38
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Thiotrichales_Thiotrichaceae_Leucothrix}	{Classification=Rare}	0.14%	100%	0.0014	1.26	18
{taxon=Bacteria_Verrucomicrobiota_Kiritimatiellae_WCHB1-41}	{Classification=Rare}	0.21%	100%	0.0021	1.26	27
{taxon=Bacteria_Verrucomicrobiota_Lentisphaeria_P.palmC41}	{Classification=Rare}	0.11%	100%	0.0011	1.26	14
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Opitutales_Opitutaceae_Diplosphaera}	{Classification=Rare}	0.19%	100%	0.0019	1.26	24
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Opitutales_Puniceicoccaceae_MB11C04 marine group}	{Classification=Rare}	0.49%	100%	0.0049	1.26	62
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Pedosphaerales_Pedosphaeraceae_SCGC AAA164-E04}	{Classification=Rare}	0.23%	100%	0.0023	1.26	29

End of Supplementary Table 11.42.

Supplementary Table 11.43. Top100 rules, with Conviction above 5, Confidence above 90%, Lift above 1, and selected non-redundant rules based on Complexity. Rules ordered by Lift. Additional metrics are provided (Support, Coverage and Count). Note that: "NO2" = NO_2^-; "NO3" = NO_3^-; "Chl" = Chl *a*; "PO4" = PO_4^-; "NH4" = NH_4^+.

Antecedent	Consequent	Support	Confidence	Coverage	Lift	Count	Conviction
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,NO2=very high,NH4=Unknown}	{Classification=Undetermined}	0.11%	93.33%	0.0012	5.79	14	12.58
{taxon=Bacteria_Planctomycetota_OM190_,PO4=very high,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Undetermined}	0.10%	92.86%	0.0011	5.76	13	11.74
{taxon=Archaea_Crenarchaeota_Nitrososphaeria_Nitrosopumilales_Nitrosopumilaceae_Candidatus Nitrosopelagicus,PO4=very high,NO2=very high}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Archaea_Crenarchaeota_Nitrososphaeria_Nitrosopumilales_Nitrosopumilaceae_Candidatus Nitrosopumilus,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Archaea_Nanoarchaeota_Nanoarchaeia_Woesearchaeales_SCGC AAA011-D5_,Si=very high,Chl=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Archaea_Thermoplasmatota_Thermoplasmata_Marine Group II,PelagicLayer=Epipelagic,PO4=very low}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Acidobacteriota_Vicinamibacteria_Vicinamibacterales_Vicinamibacteraceae_}	{Classification=Rare}	0.15%	100.00%	0.0015	1.26	19	Inf

{taxon=Bacteria_Actinobacteriota_Acidimicrobiia_Actinomarinales_Actinomarinaceae_Candidatus Actinomarina,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Actinobacteriota_Acidimicrobiia_Microtrichales_Ilumatobacteraceae_Ilumatobacter,PelagicLayer=Epipelagic,Temperature=low,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Actinobacteriota_Acidimicrobiia_Microtrichales_Microtrichaceae_,PO4=very high,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Actinobacteriota_Acidimicrobiia_Microtrichales_Microtrichaceae_Sva0996 marine group,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.15%	100.00%	0.0015	1.26	19	Inf
{taxon=Bacteria_Actinobacteriota_Actinobacteria_,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Actinobacteriota_Actinobacteria_Corynebacteriales_Nocardiaceae_Rhodococcus,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Actinobacteriota_Actinobacteria_Micrococcales_Microbacteriaceae_Candidatus Aquiluna,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.15%	100.00%	0.0015	1.26	19	Inf
{taxon=Bacteria_Actinobacteriota_Actinobacteria_Micrococcales_Microbacteriaceae_ML602J-51,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Actinobacteriota_Actinobacteria_PeM15,PelagicLayer=Epipelagic,PO4=very low,Salinity=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Actinobacteriota_Actinobacteria_Propionibacteriales_Propionibacteriaceae_Cutibacterium}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Bacteroidales_Prolixibacteraceae_,PelagicLayer=Epipelagic,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Chitinophagales,PelagicLayer=Epipelagic,Temperature=very high}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Chitinophagales_Saprospiraceae_,NH4=Unknown,Salinity=very high,WaterMass=AW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf

{taxon=Bacteria_Bacteroidota_Bacteroidia_Chitinophagales_Saprospiraceae_Aureispira,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Cytophagales_Cyclobacteriaceae_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Cytophagales_Cyclobacteriaceae_Marinoscillum,PelagicLayer=Epipelagic,NH4=Unknown,Temperature=very high}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Crocinitomicaceae_Crocinitomix,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Crocinitomicaceae_Salinirepens,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_NS10 marine group,PelagicLayer=Epipelagic,Temperature=very high,WaterMass=AW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae_Vicingus,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Flavicella,PelagicLayer=Epipelagic,Temperature=very high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Flavobacterium,PelagicLayer=Epipelagic}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Lutibacter,PelagicLayer=Epipelagic}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Maritimimonas,PelagicLayer=Epipelagic}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS2b marine group,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.16%	100.00%	0.0016	1.26	20	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_NS3a marine group,PelagicLayer=Epipelagic,WaterMass=AW}	{Classification=Rare}	0.15%	100.00%	0.0015	1.26	19	Inf

{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Pseudofulvibacter,Chl=Unknown,Salinity=high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Flavobacteriaceae_Psychroserpens,PO4=very high,Si=very high}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_NS7 marine group_,Chl=Unknown,Salinity=high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Sphingobacteriales_Lentimicrobiaceae_}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Sphingobacteriales_NS11-12 marine group_,PelagicLayer=Epipelagic,Salinity=very high,WaterMass=AW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Bacteroidota_Bacteroidia_Sphingobacteriales_Sphingobacteriaceae_Pedobacter,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Bdellovibrionota_Bdellovibrionia_Bacteriovoracales_Bacteriovoracaceae_Peredibacter,PelagicLayer=Epipelagic,WaterMass=AW}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Bdellovibrionota_Bdellovibrionia_Bdellovibrionales_Bdellovibrionaceae_OM27 clade,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.14%	100.00%	0.0014	1.26	18	Inf
{taxon=Bacteria_Bdellovibrionota_Oligoflexia_Oligoflexales_Oligoflexaceae_}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Campylobacterota_Campylobacteria_Campylobacterales_Sulfurimonadaceae_Sulfurimonas}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Chloroflexi_Anaerolineae_Anaerolineales_Anaerolineaceae_}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Cyanobacteria_Cyanobacteriia_Synechococcales_Cyanobiaceae_Synechococcus CC9902,PelagicLayer=Epipelagic,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Dadabacteria_Dadabacteriia_Dadabacteriales,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Desulfobacterota__,Chl=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf

{taxon=Bacteria_Desulfobacterota_Desulfuromonadia_Brady monadales,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Desulfobacterota_Desulfuromonadia_PB19, PO4=very high,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Firmicutes_Bacilli_Staphylococcales_Staphyl ococcaceae_Staphylococcus,NH4=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Gemmatimonadota_BD2-11 terrestrial group_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Marinimicrobia (SAR406 clade)__,PelagicLayer=Epipelagic,NH4=Unknown,Chl=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Myxococcota_Myxococcia_Myxococcales_M yxococcaceae_P3OB-42,PO4=very high,Si=very high}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Myxococcota_Polyangia_}	{Classification=Rare}	0.14%	100.00%	0.0014	1.26	18	Inf
{taxon=Bacteria_NB1-j__,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Nitrospinota_Nitrospinia_Nitrospinales_Nitro spinaceae_LS-NOB,PelagicLayer=Epipelagic,Chl=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Nitrospinota_Nitrospinia_Nitrospinales_Nitro spinaceae_Nitrospina,PelagicLayer=Epipelagic,NH4=Unknow n}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_PAUC34f__,NH4=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycispha erales_AKAU3564 sediment group_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycispha erales_Phycisphaeraceae_,NH4=Unknown}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycispha erales_Phycisphaeraceae_CL500-3,PelagicLayer=Epipelagic,Temperature=very high,WaterMass=AW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycispha erales_Phycisphaeraceae_FS140-16B-02 marine group,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf

{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycisphaerales_Phycisphaeraceae_JL-ETNP-F27,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Planctomycetota_Phycisphaerae_Phycisphaerales_Phycisphaeraceae_Urania-1B-19 marine sediment group,PO4=very high,Chl=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_,PelagicLayer=Epipelagic,NH4=Unknown,Si=low}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_Blastopirellula,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.14%	100.00%	0.0014	1.26	18	Inf
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_Pir4 lineage}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_Rhodopirellula,PelagicLayer=Mesopelagic,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Planctomycetota_Planctomycetes_Pirellulales_Pirellulaceae_Rubripirellula}	{Classification=Rare}	0.16%	100.00%	0.0016	1.26	20	Inf
{taxon=Bacteria_Planctomycetota_Planctomycetes_Planctomycetales_Gimesiaceae_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Planctomycetota_Planctomycetes_Planctomycetales_Rubinisphaeraceae_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria__,PelagicLayer=Epipelagic,Temperature=very high,WaterMass=AW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_AT-s3-44,PO4=very high,Si=very high,Chl=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Caulobacterales_Caulobacteraceae_Brevundimonas,NH4=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Defluviicoccales,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.15%	100.00%	0.0015	1.26	19	Inf

{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Kordiimonadales,PelagicLayer=Epipelagic,PO4=very low,Temperature=very high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Parvibaculales_OCS116 clade_,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.14%	100.00%	0.0014	1.26	18	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Parvibaculales_Parvibaculaceae_}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Puniceispirillales,PelagicLayer=Epipelagic,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Puniceispirillales_EF100-94H03_,NO2=very high}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Puniceispirillales_SAR116 clade_,PelagicLayer=Epipelagic,PO4=very low,Temperature=very high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Puniceispirillales_SAR116 clade_Candidatus Puniceispirillum,PelagicLayer=Epipelagic,Temperature=very high}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Lentibacter,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Octadecabacter,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.17%	100.00%	0.0017	1.26	21	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Roseobacter clade C11-7 lineage,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Tateyamaria,PelagicLayer=Epipelagic,WaterMass=AW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodobacterales_Rhodobacteraceae_Yoonia-	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf

Loktanella,PelagicLayer=Mesopelagic,PO4=very high,NO2=very high,Chl=Unknown}							
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_AEGEAN-169 marine group_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Magnetospiraceae_,PelagicLayer=Epipelagic,PO4=very low,Si=very low}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rhodospirillales_Magnetospiraceae_Magnetospira,PelagicLayer=Epipelagic,PO4=very low,Temperature=very high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales_Rickettsiaceae_Candidatus Megaira}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Rickettsiales_S25-593_}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ib,PelagicLayer=Mesopelagic,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade II_,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.15%	100.00%	0.0015	1.26	19	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade III_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade IV_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_Thalassobaculales_Nisaeaceae_OM75 clade,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf

{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Arenicellales_Arenicellaceae_Arenicella}	{Classification=Rare}	0.16%	100.00%	0.0016	1.26	20	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Comamonadaceae_Polaromonas,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Comamonadaceae_Rhodoferax,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Comamonadaceae_RS62 marine group,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Gallionellaceae_Sideroxydans,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Hydrogenophilaceae_Thiobacillus,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Methylophilaceae_OM43 clade,PelagicLayer=Epipelagic,PO4=very low,Temperature=very high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_IS-44,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Burkholderiales_Nitrosomonadaceae_Nitrosomonas,PelagicLayer=Epipelagic,Chl=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Coxiellales_Coxiellaceae_Coxiella,Temperature=very low}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Ectothiorhodospirales_Ectothiorhodospiraceae_,Salinity=very high,WaterMass=AW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Alteromonadaceae_Glaciecola,PelagicLayer=Epipelagic,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf

{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Colwelliaceae_Colwellia,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Kangiellaceae_,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Moritellaceae_Moritella}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Pseudoalteromonadaceae_Pseudoalteromonas,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Psychromonadaceae_Psychromonas,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Vibrionaceae_Photobacterium}	{Classification=Rare}	0.15%	100.00%	0.0015	1.26	19	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Francisellales_Francisellaceae_Francisella,NO2=very high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Ga0077536,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_HOC36,PelagicLayer=Mesopelagic,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Legionellales_Legionellaceae_,NH4=Unknown,Temperature=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Nitrosococcales_Methylophagaceae_,PelagicLayer=Epipelagic,Temperature=low,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Nitrosococcales_Nitrosococcaceae_Cm1-21,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	16	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Cellvibrionaceae_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Halieaceae_Halioglobus,NO2=very high}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf

{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Halieaceae_Luminiphilus,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Halieaceae_Marimicrobium,PelagicLayer=Epipelagic,Si=low}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Halomonadaceae_Halomonas,NH4=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_KI89A clade_,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Litoricolaceae_Litoricola,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Moraxellaceae_Psychrobacter,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_OM182 clade_,PelagicLayer=Epipelagic,PO4=very low,Temperature=very high}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_Porticoccus,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Pseudohongiellaceae_Pseudohongiella,PelagicLayer=Epipelagic,NH4=Unknown,Temperature=very high}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Spongiibacteraceae_BD1-7 clade,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Spongiibacteraceae_Oceanicoccus,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Spongiibacteraceae_Sinobacterium,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf

{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Thioglobaceae_Candidatus Thioglobus,PelagicLayer=Mesopelagic,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Thioglobaceae_SUP05 cluster,PelagicLayer=Epipelagic,PO4=very low,Si=very low}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_SS1-B-07-19,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Steroidobacterales_Woeseiaceae_Woeseia,PelagicLayer=Epipelagic,Temperature=very high,WaterMass=AW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Tenderiales_Tenderiaceae_Candidatus Tenderia,PelagicLayer=Epipelagic,Temperature=high}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Thiotrichales_Thiotrichaceae_,PelagicLayer=Epipelagic,NH4=Unknown,Salinity=very low,WaterMass=SW}	{Classification=Rare}	0.14%	100.00%	0.0014	1.26	18	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Thiotrichales_Thiotrichaceae_Leucothrix,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Verrucomicrobiota_Kiritimatiellae_WCHB1-41,PO4=very high,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Verrucomicrobiota_Lentisphaeria_Lentisphaerales_Lentisphaeraceae_Lentisphaera,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Verrucomicrobiota_Lentisphaeria_P.palmC41}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Arctic 97B-4 marine group,PelagicLayer=Mesopelagic,PO4=very high,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Opitutales_Opitutaceae_Diplosphaera,PelagicLayer=Epipelagic}	{Classification=Rare}	0.18%	100.00%	0.0018	1.26	23	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Opitutales_Puniceicoccaceae_Lentimonas,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown}	{Classification=Rare}	0.12%	100.00%	0.0012	1.26	15	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Opitutales_Puniceicoccaceae_MB11C04 marine group,PO4=very high,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf

{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Opitutales_Puniceicoccaceae_Pelagicoccus,NO2=very high,NH4=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Opitutales_Puniceicoccaceae_Verruc-01,PelagicLayer=Epipelagic}	{Classification=Rare}	0.13%	100.00%	0.0013	1.26	17	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Pedosphaerales_Pedosphaeraceae_SCGC AAA164-E04,PelagicLayer=Epipelagic,NH4=Unknown}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Verrucomicrobiales_DEV007_,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,Temperature=very low}	{Classification=Rare}	0.10%	100.00%	0.0010	1.26	13	Inf
{taxon=Bacteria_Verrucomicrobiota_Verrucomicrobiae_Verrucomicrobiales_Rubritaleaceae_Rubritalea,PelagicLayer=Mesopelagic,NO2=very high,Chl=Unknown,WaterMass=TAW}	{Classification=Rare}	0.11%	100.00%	0.0011	1.26	14	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Enterobacterales_Alteromonadaceae_Paraglaciecola}	{Classification=Rare}	0.20%	96.15%	0.0021	1.21	25	5.39

End of Supplementary Table 11.43.

Supplementary Table 11.44. Non-redundant and dependent rules that are taxonomically relevant. Rules were selected based on Conviction > 5, Confidence > 90%, and Lift > 1. Complexity was used to remove redundancy. Note that: "NO2" = NO_2^-; "NO3" = NO_3^-; "Chl" = Chl *a*; "PO4" = PO_4^-; "NH4" = NH_4^+.

Antecedent	Consequent	Support	Confidence	Lift	Conviction
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade I_Clade Ia,PelagicLayer=Mesopelagic,NO2=very high,Si=very high,Chl=Unknown}	{Classification=Abundant}	0.10%	92.86%	20.23	13.36
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_Porticoccaceae_SAR92 clade,NO2=very high,NH4=Unknown}	{Classification=Undetermined}	0.11%	93.33%	5.80	12.59
{taxon=Bacteria_Bacteroidota_Bacteroidia_Flavobacteriales_Cryomorphaceae,PelagicLayer=Mesopelagic,PO4=very high,NO2=very high,Chl=Unknown}	{Classification=Undetermined}	0.10%	92.86%	5.77	11.75
{taxon=Bacteria_Planctomycetota_OM190, PO4=very high,Si=very high,Chl=Unknown,Temperature=very low}	{Classification=Undetermined}	0.10%	92.86%	5.77	11.75
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade III}	{Classification=Rare}	0.68%	100.00%	1.26	Inf
{taxon=Bacteria_Proteobacteria_Gammaproteobacteria_Pseudomonadales_KI89A clade}	{Classification=Rare}	0.74%	100.00%	1.26	Inf
{taxon=Bacteria_Proteobacteria_Alphaproteobacteria_SAR11 clade_Clade IV}	{Classification=Rare}	0.76%	98.98%	1.25	20.27