

THESE DE DOCTORAT DE

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE

ECOLE DOCTORALE N° 598
Sciences de la Mer et du littoral
Spécialité : *Microbiologie*

Par

Blandine TROUCHE

**Exploring the deep ocean seafloor Bacteria and Archaea, from
microbial community structure to comparative genomics**

Thèse présentée et soutenue à **Plouzané**, le **7 juillet 2021**

Unité de recherche : Laboratoire de Microbiologie des Environnements Extrêmes (UMR 6197)

Rapporteurs avant soutenance :

Christian TAMBURINI	Directeur de recherche CNRS, MIO, Marseille
A. Murat EREN	Assistant Professor, University of Chicago

Composition du Jury :

Christian TAMBURINI	Directeur de recherche CNRS, MIO, Marseille
Ingrid OBERNOSTERER Présidente du jury	Directrice de recherche CNRS, LOMIC, Banyuls sur mer
Purificación LOPEZ-GARCIA	Directrice de recherche CNRS, ESE, Orsay
Ronnie N. GLUD	Professor, University of Southern Denmark
Sophie ARNAUD-HAOND Directrice de thèse	Chercheuse, Ifremer centre de Sète
Loïs MAIGNIEN Co-directeur de thèse	Maître de conférence, Université de Bretagne Occidentale
Invité(s) Jean-Christophe AUGUET	Chargé de recherche CNRS, MARBEC, Sète

Remerciements

J'aimerais en premier lieu remercier l'IFREMER et le projet Pourquoi pas les Abysses, et le Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation pour leur financement de ce projet de recherche via un contrat doctoral au sein de l'Université de Bretagne Occidentale. Merci également à Mohamed Jebbar de m'avoir accueillie au sein du LM2E pour ces années de thèse.

Je remercie l'ensemble des membres de mon jury de thèse de leur intérêt pour mon sujet et d'avoir accepté d'évaluer ce travail. Merci à Christian Tamburini et à Meren d'endosser le rôle de rapporteurs de thèse, et merci à Ingrid Obernosterer, Purificacion Lopez-Garcia et Ronnie N. Glud d'avoir accepté le rôle d'examineur. Un double merci à Purificacion Lopez-Garcia pour m'avoir également fait profiter de son expérience au cours de ces années de thèse dans le cadre du comité de suivi, avec l'apport de Pierre Offre et Pierre Galand, que je remercie chaudement pour leur bienveillance et leurs conseils.

Bien sûr, un grand merci à mes encadrants de thèse pour leur confiance dans la réalisation de ce projet, et leur accompagnement. Merci Sophie pour les discussions formelles et informelles, pour ton accueil à Sète, et pour ta disponibilité pour m'aider à jongler entre rédaction d'un premier papier et expérience d'un premier confinement. Merci Loïs pour ta présence tout au long de ce travail, ton exigence scientifique mais aussi ta confiance pour m'aider à progresser en autonomie, et merci pour les sorties en bateau ! Merci Jean-Christophe pour ton retour éclairé sur les aspects techniques de l'écologie microbienne. Merci encore à tous les trois pour les opportunités que vous m'avez offertes, que ce soit le départ en campagne, la participation aux workshops EBAME, ou tout simplement l'ensemble de cette thèse.

I want to address my thanks to Ronnie N. Glud and Frank Wenzhöfer for their collaboration with our team and for agreeing to have me onboard for the Atacama So261 cruise. It was an experience like no other and an amazing scientific opportunity. Thank you also to Bo Thamdrup for his advice, and of course thank you Clemens, for the scientific discussions and for the memes.

Je voudrais remercier tous ceux que j'ai côtoyés au sein du labo, celui de Sète et celui de Brest, pour leur accueil et leur amitié. A Sète je remercie bien sûr Miriam et Babett, pour les moments de travail et les moments de détente.

Je tiens à remercier tous les collègues de Brest qui m'ont permis de m'y sentir accueillie et soutenue. Un merci particulier à Stéphanie pour ton soutien moral indéfectible, et aussi au service info de l'IUEM et au SEBIMER pour toutes les pannes techniques résolues plus vite que l'éclair.

Toutes mes amitiés à l'équipe de doctorants et de stagiaires que j'ai eu la chance de côtoyer au LM2E. Je n'aurais jamais pensé m'intégrer aussi vite, et je n'ai jamais regretté d'avoir quitté le Sud pour Brest ! Alors, sans chercher à mettre de l'ordre : merci Sarah, Jordan, Coraline, Seb, Maxime, Jie, Yang et Francis, Marion, David, Maurane. Merci Marc pour les galères partagées depuis le tout début (surtout l'administratif), merci Ferial pour ton aide précieuse,

merci Flo pour toutes les blagues de papa et les astuces de bioinfo, merci Johanne pour ta disponibilité et ta patience, merci Ashley for the Thursday lunches, carpool gossips, NJ trivia and pizza judgement, et mille mercis Clarisse pour... tout ! Les pique-niques, les thés, les belotes, les cinés que pour nous, mais surtout ta bonne humeur et ton soutien indéfectibles.

L'année 2020 a été pour le moins bousculée, et je tiens à remercier mes compagnons de confinement successifs : merci Laura pour ton dynamisme et ta bonne humeur pendant le premier confinement. Merci Alan de n'avoir jamais râlé pendant mes quatre mois d'occupation constante du salon de la coloc en télétravail. Merci Seb et Oriane de m'avoir accueillie pendant le confinement (bis), ça a été un plaisir d'apprendre à vous connaître autour de soirées jeux et de discussions passionnantes sur le jeu de rôle ou le féminisme.

Un peu plus loin, merci aux rennais, Gabrielle, Pierre, Pierre-Yves, Mélodie, Mikou et Jérémy de m'avoir accueillie à bras ouverts, et aussi d'avoir la patience de me réinviter à chaque fois que la thèse m'empêche de venir vous rendre visite.

Merci au groupe M, je suis ravie d'avoir partagé ces dernières années avec vous. Un merci particulier à Elodie, Geneviève et Imane, pour votre écoute et votre empathie dans les moments moins faciles. Merci aussi pour votre dynamisme et vos propositions toujours renouvelées de week-ends, de voyages, et autres concerts.

Merci à Adélie, Anouk, Laurie, Sarah et Maxime, même s'il est parfois difficile de se voir, j'apprécie énormément la diversité de nos parcours et le soutien que vous représentez pour moi depuis le lycée.

Enfin un immense merci à mes parents et mes frères et sœurs pour leur amour et leur soutien de toujours. Si nous sommes tous devenus des scientifiques, il doit y avoir comme un gène commun... Un merci tout particulier à mes parents pour leur soutien sans faille dans toutes mes aventures, de Brest à la Nouvelle-Zélande. Merci pour vos compliments sur mes figures, complètement abstraites à vos yeux sans doute, pendant la rédaction de ce manuscrit !

Pour terminer, bien sûr un grand merci à toi Johan de m'avoir laissé te convaincre de déménager à Brest et de vivre une thèse par procuration... Merci pour ton écoute, ta patience et ton soutien précieux.

TABLE OF CONTENTS

INTRODUCTION	15
1. STUDYING THE MICROBIAL ECOLOGY OF THE DARK OCEAN: FROM THE FIRST EXPEDITIONS TO MOLECULAR ECOLOGY APPROACHES	16
1.1. <i>From the first expeditions of the 19th century to the large-scale explorations of the 21st</i>	<i>16</i>
1.1.1. Reaching the deep sea.....	16
1.1.2. Advent of molecular ecology and sequencing	17
1.1.3. Microbial ecology at the global scale.....	18
1.2. <i>Molecular tools to access the functional and taxonomic diversity of microbial communities</i>	<i>19</i>
1.2.1. Metabarcoding	19
1.2.1.1. Marker gene selection.....	19
1.2.1.2. Delineating ecologically informative units	20
1.2.1.3. Limitations and biases.....	21
1.2.2 Whole genome sequencing	22
1.2.2.1. Read-centric analysis.....	22
1.2.2.2. Assembly	22
1.2.2.3. Binning.....	23
1.2.2.4. Comparative genomics.....	24
2. LIFE IN DEEP OCEAN SURFACE SEDIMENTS.....	25
2.1. <i>Geologic and biogeographic definition of seafloor provinces.....</i>	<i>25</i>
2.1.1 Geologic classification.....	25
2.1.2. Biogeographic classification.....	27
2.1.3. Hadal trenches.....	27
2.1.3.1. Atacama and Kermadec trenches	28
2.2. <i>Metabolic reactions in deep sea sediments.....</i>	<i>29</i>
2.2.1. General diagenetic sequence.....	29
2.2.2. Organic matter input and oxygen penetration depth	32
2.2.3. Oxygen penetration depth in hadal and adjoining abyssal sediments of the Atacama and Kermadec trenches.....	33
2.3. <i>Microbial diversity in benthic sediments.....</i>	<i>34</i>
2.3.1. Microbial cell counts.....	34
2.3.2. Community composition.....	35
2.3.3. Diversity in surface sediments of hadal trenches	36
2.4. <i>Biogeographic patterns of the abyssal plains.....</i>	<i>37</i>
2.4.1. General biogeographic patterns and processes.....	37
2.4.2. Biogeography of deep sea benthic sediments.....	39
3. TAXONOMY OF ARCHAEA.....	41
3.1. <i>Brief history.....</i>	<i>41</i>
3.2. <i>Recent discussions around archaeal taxonomy</i>	<i>42</i>

OVERVIEW AND OBJECTIVES.....	47
OBJECTIFS DE LA THESE.....	49
CHAPTER 1: DEFINING STANDARDIZED METHODS FOR THE STUDY OF BENTHIC PROKARYOTIC AND EUKARYOTIC DIVERSITY	51
CONTEXT OF THIS WORK AND PERSONAL CONTRIBUTION.....	52
1. EVALUATING SEDIMENT AND WATER SAMPLING METHODS FOR THE ESTIMATION OF DEEP-SEA BIODIVERSITY USING ENVIRONMENTAL DNA.....	53
RESUME DE L'ARTICLE EN FRANÇAIS	53
2. AN ASSESSMENT OF ENVIRONMENTAL METABARCODING PROTOCOLS AIMING AT FAVORING CONTEMPORARY BIODIVERSITY IN INVENTORIES OF DEEP-SEA COMMUNITIES	69
RESUME DE L'ARTICLE EN FRANÇAIS	69
3. BIOINFORMATIC PIPELINES COMBINING DENOISING AND CLUSTERING TOOLS ALLOW FOR MORE COMPREHENSIVE PROKARYOTIC AND EUKARYOTIC METABARCODING.....	83
RESUME DE L'ARTICLE EN FRANÇAIS	83
4. COMPARISON OF TWO 16S rRNA AMPLICON PRIMERS AND METAGENOMIC DATA FOR DEEP-SEA BENTHIC ARCHAEAL DIVERSITY STUDIES	104
RESUME DE L'ARTICLE EN FRANÇAIS	105
ABSTRACT	107
INTRODUCTION	108
MATERIAL & METHODS.....	111
1. <i>Environmental samples collection</i>	111
2. <i>DNA extraction</i>	112
3. <i>Libraries construction and sequencing</i>	112
3.1. Metabarcoding.....	112
3.2. Metagenomics	113
4. <i>In silico primer specificity evaluation with TestPrime</i>	113
5. <i>Amplicon datasets bioinformatic analysis</i>	114
6. <i>Metagenomic reads analysis</i>	115
6.1. Ribosomal SSU sequences (miTags)	115
6.2. Single copy core genes	116
7. <i>Comparison of the datasets</i>	116
RESULTS.....	118
1. <i>In silico primer specificity evaluation using SILVA v138 16S rRNA database</i>	118
2. <i>Datasets description</i>	119
2.1. Universal primers dataset.....	120
2.2. Archaeal primers dataset	121
2.3. miTAGs dataset and SCG profiles	121
3. <i>Comparative taxonomic profiles on deep-sea benthic samples</i>	122

4. <i>Multi-dimensional analysis based on phylogenetic distance</i>	127
DISCUSSION	129
CONCLUSION.....	133
SUPPLEMENTARY FIGURES.....	134

CHAPTER 2: DIVERSITY AND BIOGEOGRAPHY OF BATHYAL AND ABYSSAL SEAFLOOR

BACTERIA AND ARCHAEA ALONG A MEDITERRANEAN - ATLANTIC GRADIENT 140

RESUME DE L'ARTICLE EN FRANÇAIS	142
ABSTRACT	143
INTRODUCTION	144
MATERIAL & METHODS.....	146
1. <i>Sample collection and processing</i>	146
1.1. Cruises and locations.....	146
1.2. Sampling protocol.....	147
2. <i>DNA extraction</i>	148
3. <i>Libraries construction and sequencing</i>	148
4. <i>Bioinformatic analysis</i>	149
5. <i>Sediments characterization</i>	150
6. <i>Statistical analysis</i>	151
RESULTS.....	153
1. <i>16s rRNA gene amplicon processing</i>	153
2. <i>Description of sampling sites</i>	154
3. <i>Distance-decay relationship between deep sea sediment communities</i>	155
4. <i>Distance-decay relationship depending on sediment horizons</i>	157
5. <i>Environmental parameters structuring microbial communities</i>	158
6. <i>Exploring the link between surface and subsurface communities at local scale</i>	162
DISCUSSION	164
CONCLUSION.....	169
SUPPLEMENTARY FIGURES AND TABLES	172

CHAPTER 3: DISTRIBUTION OF ARCHAEA AND PUTATIVE ASSOCIATIONS OF MEMBERS OF THE NANOARCHAEOTA PHYLUM IN ABYSSAL AND HADAL SURFACE SEDIMENTS

REVEALED BY NETWORK ANALYSIS 179

RESUME DE L'ARTICLE EN FRANÇAIS	181
ABSTRACT	182
INTRODUCTION	183
MATERIAL & METHODS.....	186
1. <i>Sample collection and processing</i>	186
2. <i>DNA extraction, library construction and sequencing</i>	188

3. <i>Bioinformatic processing</i>	188
4. <i>Statistical analysis</i>	189
RESULTS.....	190
1. <i>Dataset description</i>	190
2. <i>Overview of archaeal diversity</i>	190
3. <i>Archaeal co-occurrence network</i>	193
4. <i>Putative associations of Woesearchaeales</i>	196
DISCUSSION	198
1. <i>Influence of habitat, sediment depth and trench of origin on archaeal benthic communities</i> 198	
2. <i>Distribution and modules of Nitrososphaeria</i>	200
3. <i>Putative associations of Woesearchaeales</i>	201
CONCLUSION AND PERSPECTIVES	203
SUPPLEMENTARY FIGURES	204

**CHAPTER 4: CLADE DISTRIBUTION AND GENOMIC VARIATION OF AMMONIA OXIDIZING
ARCHAEA IN ABYSSAL AND HADAL SURFACE SEDIMENTS.....208**

RESUME DE L'ARTICLE EN FRANÇAIS	210
ABSTRACT	211
INTRODUCTION	212
MATERIAL & METHODS.....	215
1. <i>Sampling sites, slicing scheme and DNA extraction</i>	215
2. <i>Assembly and binning</i>	215
3. <i>Reference genomes</i>	216
4. <i>Phylogenetic placement of amoA reconstructed genes</i>	216
5. <i>Taxonomic placement of MAGs</i>	217
6. <i>Single nucleotide and single amino acid variant analyses</i>	217
RESULTS AND DISCUSSION	219
1. <i>Distribution of AOA clades in abyssal and hadal benthic sediments</i>	219
2. <i>Phylogenomic placement and distribution of MAGs affiliated to class Nitrososphaeria</i>	224
3. <i>Sequence variability of AOA MAGs highlights differences in selective pressure</i>	229
CONCLUSION.....	231
SUPPLEMENTARY FIGURES	232
.....	233

GENERAL DISCUSSION234

1. LARGE-SCALE ECOLOGICAL STUDY OF THE DEEP OCEAN SEAFLOOR IN THE AGE OF NGS.....	235
1.1. <i>Accessing archaeal diversity with Next Generation Sequencing</i>	235

1.2. <i>Expanding the database of metagenomes to study the global distribution of Bacteria and Archaea</i>	237
1.3. <i>A matter of scales</i>	238
1.4. <i>The importance of a holistic approach for ecosystem characterization</i>	240
2. MOLECULAR APPROACHES TO UNCOVER NEW DIVERSITY: LIMITS AND PERSPECTIVES	241
2.1. <i>Challenges in linking 16S and metagenomic inventories of diversity</i>	241
2.2. <i>Limitations due to lack of completeness in the databases</i>	242
2.3. <i>Discussions around archaeal taxonomy</i>	242
3. PERSPECTIVES FOR DEEP SEA RESEARCH.....	244
3.1. <i>Importance of experimental evidence to complement molecular results</i>	244
3.2. <i>Establishment of long-term observatories</i>	245
CONCLUSION AND PERSPECTIVES	247
REFERENCES.....	252
SUPPLEMENTARY MATERIAL	278
SUPPLEMENTARY MATERIAL FOR CHAPTER 1.....	279
1. <i>Amplicon libraries preparation and sequencing</i>	279
2. <i>Metagenomic libraries preparation and sequencing</i>	280
SUPPLEMENTARY MATERIAL FOR CHAPTER 2.....	282
1. <i>PCR amplification</i>	282
2. <i>Sequencing</i>	282
3. <i>Sediment characterization</i>	283
SUPPLEMENTARY MATERIAL FOR CHAPTER 4.....	285

List of figures

FIGURE 1: SEQUENCING COST PER MEGABASE OF DNA - AUGUST 2020.....	18
FIGURE 2: SCHEMATIC VIEW OF A STYLIZED CROSS SECTION OF DARK OCEAN HABITATS (TOP) AND SEDIMENT ZONATION (BOTTOM).....	26
FIGURE 3: GEOGRAPHIC LOCATION OF THE HADAL TRENCHES OF A) THE SOUTHWEST PACIFIC OCEAN AND B) THE SOUTHEAST PACIFIC, ATLANTIC AND SOUTHERN OCEANS. (JAMIESON, 2015)	29
FIGURE 4: SCHEMATIC ILLUSTRATION OF THE METABOLIC PROCESSES TAKING PLACE WITH DEPTH IN SEAFLOOR SEDIMENTS. (PARKES ET AL, 2014).....	32
FIGURE 5: EVOLUTION OF THE ARCHAEOAL TREE OF LIFE OVER THE YEARS. (SPANG ET AL., 2017).....	42
FIGURE 6: RANK NORMALIZED ARCHAEOAL GTDB TAXONOMY PROPOSED BY RINKE ET AL, 2021.....	44
FIGURE 7: RECLASSIFICATION OF THAUMARCHAEOTA MEMBERS PROPOSED BY RINKE ET AL (2021).	46
FIGURE 8: COMPOSITION OF DATASETS AT DOMAIN AND PHYLUM LEVEL.	124
FIGURE 9: DENDROGRAM OF THE TAXA REPRESENTING MORE THAN 0.025% OF THE DATASETS.....	127
FIGURE 10: EDGE PRINCIPAL COMPONENTS ANALYSIS OF THE THREE DATASETS (UNIVERSAL PRIMERS DATA, ARCHAEOAL PRIMERS AND MITAGS) BASED ON PLACEMENT OF THE SEQUENCES IN THE SILVA REFERENCE TREE.	128
FIGURE 11: DESCRIPTION OF SAMPLING SITES:	154
FIGURE 12: PAIRWISE BRAY-CURTIS COMMUNITY SIMILARITY BETWEEN SAMPLES WITH RESPECT TO GEOGRAPHIC DISTANCE (KM) AND ENVIRONMENTAL SIMILARITY	156
FIGURE 13: PAIRWISE BRAY-CURTIS COMMUNITY SIMILARITY WITH RESPECT TO GEOGRAPHIC DISTANCE (KM) BETWEEN SAMPLES.....	158
FIGURE 14: NON-METRIC MULTIDIMENSIONAL SCALING ORDINATION PLOT OF BRAY-CURTIS DISTANCE BETWEEN SAMPLES.....	161
FIGURE 15: LOCAL BIOGEOGRAPHIC PATTERNS.	163
FIGURE 16: <i>MAP OF THE STUDY AREAS IN THE SOUTH PACIFIC OCEAN</i>	187
FIGURE 17: PERCENTAGE OF SEQUENCES IDENTIFIED AS ARCHAEOA IN EACH SAMPLE, GROUPED BY TRENCH, ZONE, AND HORIZON DEPTH.....	192
FIGURE 18: ARCHAEOAL TAXONOMIC PROFILES AT CLASS LEVEL, GROUPED BY TRENCH, ZONE, AND HORIZON DEPTH. TAXONOMIC ASSIGNMENT WAS BASED ON SILVA 138.....	192

FIGURE 19: CO-OCCURRENCE NETWORK AND MODULARITY ANALYSIS.....	195
FIGURE 20: TAXONOMIC COMPOSITION OF THE MODULES AT CLASS LEVEL.	196
FIGURE 21: <i>VISUALISATION OF THE EDGES OF WEIGHT OVER 0.7 IN MODULES 3 (A) AND 4 (B).</i>	197
FIGURE 22: GLOBAL PHYLOGENETIC TREE OF AMO _A GENES OBTAINED FROM ALVES ET AL (2018), WITH PLACEMENT OF AMO _A GENES IDENTIFIED FROM CO-ASSEMBLIES AND EXTRACTED FROM REFERENCE MAGs.	220
FIGURE 23: COVERAGE OF AMO _A GENES IDENTIFIED IN CO-ASSEMBLIES (DETECTION OVER 0.9).	223
FIGURE 24: PHYLOGENOMIC TREE OF ORDER NITROSOSPHAERALES BUILT FROM THE GTDB DATABASE	226
FIGURE 25: MEAN COVERAGE OF NITROSOSPHAERIA MAGS IN OUR SAMPLES.	228
FIGURE 26: VIOLIN PLOTS SHOWING THE ESTIMATION OF THE NUMBER OF SNVs BY KBP AND THE RATIO OF SAAV TO SNV IN EACH SAMPLE FOR 11 MAGS FROM FOUR DIFFERENT AMO _A CLADES.	230
FIGURE 27: MEAN COVERAGE BY SAMPLE OF THE 90 ARCHAEL MAGS RECONSTRUCTED DURING THIS PROJECT.	237
FIGURE 28: SEASONALITY OF THE BACTERIOPLANKTON COMMUNITIES OF THE BAY OF BREST (LEMONNIER, 2019).	246

List of tables

TABLE 1: PRIMER SEQUENCES	113
TABLE 2: POTENTIAL COVERAGE OF PRIMER PAIRS DETERMINED USING TESTPRIME 1.0 AGAINST THE SILVA REFNR DATABASE V138, AT DOMAIN LEVEL, AND PHYLUM LEVEL FOR ARCHAEA.	119
TABLE 3: LIST OF SAMPLING STATIONS AND THEIR CHARACTERISTICS.....	147

List of supplementary figures

FIGURE S1: RAREFACTION CURVES FOR THE 2 METABARCODING DATASETS, COLORED BY SEDIMENT HORIZON.
134

FIGURE S2: HEATMAP OF THE RELATIVE PROPORTION OF EACH PHYLUM IN THE FOUR DATASETS: ARCHAEL
 METABARCODING DATASET, miTAGs ASSIGNED USING THE NBC IN DADA2, miTAGs ASSIGNED IN
 PHYLOFLASH AND UNIVERSAL METABARCODING DATASET.134

FIGURE S3: HEATMAP OF THE RELATIVE PROPORTION OF EACH TAXONOMIC CLASS IN THE FOUR DATASETS.
135

FIGURE S4: HEATMAP OF THE RELATIVE PROPORTION OF EACH TAXONOMIC ORDER IN THE FOUR DATASETS.
136

FIGURE S5: HEATMAP OF THE RELATIVE PROPORTION OF EACH TAXONOMIC FAMILY IN THE FOUR DATASETS.
138

FIGURE S6: HEATMAP OF THE RELATIVE PROPORTION OF EACH TAXONOMIC GENUS IN THE FOUR DATASETS.
138

FIGURE S7 : RAREFACTION CURVES FOR EACH SAMPLE IN THE METABARCODING DATASET172

FIGURE S8 : TAXONOMIC PROFILES AND ALPHA-DIVERSITY ESTIMATES.174

FIGURE S9: TAXONOMY OF THE 50 MOST ABUNDANT ALBORAN SEA BIOMARKER ASVs FOR (A) THE OVERALL
 SUBSURFACE BIOMARKER SET AND (B) THE SITE-SPECIFIC BIOMARKER SETS175

FIGURE S10 : VARIATION PARTITIONING ANALYSIS OF THE DATA USING BRAY-CURTIS DISSIMILARITY.176

FIGURE S11: EVOLUTION OF SEDIMENT CHARACTERISTICS IN THE THREE DEPTH ZONES TARGETED BY THE
 LONGITUDINAL SAMPLING SCHEME.177

FIGURE S12: ARCHAEL ALPHA DIVERSITY ESTIMATES WITH SHANNON INDEX ORGANIZED BY TRENCH, ZONE,
 AND HORIZON DEPTH.....204

FIGURE S13: ARCHAEL TAXONOMIC PROFILES AT CLASS LEVEL, GROUPED BY SAMPLING SITE, AND HORIZON
 DEPTH.....205

FIGURE S14: RELATIVE ABUNDANCE OF THE 50 MOST ABUNDANT NODES OF MODULE 1 IN ABYSSAL SITES (A9,
 A7, K7).....206

FIGURE S15: DISTRIBUTION PROFILES OF THE ARCHAEL MODULES IN SAMPLES ORGANIZED BY SITE AND
 INCREASING HORIZON DEPTH.....207

FIGURE S16: HEATMAPS OF COVERAGE AND SEQUENCE IDENTITY FOR *AMO*A CLADES NP-GAMMA-2.1 (A) AND NP-GAMMA-2.2 (B).232

FIGURE S17: EVOLUTION OF THE DENSITY OF SNVs IN A MAG WITH MEAN COVERAGE OF THIS MAG IN SAMPLES.233

List of supplementary tables

TABLE S1: VALUES OF LINEAR REGRESSION PARAMETERS COMPUTED FOR EACH SEDIMENT HORIZON ON THE WHOLE DATASET.178

TABLE S2: VALUES OF LINEAR REGRESSION PARAMETERS COMPUTED FOR THE ALBORÁN SEA SAMPLES FOR EACH SEDIMENT HORIZON.178

TABLE S3: METAGENOME INFORMATION: SEQUENCE NUMBER AND CO-ASSEMBLY GROUPS.285

TABLE S4: MAG INFORMATION : LENGTH, GC CONTENT, COMPLETION, REDUNDANCY, TAXONOMY287

List of abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
MAG	Metagenome-assembled genome
SCG	Single-copy core gene
SAG	Single-cell amplified genome
Mbsl	Meters below sea level
cmbsf	Centimeters below sea level
GTDB	Genome Taxonomy Database
RED	Relative evolutionary divergence
AOM	Anaerobic Oxidation of Methane

SMTZ	Sulfate Methane Transition Zone
SPG	South Pacific Gyre
MGI	Archaeal Marine group I
MCG	Miscellaneous Crenarchaeotal Group
TAR	Taxa-area relationship
DDR	Distance-decay relationship
SSU rRNA	Small subunit ribosomal RNA
COI	Cytochrome oxidase I
ASV	Amplicon sequence variant
OTU	Operational taxonomic unit
RDP	Ribosomal Database Project
NBC	Naive Bayesian classifier
<i>amoA</i>	Ammonia monooxygenase subunit A
AOA	Ammonia oxidizing Archaea
AOB	Ammonia oxidizing Bacteria
NGS	Next generation sequencing
Bp	Base pair
ORF	Open reading frame
SNV	Single nucleotide variant
SAAV	Single amino acid variant
GOS	Global Ocean Survey
DCO	Deep Carbon Observatory
IODP	International Ocean Drilling Project
OSD	Ocean Sampling Day
MGII	Archaeal Marine Group II
DHVEG-6	Deep Sea Hydrothermal Vent Group 6
PCR	Polymerase Chain Reaction

INTRODUCTION

INTRODUCTION

1. Studying the microbial ecology of the dark ocean: from the first expeditions to molecular ecology approaches

1.1. From the first expeditions of the 19th century to the large-scale explorations of the 21st

1.1.1. Reaching the deep sea

Covering 70% of the surface of the Earth, the ocean contains some of the largest habitats on the planet. However, its depths have long remained understudied. In 1842, Edward Forbes proposed the azoic hypothesis, postulating that biodiversity decreases with depth in the ocean and that life cannot exist deeper than 550 meters. However, from 1850 onward, dredging operations by Michael Sars in the Norwegian fjords recovered evidence of life, in particular crinoids, down to 820 meters. This prompted global exploration efforts, starting with the HMS *Challenger* expedition in the 1870s, led by Sir Charles Wyville Thomson.

The 20th century saw the development of numerous crucial instruments for the study of the deep ocean, including sonars to measure ocean depth and detect underwater objects, and submersibles. Between 1930 and 1934, Otis Barton and William Beebe conducted the first manned dives with the intent of observing deep-sea animals in their environment, off the coast of Bermuda, and reached 923 meters. Thirty years later in 1960, Jacques Piccard and Don Walsh were the first to reach the bottom of the Mariana Trench at 10,911 meters. Building on these technological advances, submersible vehicles were developed, and led to the discovery in 1977 of the first hydrothermal vent and its unexpected oasis of life along the Galapagos Ridge (Corliss et al., 1979). These exciting discoveries paved the way for further studies of the biodiversity found at the bottom of the ocean, with the development of technical gear capable of reaching the deepest parts of the ocean now encompassing remotely operated underwater vehicles (ROVs) and independent free-falling systems, or landers (Jamieson et al., 2018).

INTRODUCTION

1.1.2. Advent of molecular ecology and sequencing

In parallel with the technological advances that made it possible to reach high depths in the ocean, a number of molecular ecology discoveries came to be very significant for deep sea microbiology studies, first and foremost the resolution of the double helix structure of DNA (Watson and Crick, 1953) and the ensuing “central dogma” of molecular biology, linking information with DNA sequence (Crick, 1958). Further studies based on DNA properties revolutionized the taxonomic classification of microorganisms, with Woese and others starting to use the 16S small subunit ribosomal RNA gene to systematically classify prokaryotes (Fox et al., 1977).

The Sanger sequencing method first proposed by Sanger et al. (1977) soon became a widely used technique to obtain DNA sequences from microorganisms, especially after the refining of the Polymerase Chain Reaction (PCR) (Mullis et al., 1986). Necessitating a PCR amplification and cloning step before sequencing, this fastidious method allowed access to around 100 sequences by sample, thus predominantly grasping the dominant taxa. It was used to produce the first complete genome sequence of *Haemophilus influenzae* (Fleischmann et al., 1995), and fueled the efforts of the Human Genome Project between 1990 and 2003 (International Human Genome Sequencing Consortium, 2001).

The second generation of sequencing techniques, also called next-generation sequencing (NGS), were developed in the 1990s and early 2000s and had a huge impact on the field of microbial ecology (van Dijk et al., 2014). Still relying on neo-synthesis of DNA molecules, these techniques replaced biological cloning in *E. Coli* by physical cloning on microbeads in the case of pyrosequencing, and nano-sized clusters on flow cells for the Illumina technology. Sequencing capacity increased rapidly, and the cost of sequencing fell (Fig. 1), generating an increasing subsampling and coverage of prokaryotic communities through the sequencing of environmental DNA.

In 1985, Staley and Konopka had estimated that less than 1% of microorganisms in the environment were potentially cultivable. With Illumina sequencers allowing access to a wider

INTRODUCTION

than ever extent of the microbial diversity of an environmental sample, progresses in microbial ecology were strongly linked with this increase in affordability and rate of sequencing.

More recently, a third generation of sequencing methods emerged, characterized by their capacity to produce very long sequences from single molecules of DNA, eliminating the amplification step. However, read quality and sequencing depth are often poor, and this type of data requires specific bioinformatic error correction steps or association with short NGS reads (Weirather et al., 2017).

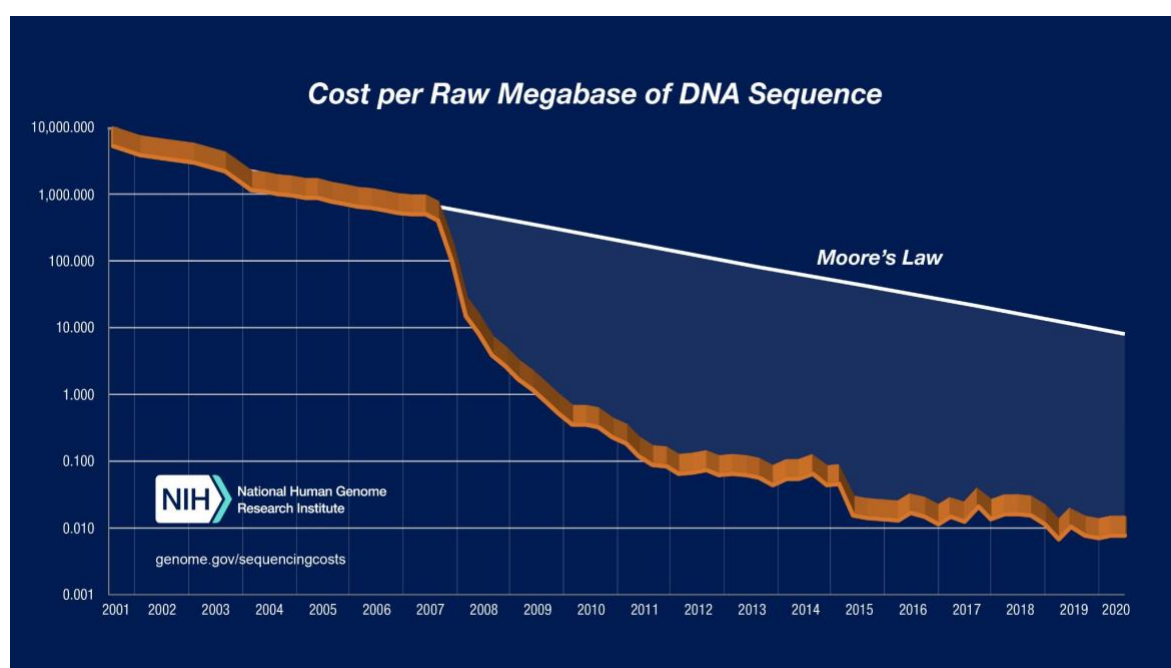


Figure 1: Sequencing cost per megabase of DNA - August 2020. Wetterstand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 06/05/2021.

1.1.3. Microbial ecology at the global scale

The advent of the sequencing techniques presented above opened the way for the inventory of diverse microbiomes such as the human microbiome (Turnbaugh et al., 2007) or soil microbiome (Fierer and Jackson, 2006). In the marine realm, several large-scale projects were developed to target specific biomes: Ocean Sampling Day (OSD) focusing on coastal

INTRODUCTION

environments, the Global Ocean Survey (GOS) by the Craig Venter Institute and Tara Oceans expeditions sampling mostly the euphotic zone around the world. Deep sea sediments, and in particular subsurface sediments, were also the object of international collaboration projects, e.g. the Deep Carbon Observatory (DCO) and the International Ocean Drilling Project (IODP). All of these projects relied at least in part on two applications of high-throughput sequencing: the sequencing of specific marker genes or amplicons, also called metabarcoding, and whole genome sequencing.

1.2. Molecular tools to access the functional and taxonomic diversity of microbial communities

1.2.1. Metabarcoding

1.2.1.1. Marker gene selection

Metabarcoding is a very widely used technique in microbial ecology to characterize taxonomic diversity. It relies on the PCR amplification and subsequent sequencing of a marker gene, that can then be compared to a reference database to phylogenetically place the organism.

To effectively achieve this goal, the ideal marker gene must be stable enough to be present in all microorganisms, but presenting enough variation from one organism to the next to discriminate between them at the species level.

The most commonly used marker gene when working with Bacteria and Archaea is the 16S rRNA gene. Because of its role in the translation process, it is ubiquitous. Its function depending on its structure, it presents some very conserved regions with a slow evolution rate that can be used to design PCR primers, allowing optimized recognition and capture. Conversely, it also presents nine hypervariable regions where less selection pressure is applied, that are used to discriminate between taxa.

INTRODUCTION

The full sequence of the 16S rRNA gene is about 1500 bp long. Sequencing of the full gene would permit a species-level resolution of the composition of a community. However, the widely used next-generation sequencing technologies currently offer amplicon lengths of a few hundred base pairs. As a result, the targeted sequences to be amplified in the 16S rRNA gene usually encompass one or two hypervariable regions between V1 and V9.

The choice of which region to amplify depends on the desired outcome. No single region can differentiate among all Bacteria and Archaea, nor do all of the hypervariable regions have the same resolution power (Fuks et al., 2018; Willis et al., 2019; García-López et al., 2020).

1.2.1.2. Delineating ecologically informative units

The next challenging step is the bioinformatic analysis of the large datasets produced, to account for sequence errors and intra-specific diversity, through the denoising and/or clustering of sequences into ecologically relevant units. The first definition of microbial species from a molecular point of view was based on DNA/DNA hybridization experiments, where a threshold of 70% reassociation of the single-stranded DNA molecules sorted two microorganisms as the same species (Wayne et al., 1987). When 16S approaches became more widespread, a threshold of 97% sequence similarity was proposed to delineate species (Stackebrandt and Goebel, 1994). This threshold takes into account intra-species variability and possible sequencing errors introduced, since they are difficult to differentiate from actual variants. Indeed, microbial strains whose genomes contain multiple copies of the 16S gene not coalesced through concerted evolution have been shown to possibly lead to an overestimation of the community's diversity (Kang et al., 2010; Pei et al., 2010). However, studies have questioned the 97% threshold and its representation of true diversity, putting forward a value of 99% instead (Kim et al., 2014; Edgar, 2018).

In recent years, new algorithms have been developed to access fine-scale variations that could be masked by clustering approaches. These high-resolution methods are based on

INTRODUCTION

information entropy analysis (oligotyping and MED, Eren et al., 2013a, 2014), sequence identity and distribution profiles (SWARM, Mahé et al., 2015) or sequencing error correction (DADA2, Callahan et al., 2016)

More particularly, DADA2's error correction algorithm is based on the assumption that sequencing errors are randomly distributed, while actual variants are not. Ultimately, after the correction step, also termed denoising, DADA2 yields a collection of unique sequences, or Amplicon Sequence Variants (ASVs), instead of clusters of closely related sequences, i.e. Operational Taxonomic Units (OTUs). In addition to resolving fine-scale variation, the denoising process and generation of ASV-level observation tables increases the comparability of datasets (Callahan et al., 2017), and depending on the aim of the study, clustering algorithms could still be applied after such error correction/denoising.

1.2.1.3. Limitations and biases

Primer pairs used to amplify the chosen region of the 16S rRNA gene often have more affinity with certain groups of microorganisms. Indeed, studies have highlighted the bias induced by a single nucleotide mismatch in a primer sequence (Bru et al., 2008; Eloë-Fadrosh et al., 2016). Archaea in particular have been shown to be underestimated when using universal primers (Baker et al., 2003; Klindworth et al., 2013). It thus appears necessary to adapt the primer set used to the study's aim, though this might result in an impossibility to compare datasets generated through different methods.

Taxonomic affiliation of 16S sequences relies on comparison with pre-existing databases. Three main databases are available for 16S rRNA: Greengenes (McDonald et al., 2012), the Ribosomal database project (RDP, Wang et al., 2007) and Silva (Quast et al., 2013). The last one is the largest, with manually curated taxonomic rank assignment. It is only as complete

INTRODUCTION

as researchers have been able to make it over the years however, and deep sea sediments are one of the least characterized environments, suggesting significant gaps can be expected.

Finally, marker gene approaches do not reflect the whole genome content, and are thus not adapted to the study of functional diversity or microbial activity. Because of these limitations, recent studies in microbial ecology tend to take advantage of increasing sequencing power and decreasing associated costs to turn to metagenomics and whole genome sequencing.

1.2.2 Whole genome sequencing

1.2.2.1. Read-centric analysis

Metagenomic data can be analyzed in a read-centric fashion, based on the raw reads obtained from sequencing. For example, MG-RAST (Meyer et al., 2008) is a web-available tool that computes phylogenetic and functional summaries of metagenomes based on the comparison of short reads to databases. Phyloflash (Gruber-Vodicka et al., 2020) and SortMeRNA (Kopylova et al., 2012) are other examples of read-centric tools that identify and extract sequences affiliated to small-subunit rRNA genes to produce a taxonomic profile of the sample. Beta-diversity can also be investigated *de novo*, by computing distances between samples based on k-mer profiles, such as with Simka (Benoit et al., 2016). This can be interesting to avoid the challenges and biases inherent to the assembly step (see below).

1.2.2.2. Assembly

NGS produces short reads (around 150 bp for Illumina HiSeq/NextSeq). As a consequence, reconstruction of genes and genomes start with an assembly step, with tools such as Metaspades (Nurk et al., 2017), Megahit (Li et al., 2015) or IDBA-UD (Peng et al., 2012). Ideally, the aim is to reconstruct full genomes, however *de novo* assembly of a metagenome

INTRODUCTION

is challenging. For instance, sequences with high conservation between genomes break the assembly graph and low abundance organisms may not be sequenced with a depth sufficient to be re-assembled (Lapidus and Korobeynikov, 2021). As a result, getting a complete genome from a metagenome simply by assembling reads is fairly impossible with the current softwares, but assembly is useful in order to generate contigs, i.e. sequences longer than the reads first obtained, that make further analyses possible. To help with the recovery of low abundance organisms, coassembly can also be performed by merging sequences obtained from different samples.

1.2.2.3. Binning

Binning is one of the possible steps following assembly. It aims at separating contigs into discrete clusters called bins, representing putative genomes. Several automatic binning algorithms have been developed in recent years such as MaxBin2 (Wu et al., 2016), MetaBAT (Kang et al., 2015, 2019), CONCOCT (Aneberg et al., 2014) or Binsanity (Graham et al., 2017). Clustering of the sequences is based on sequence composition through tetranucleotide frequency (or k-mer profiles), and differential coverage.

Indeed, genomes have specific k-mer profiles, and k-mer composition is stable inside a genome while it varies among microorganisms. On the other hand, coverage, the number of times a sequence is found in the data, is assumed to reflect the abundance of organisms in the sample, which justifies the use of coverage profiles to define clusters of sequences.

However, automatic binning can result in conflation errors (more than one genome in a bin) or fragmentation errors (genome separated into multiple bins). This is assessed by detecting a set of single-copy core genes, previously determined from comparative analysis of database genomes (Darling et al., 2014; Lee, 2019). These genes are expected to be present only once in each microorganism, and thus are used to estimate the completion and redundancy of the bins (CheckM, Parks et al., 2015).

INTRODUCTION

Recently, automatic methods to refine bins generated by one of the softwares mentioned above have been proposed. DAS Tool (Sieber et al., 2018) compares the results of multiple binning softwares based on marker genes, and refines them by combination or rejection. GraphBin (Mallawaarachchi et al., 2020) uses the de Bruijn graph inherited from the assembly step to refine the proposed bins and identify mis-binned contigs. Finally, manual curation of the binning results using visualization softwares such as Anvi'o (Eren et al., 2015) can be applied to improve bin quality.

The refined bins can be called MAGs for Metagenome-Assembled Genomes. The Genomic Standards Consortium proposed guidelines for reporting new genome sequences, in order to facilitate more robust genomic analyses (The Genome Standards Consortium et al., 2017). In particular, they defined MAG quality thresholds based on the minimum information available. According to these standards, MAGs are considered to be high-quality drafts when they are over 90% complete and less than 5% redundant, with presence of the 23S, 16S and 5S rRNA genes and at least 18 tRNAs. When MAG completion is over 50% and contamination under 10%, it is classified as a medium-quality draft.

1.2.2.4. Comparative genomics

Once the MAGs have been reconstructed, further analysis can include a reassembly step to attempt “closing” the genomes.

If contigs have been scanned for Open Reading Frames (ORFs) and genes have been identified, functional annotation of these genes can be attempted with databases such as KEGG (Kanehisa and Goto, 2000) or COG (Tatusov, 2000) to infer the metabolic capabilities of the microorganism.

Comparison of the sequences of closely related MAGs can also yield insights into the core and accessory parts of the putative genome. Corresponding to genes found in all the bins or taxa, the core genome can be hypothesized as being essential to the organisms' survival in

INTRODUCTION

their environment. Conversely, the accessory genome is differentially present among populations, and possibly makes them specifically adapted to a condition or a metabolite usage. However, in poorly characterized environments such as deep-sea sediments, databases lack references for many genes, and functional annotation remains difficult.

2. Life in deep ocean surface sediments

2.1. Geologic and biogeographic definition of seafloor provinces

The ocean can be broadly divided in two realms: the pelagic and the benthic. The pelagic realm refers to the water column, and the benthic zone is composed of the first few layers of seafloor sediments. The interface of these two zones is called the benthic boundary layer, and includes the bottom water and sediment layer directly in contact with it.

2.1.1 Geologic classification

Seafloor sediments can be categorized based on their proximity to tectonic plates or land boundaries. The most studied sediments are those of the continental margin, which account for about 20% of the total aerial coverage (Orcutt et al., 2011).

These margins are split between passive continental margins, which do not coincide with a tectonic plate margin (e.g. Atlantic Ocean coasts), and active margins, which fall along a tectonic plate boundary where subduction occurs, like on the Western coast of South America. Passive continental margins usually exhibit a bathymetric profile with three subdivisions: first the continental shelf, where the depth of the water does not exceed 200m, then the continental slope, where the seafloor plunges from 200m below the surface to 1,000 to 2,000m below the surface (Fig. 2). In contrast with this, active margins seafloor drops much more abruptly to

INTRODUCTION

great depths in the subduction, or hadal, trenches. Most of the seafloor underlying the open ocean is a flat expanse called the abyssal plain, which has an average depth of around 4000m (Orcutt et al., 2011) and accounts for as much as 79% of the seafloor worldwide (Schrenk et al., 2010). The sedimentation rate over the abyssal plain is generally much lower than the rate closer to shore, over the continental margins that also benefit from terrestrial input and possibly upwelling processes, and thus the sediments, on average, are thinner.

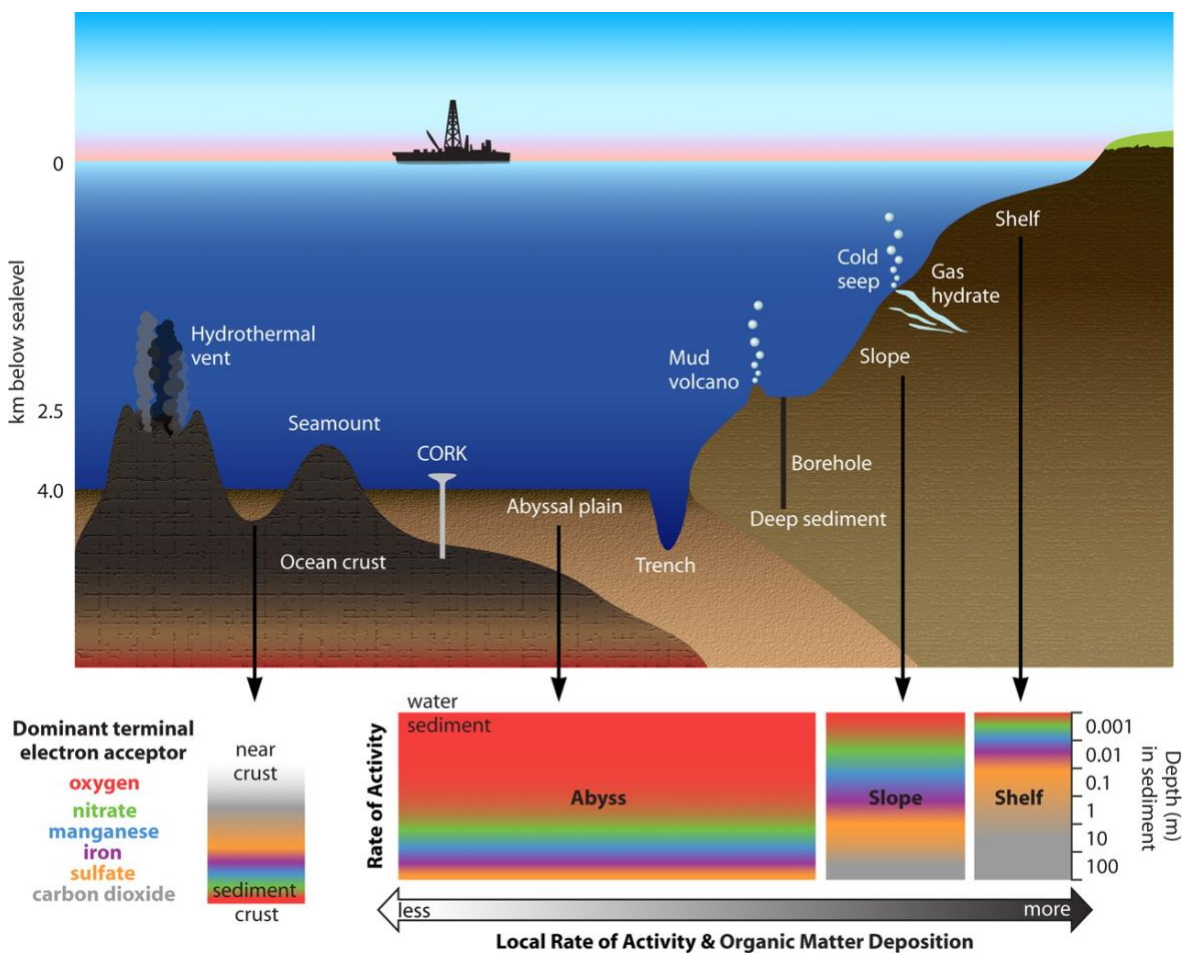


Figure 2: Schematic view of a stylized cross section of dark ocean habitats (top) and sediment zonation (bottom). NB: the upper panel is not drawn to scale, and the scale for the bottom one is logarithmic. (Orcutt et al., 2011)

INTRODUCTION

2.1.2. Biogeographic classification

Biologists have traditionally delineated seafloor regions based on environmental characteristics, most importantly ocean depth (Watling et al., 2013; Costello et al., 2017).

The first region, like the continental shelf, extends down to 200m and is found below the photic zone. The mesopelagic or “twilight” zone, sometimes also called the upper bathyal zone, extends from the end of the photic zone down to around 1000 m. Finally, the lower bathyal (1000 - 3500/4000 mbsl), abyssal (~4000 - 6000 mbsl) and hadal zones (> 6000 mbsl) compose the rest of the seafloor regions. These last three zones are characterized by low environmental variation, no light penetration, low temperatures and high pressure, and the existence of an ecological delineation between lower bathyal and abyssal regions in particular has been questioned (Costello and Breyer, 2017).

2.1.3. Hadal trenches

Hadal trenches are formed by the subduction of one tectonic plate under another and thereby are much longer than they are wide. They are the deepest places on Earth, the Challenger Deep in the Mariana Trench reaching a depth of 10 898 m.

Depending on the precise definition used, the number of hadal environments varies. Jamieson (2015) defined trenches as distinct, single elongated areas deeper than 6500 m generally formed by subduction or faulting, and troughs as large areas deeper than 6500 m which are not formed at converging plate boundaries. Based on these definitions, the author identified 33 hadal trenches and 13 troughs. Still, they represent less than 1% of the seafloor surface habitat (Jamieson, 2015).

In the trenches, hydrostatic pressure increases linearly with depth, and the typical temperature range is 1.0 - 2.5°C. This temperature is comparable to the ones measured on the continental margins around 3000 mbsl, due to the adiabatic heating created by the pressure increase

INTRODUCTION

(Jamieson et al., 2010). Salinity at the bottom of the trenches remains around 34-35 ppm, and overall, with the exception of hydrostatic pressure, the physical characteristics of the trenches do not differ sharply from those of the abyssal zone.

2.1.3.1. Atacama and Kermadec trenches

Part of this manuscript will be dedicated to the study of microbial communities found in the sediments of the Atacama and Kermadec trenches. These trenches will be shortly introduced here.

The Peru-Chile trench system extends for about 5900 km with a width of 100km off the coast of Chile (Fig. 3B). The Atacama trench is one of the deepest parts of this system, reaching over 8000m in depth, and is located between 20° and 30°S, about 160 km away from the coast (Danovaro et al., 2003).

The Kermadec trench is located in the southwestern part of the Pacific Ocean, north-east of New Zealand's northern island (Fig. 3A). It forms a near linear system with the Tonga trench to the north, from which it is separated by the Louisville Seamount Ridge (Jamieson, 2015). It extends for about 1000 km between 26° and 37°S, and is one of the deepest trenches on Earth, reaching depths of more than 10 000m (Ballance et al., 1999).

INTRODUCTION

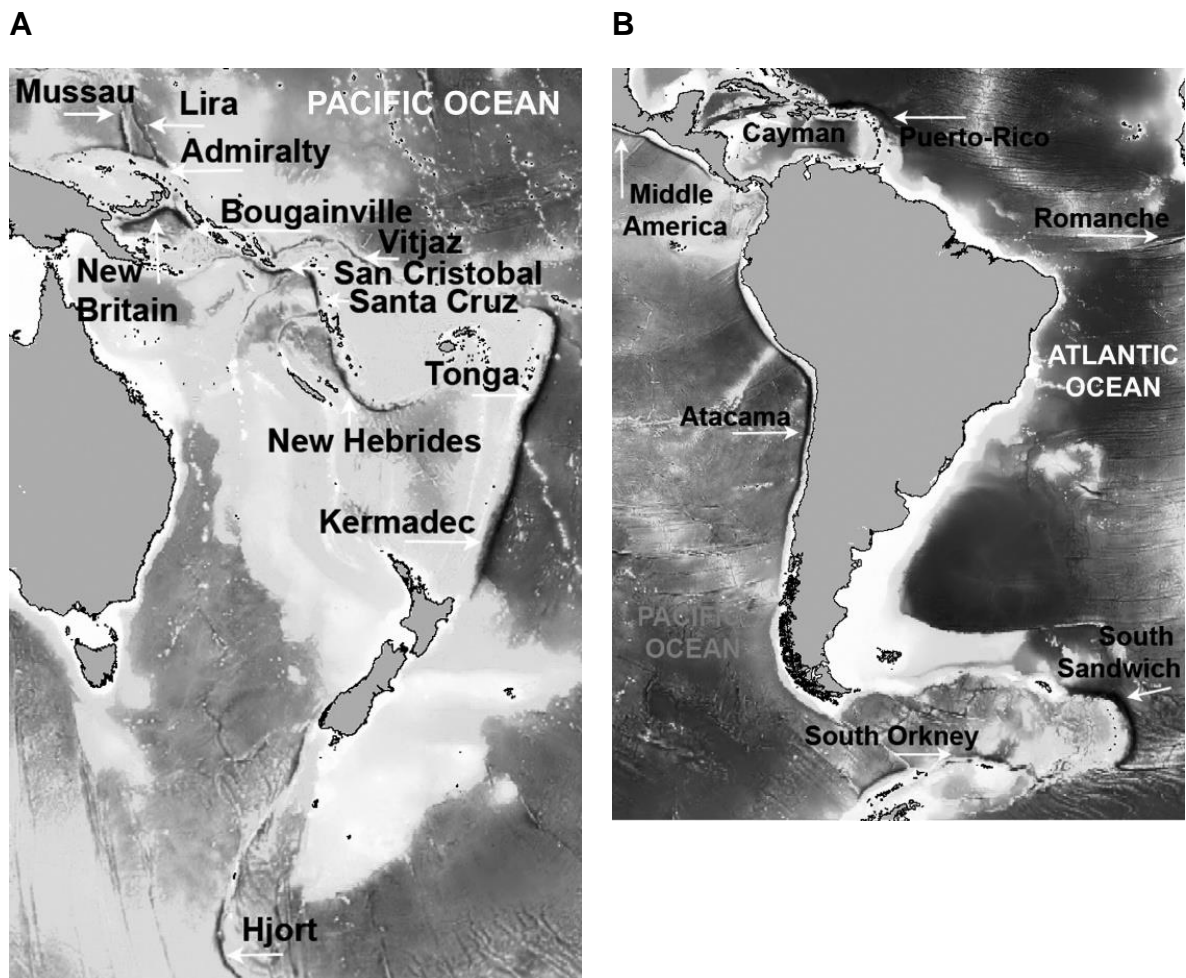


Figure 3: Geographic location of the hadal trenches of A) the southwest Pacific Ocean and B) the southeast Pacific, Atlantic and Southern Oceans. (Jamieson, 2015)

2.2. Metabolic reactions in deep sea sediments

2.2.1. General diagenetic sequence

The physico-chemical characteristics of deep sea sediments can vary greatly from one location to the next, with the thickness of the sediment layer varying from relatively thin near newly formed crust at mid ocean ridges and under low-productivity zones of the ocean, to

INTRODUCTION

1000s of m thick at highly productive continental margins (Jahnke, 1996; Divins, 2003; Olson et al., 2016).

The composition of sediments depends on the origin of the particulate matter that lands on the seafloor, be it organic matter sinking from the ocean surface, terrestrial particles carried by rivers to the open ocean, hydrothermal effluents, or wind-blown particles from land. Due to this, the sediment particle size will also vary, which will affect transport of fluids and chemical compounds inside the sediment layers. The transport of chemical compounds in marine sediments mostly occurs via molecular diffusion against chemical gradients, which limits their availability to microorganisms compared to more actively moved sites or matrices. In surface sediments, bioirrigation and bioturbation by macrofauna enhance oxygen exchanges between sediments and the overlying water column (Pischedda et al., 2008).

Generally, there is a discrimination between surficial sediments and deep or subsurface sediments. Shallow sediments usually exhibit strong geochemical gradients and higher rates of microbial activity while deep sediments display lower cell densities and more stable gradients (Parkes et al., 1994; Wellsbury et al., 1997; D'Hondt et al., 2004). In both cases, microorganisms rely on chemotrophy, obtaining energy from chemical redox reactions.

The most important substrate in most marine sediments is organic matter that is remineralized by oxidation back to carbon dioxide. The rates of remineralization depend on the quantity and quality of organic matter (its freshness, its origin, etc), and the availability of terminal electron acceptors (Hedges et al., 1988; Canfield, 1994; Niggemann et al., 2007). Utilization of electron acceptors tends to go from highest redox potential to lowest, following thermodynamic energy yield (Froelich et al., 1979). Given that the availability of reactive compounds is limited by deposition and diffusion, this results in a vertical redox stratification. This pattern of redox processes occurring with depth in sediments is referred to as the diagenetic sequence (Fig 4 extracted from Parkes et al., 2014).

INTRODUCTION

The electron acceptor with highest redox potential is oxygen, produced in the photic zone through photosynthesis then transported to depth via ocean mixing. It is used in the first layers of sediments, the oxic zone, to respire organic matter (Froelich et al., 1979). This results in production of ammonium through ammonification, which is partly aerobically oxidized by nitrification, leading to the formation of nitrate. All of these compounds partially diffuse to the bottom waters.

In sediments where oxygen is depleted, nitrate becomes the most favorable terminal electron acceptor, and is involved in the processes of denitrification (nitrate reduction) and anaerobic ammonium oxidation (anammox) after reduction to nitrite (Thamdrup, 2012). Both of these processes end with production of dinitrogen and are thus sinks of nitrogen. This nitrate reduction zone just below the oxic layers is sometimes referred to as the nitrogenous zone (Canfield and Thamdrup, 2009).

The next step in the diagenetic sequence is the reduction of metal oxides, more particularly manganese and iron compounds, in the next layers of the suboxic zone, proposed to be more precisely renamed to the manganous and ferruginous zones (Canfield and Thamdrup, 2009). Then comes the anaerobic sulfate reduction zone, or sulfidic zone, and finally the methanogenic zone (Fig 4).

In the methanogenic zone, carbon dioxide, the final oxidized carbon product of organic matter degradation and remineralization, is consumed during methanogenesis and other autotrophic metabolisms such as acetogenesis (Wellsbury et al., 2002; Newberry et al., 2004; Parkes et al., 2005; Hinrichs et al., 2006; Lever et al., 2010). Marine sediments are the largest reservoir of methane on Earth (Kvenvolden, 1993), with a majority of marine methane being estimated to come from reduction of CO₂, carbon monoxide, acetate or formate by methanogenic archaea (Orcutt et al., 2011).

Methane diffuses upward from the methanogenic zone, and is usually oxidized with sulfate because it is the first electron acceptor that becomes available. Sulfate-dependent methane oxidation happens in defined anoxic zones in marine sediments called sulfate-methane

INTRODUCTION

transition zones (SMTZ) and is one of the processes referred to as anaerobic methane oxidation (AOM) (Devol et al., 1984; Iversen and Jorgensen, 1985; Hinrichs and Boetius, 2002).

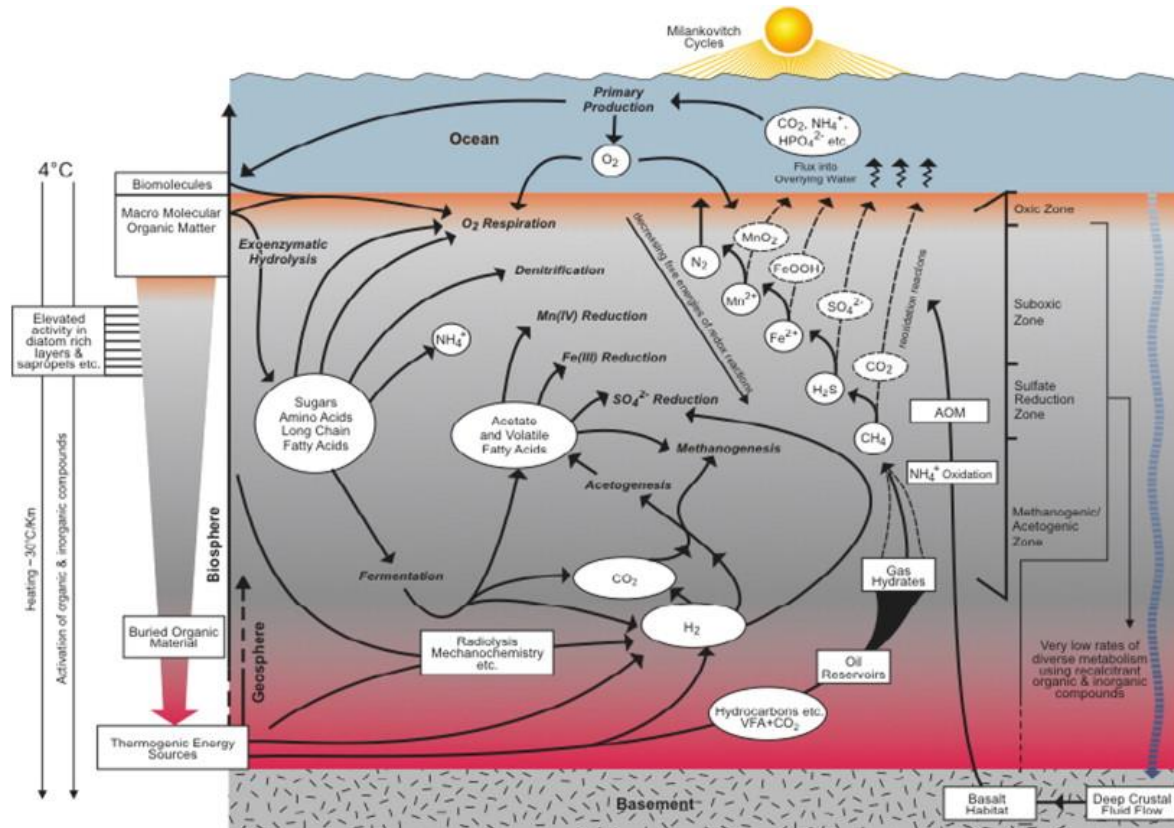


Figure 4: Schematic illustration of the metabolic processes taking place with depth in seafloor sediments. (Parkes et al., 2014)

2.2.2. Organic matter input and oxygen penetration depth

Input of organic matter to seafloor sediments comes mainly from sinking particulate matter generated in surface waters through primary production. During its transport through the water column, it is altered and consumed by microbes, resulting in a lower percentage of labile organic matter reaching the deep seafloor compared to shallower environments (Arndt et al., 2013).

Typically, the open ocean overlying abyssal plains displays low primary production in the surface waters, which, in combination with the great oceanic depth, leads to a low supply of

INTRODUCTION

organic matter to the sediments causing low cell densities and low rates of carbon remineralization. This means that the oxygen present in the sediments will be consumed slowly and diffuse deeper, staying the dominant terminal electron acceptor until deeper layers (Glud, 2008). For example, the open ocean gyres such as the South Pacific Gyre (SPG) seem to be particularly carbon- and nutrient-starved (D'Hondt et al., 2009).

Conversely, in highly active sediments where organic matter is not limiting, the rates of remineralization depend on the availability of terminal electron acceptors that are consumed faster than they are supplied, so burial of a larger percentage of the organic matter is observed. In this case, oxygen is depleted quickly and disappears after the first few millimeters/centimeters (reviewed in Glud, 2008; Wenzhöfer et al., 2016).

For these reasons, the oxygen consumption rate is often used as a measure of biological activity and benthic carbon mineralization in sediment layers (Nøhr Glud et al., 1994).

2.2.3. Oxygen penetration depth in hadal and adjoining abyssal sediments of the Atacama and Kermadec trenches

In addition to geographic location, hadal trenches can be distinguished by the surface primary production of the water mass they are found under. The Atacama trench presented above is situated in an area characterized by important upwelling events leading to very high primary production levels (Fossing et al., 1995). The Kermadec trench region experiences lower surface primary production levels, i.e. ~400 mg C.m⁻².d⁻¹ against ~900 mg C.m⁻².d⁻¹ in Atacama in Glud et al.'s study (2021) corresponding to the sampling cruises presented in this manuscript. The Atacama trench system is also located near the Chilean coastline and subjected to transfer of material from the adjacent continental desert by winds. Measurements of functional chlorophyll-a, phytodetritus and labile organic carbon deposited on its seafloor sediments were of concentrations similar to those of highly productive shallow coastal areas (Danovaro et al., 2003).

INTRODUCTION

There are major differences in organic matter input as well between hadal trenches and abyssal plains. Firstly, the specific topography of trenches generates a funneling effect that concentrates organic matter along the trench axis (Turnewitsch et al., 2014). Additionally, the seismic activity associated with tectonic subduction zones can lead to the relocation of important quantities of organic-rich surface sediments to the bottom of the trench through mass-wasting events (Kioka et al., 2019).

These contrasting characteristics are reflected in *in-situ* measures of oxygen penetration depth and uptake (Glud et al., 2021). While there are variations in activity along the trench axis, benthic oxygen consumption is always higher inside the trench compared to the adjacent abyssal plains. Differences in surface production are also echoed in the higher oxygen consumption rates of the Atacama region compared to the Kermadec region, and of the abyssal landward site compared to the “open ocean” abyssal site near the Atacama trench.

2.3. Microbial diversity in benthic sediments

2.3.1. Microbial cell counts

Microbial cell abundance in seafloor sediment varies between sites up to 5 orders of magnitude. This variation is strongly correlated with sedimentation rate and distance from land (Kallmeyer et al., 2012), highlighting the link between microbial biomass and organic carbon input from vertical sources and lateral transport. Global abundance of Bacteria and Archaea in seafloor sediments has been estimated to be on the order of 4.9×10^{28} cells in the benthic layer (top 50 cm) and 2.9×10^{29} globally (Kallmeyer et al., 2012; Danovaro et al., 2015).

In the sediment column, the abundance of cells decreases logarithmically with increasing sediment depth (Parkes et al., 2014). The cells appear to be alive and metabolically active, but with very widely varying activity rates that are much lower than anything known through cultivation (Schippers et al., 2005; Teske, 2005). In addition to this decrease in cell counts, a

INTRODUCTION

general decrease in microbial richness with sediment depth has been observed (Petro et al., 2017). It is more pronounced for Bacteria than Archaea, possibly because the latter are more adapted to the energy-limited conditions encountered in the deep seafloor. It seems clear that Archaea play a prominent role in subsurface communities, and it has been suggested that they even come to dominate deep oxic sediment communities of the lower bathyal and abyssal zones (Biddle et al., 2006; Lipp et al., 2008; Vuillemin et al., 2019).

Microbial cell counts in hadal trench surface sediments have highlighted higher abundances than in the adjacent abyssal sites, and interestingly no logarithmic decrease with sediment depth. Instead, a number of subsurface peaks were observed, closely correlated with fluctuations in total organic carbon content (Hiraoka et al., 2020; Schaubberger et al., 2021a).

2.3.2. Community composition

Microbial communities of benthic sediments have been shown to be distinct from those of the water column as well as subseafloor (Zinger et al., 2011; Bienhold et al., 2016; Walsh et al., 2016). There is also a strong distinction between communities found in oxic and anoxic sediments (Hoshino et al., 2020), as well as a stratification of community diversity with sediment depth (Durbin and Teske, 2011; Hiraoka et al., 2020; Schaubberger et al., 2021b), probably related to the electron acceptor cascade described before.

As a result, large-scale trends in community composition at higher taxonomic level can be observed. Oxic surficial sediment communities are mostly composed of bacterial lineages Alpha-, Gamma-, and Deltaproteobacteria (Desulfobacterota in SILVA 138), Acidobacteria, and Actinobacteria (Durbin and Teske, 2011; Orcutt et al., 2011; Bienhold et al., 2016). Archaeal communities in these first layers are dominated by Thaumarchaeota (previously Marine Group I; phylum Crenarchaeota, class Nitrososphaeria in SILVA 138).

Some of these groups, notably Alpha-, Gammaproteobacteria and Thaumarchaeota, are also characteristic of deep water communities, and decrease in relative abundance with sediment depth (Bienhold et al., 2016; Peoples et al., 2018; Lloyd et al., 2020). Instead, deeper in the sediments, Chloroflexi, Planctomycetes and Atribacteria (OP9/JS1 group) become prominent

INTRODUCTION

lineages (Durbin and Teske, 2011; Bienhold et al., 2016; Vuillemin et al., 2020). A number of uncultivated archaeal groups also become more abundant in deeper anoxic sediments, including candidate phylum Bathyarchaeota (previously Miscellaneous Crenarchaeotic Group MCG), Euryarchaeota and members of superphyla Asgard and DPANN, in particular candidate phylum Woesearchaeota (Peoples et al., 2019; Hiraoka et al., 2020; Hoshino et al., 2020; Schaubberger et al., 2021b).

With the development of high-throughput sequencing methods, there has been an explosion in the study of Archaea, leading to the description of several new phyla and super-phyla (reviewed in Adam et al., 2017). Conversely, the number of isolates and newly described archaeal species increased steadily but slowly, due to the specific challenges faced by culturing approaches, for example very slow growth (Imachi et al., 2020). Naming and classification of the ever-widening diversity of archaeal lineages over the years outside the traditional framework and guidelines of cultured type strains has resulted in some confusion. Recent efforts towards a standardized definition of the phylogeny and taxonomy of archaeal taxa based on genomic data have been started and will be described in the last part of the introduction for reference in the rest of the manuscript.

2.3.3. Diversity in surface sediments of hadal trenches

In benthic sediments of hadal trenches, phylum level patterns of diversity are overall the same as described above. At this resolution, the influence of the high hydrostatic pressure characteristic of hadal depths on community structure is unclear. However, zonation by sediment depth and oceanographic realm (i.e. abyssal or hadal) is observable and reflects geochemical conditions (sedimentation rate and redox stratification) (Hiraoka et al., 2020; Schaubberger et al., 2021b).

Peoples et al. (2019) and Schaubberger et al. (2021b) compared the degree of endemism between the Mariana and Kermadec trench and the Atacama and Kermadec trench

INTRODUCTION

respectively. They both showed connectivity between the trenches, even at ASV level, emphasizing the importance of oceanic depth over limitations on microbial dispersal and differences in surface primary productivity.

2.4. Biogeographic patterns of the abyssal plains

2.4.1. General biogeographic patterns and processes

Biogeography is the study of the distribution of biodiversity over space and time: what microorganisms can be found where and at what abundance, but also what processes shape the patterns that can be observed (MacArthur and Wilson, 1967; Martiny et al., 2006). It depends on the observation that spatial distribution of microorganisms across habitat types is non-random (Cho and Tiedje, 2000; Oda et al., 2003).

Depending on the distribution of a microorganism, it can be categorized as ubiquitous, or cosmopolitan, if it is widespread. Conversely, endemic species will have uneven distributions, restricted to a particular location, region or habitat type. Allopatric species are species occurring in separate and non-overlapping areas, while sympatric species share the same environment.

At the ecological level, biogeographic patterns result from four main processes: selection, diversification, dispersal, and drift (Vellend, 2010; Hanson et al., 2012; Nemergut et al., 2013). Selection refers to the alteration of species relative abundance based on fitness differences (ability to survive, grow and reproduce). This process is deterministic and can be influenced by environmental biotic (e.g. competition, mutualism...) and abiotic factors (temperature, pH, salinity...). The stochastic counterpart to selection is drift, the variation in species frequencies owing to chance demographic fluctuations.

Diversification (or speciation in Vellend's framework) is the introduction of new genetic variation. Widely accepted as a stochastic process, its effects are difficult to detect and can

INTRODUCTION

be intertwined with the results of selection (Vellend et al., 2014; Mittelbach and Schemske, 2015).

Finally, dispersal is defined as the movement of organisms across space, to which is sometimes added the successful establishment of a species in a new location (Hanson et al., 2012). For microorganisms, this successful colonization is difficult to establish since it would necessitate detection of metabolic activity. Thus, simple detection of the microorganism is usually considered synonymous to establishment.

In the case of microorganisms, dispersal is considered to be mostly passive, and thus to be a largely stochastic process (Zhou and Ning, 2017). However, some factors may influence dispersal efficiency for specific taxa. Unlike for macroorganisms, size does not seem to correlate with dispersal efficiency of microorganisms, but population density is an important parameter (Martiny et al., 2006). A high abundance will mean that more propagules, the smallest unit of dispersal necessary for microorganisms to colonize a new location, can be transported, leading to a theoretically better chance to reach further habitats. Other factors can impact the passive dispersal of microorganisms, such as habitat type, free water microorganisms for example being more mobile than sediment communities, physical barriers such as ridges, and the ability to form spores that increases organism resilience to transport (Locey et al., 2020). In marine sediments, dormancy and endospore formation are common (Wörmer et al., 2019).

Two important relationships usually examined are the taxa-area relationship (TAR) and the distance-decay relationship (DDR). The TAR postulates that the bigger the sampling area is, the higher the number of taxa detected will be (Horner-Devine et al., 2004). On the other hand, the distance-decay relationship is the name given to the decay of community similarity with geographic distance (Nekola and White, 1999; Whitaker, 2003).

This last pattern can be traced back to the four processes defined above, as shown by Hanson et al. (2012). When selective factors are organized spatially, i.e. in a gradient, selection will differentiate communities between locations. If dispersal of microorganisms is limited and

INTRODUCTION

cannot ensure homogenization, the slope of the DDR will be steeper, a phenomenon reinforced by random drift processes happening at all locations. Finally, diversification increases the variance in community similarity at all distances and lowers the height of the DDR by increasing local genetic diversity.

In addition to the four processes detailed above, it is interesting to disentangle the influence of historical legacies and of contemporary environmental factors in creating biogeographic patterns. Historical effects result from the application of all previously detailed processes, but are visible as distance effects where genetic isolation is maintained through dispersal limitation (Martiny et al., 2006). It is thus possible to disentangle historical and contemporary influences by observing the evolution of community similarity with geographic distance and with environmental distance.

Finally, when investigating geographic variations in community structure, it is also useful to investigate the variation of microbial community composition in terms of functional structure and taxonomic variation within functional groups. Functional structure is expected to be more likely influenced by environmental conditions whereas taxonomy might result more from biogeographic history, biotic interactions, and stochastic processes (Green et al., 2008; Raes et al., 2011).

2.4.2. Biogeography of deep sea benthic sediments

Abyssal plains are a rather uniform environment in terms of low temperature, high hydrostatic pressure, absence of light, low supply of organic matter and low permeability of the sediments. As stated before, phylum-level patterns of diversity in benthic sediments seem to reflect the existence of a core deep sea sediment microbiome, distinct from other deep sea environments (Orcutt et al., 2011; Walsh et al., 2016). However, studies of benthic community structure at finer taxonomic scale have shown clear geographic structuration at local and regional scale

INTRODUCTION

(Jacob et al., 2013; Buttigieg and Ramette, 2015; Liu et al., 2020; Li et al., 2021). At global scale, Bienhold et al. (2016) have shown the existence of a core community of a few very abundant organisms, while the rare taxa exhibited a high degree of endemism. In agreement with this observation, they also highlighted a positive range-abundance relationship. Finally, they observed a distance-decay relationship, illustrating a lack of dispersal and mixing for populations of deep sea sediments over large scales.

Environmental factors shaping microbial communities were often found related to energy availability, such as bathymetric depth or indicators of productivity regimes, e.g. total organic carbon content or phytodetrital pigment concentrations (Jacob et al., 2013; Buttigieg and Ramette, 2015; Bienhold et al., 2016).

In addition to these patterns and as mentioned above, microbial communities display a strong stratification with sediment depth. A number of studies have focused on the processes responsible for this vertical distribution of sedimentary microorganisms, and have suggested that subseafloor community assembly starts in the very first layers of sediment (Petro et al., 2019), and that microorganisms inherited from the water column are subsampled through selective survival (Walsh et al., 2016; Jochum et al., 2017; Petro et al., 2017; Starnawski et al., 2017; Kirkpatrick et al., 2019; Marshall et al., 2019).

3. Taxonomy of Archaea

3.1. Brief history

Relatively to the study of microorganisms, whose first observation can be traced back as far as the 17th century, Archaea have only been very recently “discovered”. Originally called “Archaeobacteria” (Woese and Fox, 1977), they were proposed as the third domain of life under their current name by Woese et al. (1990). Then thought to be divided between Euryarchaeota and Crenarchaeota, description of their wide diversity is still ongoing.

The first complete genome of an archaea, *Methanococcus jannaschii*, was published in 1996 by Bult et al.. From the 1990s on, most of the detection of unknown archaeal diversity was achieved through environmental surveys and culture-independent approaches, since cultivation of archaeal organisms proved challenging (reviewed in Sun et al., 2019; Baker et al., 2020). New phyla were proposed over the years such as Nanoarchaeota (Huber et al., 2002a), Korarchaea (Barns et al., 1994; Elkins et al., 2008), and Thaumarchaeota (Brochier-Armanet et al., 2008). These descriptions served to revise and amend the archaeal tree of life (Fig. 5). The proposed and putative phyla were split between phylum Euryarchaeota and three superphyla: Asgard (Zaremba-Niedzwiedzka et al., 2017), DPANN (Rinke et al., 2013; Castelle et al., 2015; Castelle and Banfield, 2018), and TACK (also called Proteoarchaeota) (Guy and Ettema, 2011; Petitjean et al., 2015).

Description of new archaeal lineages is still ongoing at an important rate, as shown by the recent reconstruction of 15 MAGs belonging to new putative phylum Brockarchaeota (De Anda et al., 2021). In the same way, a high number of archaeal clades are represented only by MAGs or SAGs (Single-cell assembled genomes), or even in some cases only by 16S sequences (Spang et al., 2017). Indeed, when considering both Archaea and Bacteria, the phylogenetic diversity of uncultivated microorganisms has been estimated to make up for more than 85% of known microbial diversity (Rinke et al., 2013).

INTRODUCTION

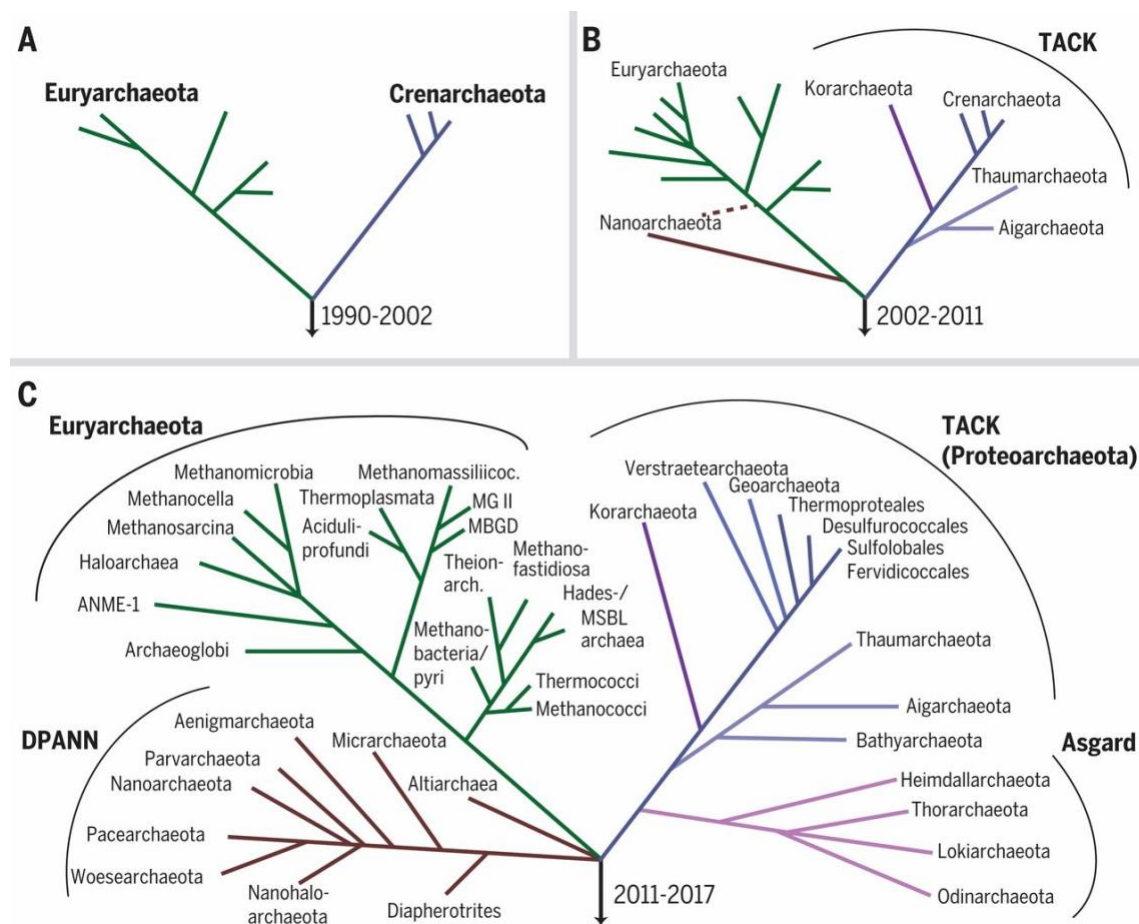


Figure 5: Evolution of the archaeal tree of life over the years. (Spang et al., 2017)

3.2. Recent discussions around archaeal taxonomy

Recently, an important conversation regarding the nomenclature and taxonomy of uncultivated microorganisms has been started (Hugenholtz et al., 2021).

Originally inheriting from traditional practices of basing taxonomy on phenotypic characteristics, the end of the 20th century saw a shift towards genetic information as the basis to attempt reconstruction of evolutionary relationships between species and thus infer phylogenetic placement and taxonomic affiliation (Lane et al., 1985; Olsen and Woese, 1993). Currently, the standards of definition of a microbial name require that it be attributed only to taxon represented by a type strain (cultured isolate) (Parker et al., 2015). This means that the wealth of genomic sequences added to the databases, much faster than single strains can be

INTRODUCTION

isolated and characterized, are used to refine phylogenetic trees, but do not receive names cohesive with the International Code of Prokaryotes and universally accepted.

In March 2020, the International Committee on Systematics of Prokaryotes rejected the use of sequence data as type material for naming prokaryotes (Sutcliffe et al., 2020). Arguments against this proposal included the lack of uniformly applied genome quality standards, the absence of observed phenotypic traits, and the potential for disorganized nomenclature (Bisgaard et al., 2019; Overmann et al., 2019). Another option put forward was the definition of a separate nomenclatural code for uncultivated Bacteria and Archaea (Murray et al., 2020). Parks et al. (2018) developed the Genome Taxonomy Database (GTDB) in an attempt to extract phylogenetic information from as many high-quality genome sequences as possible, both from cultured and uncultivated taxa, and to propose a standardized approach to taxonomy. Originally limited to Bacteria, the GTDB was later expanded to include both Bacteria and Archaea (Parks et al., 2020; Rinke et al., 2021). A comparison of the proposed archaeal taxonomy with NCBI phylum affiliations is presented in Fig. 6 extracted from Rinke et al. (2021).

INTRODUCTION

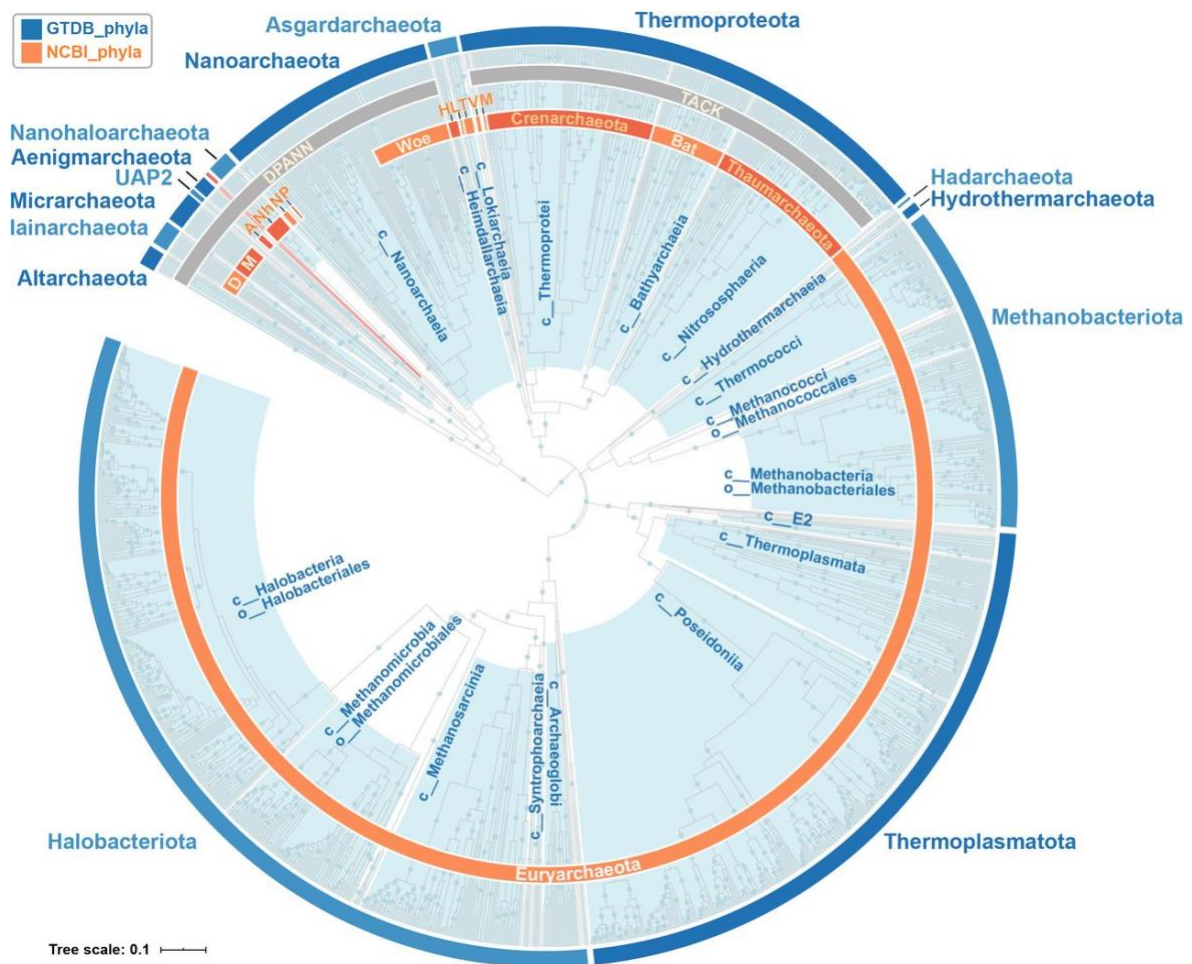


Figure 6: Rank normalized archaeal GTDB taxonomy proposed by Rinke et al. (2021). The dark blue ring illustrates the defined phyla, and the orange ring corresponds to the r89 NCBI phyla with two or more taxa. Light blue clades indicate the classes, and are labeled if they contain more than 10 taxa. The scale bar indicates 0.1 relative evolutionary divergence (RED). The following abbreviations are used: Bat (Ca. Bathyarchaeota), M (Ca. Marsarchaeota), V (Ca. Verstraetearchaeota), T (Ca. Thorarchaeota), L (Ca. Lokiarchaeota), H (Ca. Heimdallarchaeota), Woe (Ca. Woearchaeota), P (Ca. Parvarchaeota), N (Nanoarchaeota), Nh (Ca. Nanohaloarchaeota), A (Ca. Aenigmarchaeota), M (Ca. Micrarchaeota), D (Ca. Diapherotrites).

Rinke et al. proposed 16 archaeal phyla based on 122 concatenated conserved single copy marker proteins and normalized ranks. Interestingly, the DPANN superphylum was split between 9 phyla while the Asgard and TACK superphyla were reclassified as single phyla, with proposed names Asgardarchaeota and Thermoproteota. This entailed a reclassification

INTRODUCTION

of Thaumarchaeota as a class-level lineage, named Nitrososphaeria after the validly defined lineage of *Nitrososphaera viennensis* (Stieglmeier et al., 2014) (Fig. 7, extracted from Rinke et al., 2021).

In the most recent version of the SILVA database (Quast et al., 2013), SILVA v138 released in December 2019, the GTDB classification was used as an additional resource for taxonomy curation. However, this curation was based on a previous iteration of the GTDB and SILVA 138 includes only the following 15 archaeal phyla: Aenigm-, Alti-, Asgard-, Cren-, Eury-, Had-, Hydrotherm-, Iain-, Kor-, Micr-, Nano-, Nanohaloarchaeota, Thermoplasmatota, Halobacterota and uncultured. In this iteration, the name Crenarchaeota has been used as a placeholder name for the TACK superphylum, to the exception of Korarchaea, which might be confusing because of the history of domain Archaea. Class Nitrososphaeria has been adopted by the SILVA database as well, though some specific clades with non-systematic names were not renamed.

In the following work, we relied on SILVA 138 and the GTDB for taxonomic assignment and phylogenetic placement, in an attempt to take advantage of the most recent databases. We thus used the proposed names in the text for consistency, but we tried to provide context for these yet unusual designations.

INTRODUCTION

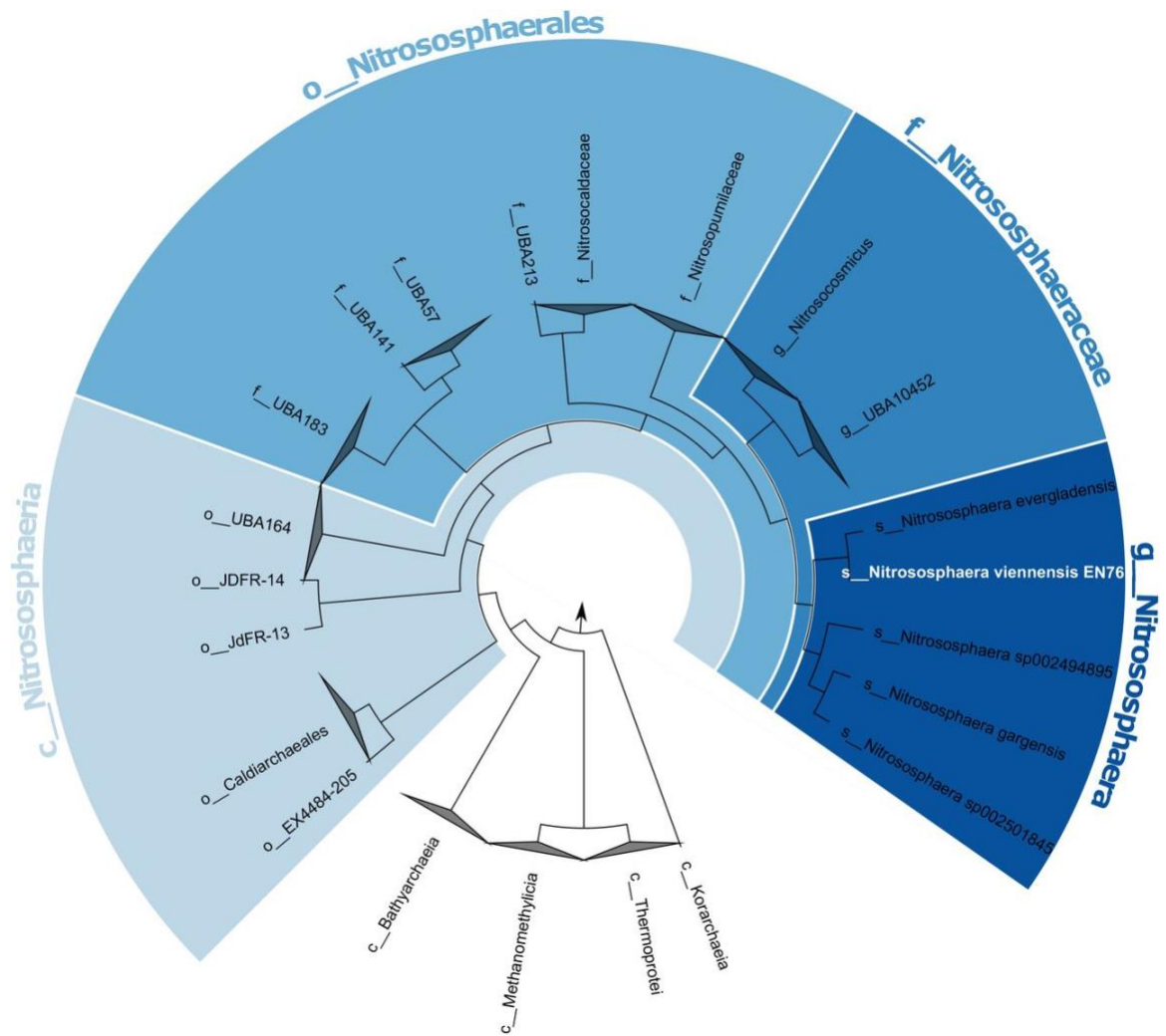


Figure 7: Reclassification of Thaumarchaeota members proposed by Rinke et al. (2021).

OVERVIEW AND OBJECTIVES

Study and characterization of deep-sea ecosystems has been the subject of strong interest in the last decades thanks to the development of pressure-resistant equipment and the revolution brought by NGS and cultivation-independent approaches.

Deep sea sediments constitute one of the largest habitats on Earth, and are hypothesized to be treasure troves of undiscovered taxonomic and functional microbial diversity. The benthic zone, the first few layers of seafloor sediment, are of particular interest because of their crucial role in biogeochemical cycles through early diagenesis of sinking organic matter and their position at the interface between pelagic and subsurface communities. However, a big part of deep sea research focuses on hotspots of diversity such as hydrothermal vents and cold seeps, so benthic communities are still sparsely described. It is thus important to characterize these environments, especially since they are or will be subjected to anthropogenic direct and indirect impacts, through plastic pollution, deep sea mining or changes in ocean biogeochemistry as a result of climate change.

The “Pourquoi pas les Abysses ?” project was developed by Ifremer starting in 2016 to make use of the possibilities offered by NGS and environmental DNA sequencing to bring light to the biodiversity of “the last frontier on Earth”. In the scope of this project, this PhD aimed at studying the bacterial and archaeal diversity found in benthic ecosystems of the deep ocean and the specificities of microbial life in these extreme conditions.

The first part of this thesis, in collaboration with Miriam Brandt, was dedicated to the establishment of standardized protocols for deep sea sedimentary diversity exploration through environmental DNA sequencing. The first three methodological studies presented here, part of Miriam Brandt’s PhD work, served to define appropriate sampling techniques, extraction methods, and bioinformatic pipeline to assess the diversity of benthic Bacteria,

Archaea, Protista and Metazoa. The last methodological study, original to this work, compared four molecular approaches to access the taxonomic diversity of benthic Archaea: metabarcoding with universal or domain-specific primers, SSU rRNA sequences extracted from unassembled metagenomic data, and single copy core gene profiles.

In the second part of this work, the methods implemented above served to characterize the microbial communities at the transition between the Mediterranean Sea and Atlantic Ocean along a longitudinal gradient. We explored biogeographic patterns, and the influence of historical and contemporary processes on community structure.

Finally, the last part of this thesis focused on the archaeal diversity found in samples collected during the first year of this PhD from two hadal trenches in the southern Pacific Ocean. In Chapter 3, we conducted a general taxonomic characterization of Archaea in this environment through co-occurrence network analysis, and identified putative associations of the members of the presumed symbiotic order Woesearchaeales (*Ca.* phylum Woesearchaeota). In Chapter 4, we used a metagenomic approach to reconstruct 53 MAGs affiliated to Thaumarchaeota (class Nitrososphaeria in the GTDB). We investigated the clade-level distribution of these populations in the abyssal and hadal zone with increasing depth in the sediments and relied on gene-level variability to explore possible selective pressures.

OBJECTIFS DE LA THESE

Un fort intérêt a été porté à l'étude et la caractérisation des écosystèmes marins profonds ces dernières décennies grâce au développement d'équipements résistants à la pression hydrostatique et à la révolution amenée par les techniques de séquençage massif nouvelle génération.

Les sédiments marins profonds constituent l'un des plus grands habitats sur Terre, et renferment, selon toute vraisemblance, une importante diversité taxonomique et fonctionnelle encore inconnue. La zone benthique, constituée des premières couches de sédiments des fonds marins, présente un attrait particulier à cause de son rôle crucial dans les rôles biogéochimiques à travers les premières étapes de la diagénèse, et à cause de sa localisation à l'interface entre les communautés pélagiques et les communautés de subsurface. Cependant, une part importante de la recherche sur les fonds marins se concentre sur les « hotspots » de diversité que constituent les cheminées hydrothermales ou les suintements froids, et les communautés benthiques sont encore peu décrites. Il est donc important de caractériser ces environnements, d'autant plus qu'ils sont ou seront impactés par les activités anthropiques, de façon directe ou indirecte, à travers la pollution plastique, l'extraction minière sous-marine ou les variations de biogéochimie de l'océan résultant du changement climatique.

Le projet « Pourquoi pas les Abysses ? » a été développé par l'Ifremer à partir de 2016 dans le but de tirer parti des possibilités offertes par le séquençage massif de l'ADN environnemental pour mettre en lumière la biodiversité de cette « dernière frontière sur Terre ». Dans le cadre de ce projet, cette thèse a eu pour but d'étudier la diversité bactérienne et archéale des écosystèmes benthiques de l'océan profond et les spécificités de la vie microbienne dans les conditions extrêmes qui les caractérisent.

La première partie de cette thèse, en collaboration avec Miriam Brandt, a été consacrée à la mise en place de protocoles standards pour l'inventaire de la biodiversité sédimentaire via le séquençage d'ADN environnemental. Les trois premières études méthodologiques présentées ici, réalisées dans le cadre de la thèse de Miriam Brandt, ont servi à définir les méthodes appropriées d'échantillonnage, d'extraction et d'analyse bioinformatique pour la description de la diversité benthique des Bactéries, Archées, Protistes et Métazoaires. La quatrième et dernière étude méthodologique, réalisée pour ce travail de thèse, a permis de comparer quatre méthodes moléculaires pour accéder à la diversité taxonomique des Archées benthiques : le métabarcoding à l'aide d'amorces universelles ou spécifiques des Archées, l'extraction de séquences d'ARNr SSU à partir de données métagénomiques brutes, et les profils de gènes centraux à copie unique.

Dans la seconde partie de ce travail de thèse, les protocoles standards résultant des études techniques ont été appliqués à la caractérisation des communautés microbiennes de la transition entre Méditerranée et Atlantique suivant un gradient longitudinal. Nous avons ainsi exploré les schémas biogéographiques et l'influence des processus contemporains et historiques sur la structure des communautés.

Enfin, la dernière partie de cette thèse a été consacrée à l'étude de la diversité archéenne présente dans les échantillons issus de deux fosses hadales du Pacifique Sud, collectés pendant la première année de doctorat. Dans le Chapitre 3, nous avons réalisé une description taxonomique générale des Archées de cet environnement à l'aide de l'analyse d'un réseau de cooccurrence, et nous avons identifié des associations putatives impliquant les membres de l'ordre présumé symbiotique Woearchaeales (*Ca.* Phylum Woearchaeota). Dans le Chapitre 4, nous avons utilisé une approche métagénomique pour reconstruire 53 MAGs affiliés aux Thaumarchées (classe Nitrososphaeria dans GTDB). Nous avons examiné la distribution par clade de ces populations dans les zones abyssale et hadale, à travers les premières couches de sédiments, et nous sommes appuyés sur la variabilité génique pour explorer les possibles pressions de sélection rencontrées.

CHAPTER 1

Defining standardized methods
for the study of benthic prokaryotic
and eukaryotic diversity

Context of this work and personal contribution

Metabarcoding is a fairly standard practice in microbiology for the study of environmental communities. However, due to the wider scope of the “Pourquoi pas les Abysses ?” project, it was necessary to define standardized methods for the characterization of benthic diversity of multiple biotic compartments: metazoans, protists, Bacteria and Archaea.

The first three papers presented in this chapter were published by Miriam Brandt as part of her PhD dissertation. They aimed at finding methodological trade-offs for the reliable characterization of each biotic compartment without compromising comparability between datasets. I was involved in the bioinformatic and statistical analyses, which I ran for the 16S metabarcoding datasets in these three studies. I also took part in the implementation and testing of the bioinformatic pipeline described in part 3 of this chapter, and participated in the writing of the final papers.

Finally, I performed most of the work presented in the fourth study with the help of Ferial Boudarka during her Masters 1 and 2 internships, with the exception of part of the extraction and sequencing steps.

1. Evaluating sediment and water sampling methods for the estimation of deep-sea biodiversity using environmental DNA

Résumé de l'article en français

La biodiversité présente dans les sédiments des fonds marins est encore peu caractérisée, malgré l'étendue de cet habitat sur notre planète. Le séquençage de l'ADN extrait d'échantillons environnementaux, et en particulier le métabarcoding, offre des perspectives très intéressantes pour la caractérisation rapide et à grande échelle de la diversité taxonomique présente dans un échantillon. Cependant, pour être reproductible, cette approche nécessite des méthodes d'échantillonnage standardisées et un choix attentif du substrat environnemental. L'étude présentée ici a pour but d'optimiser la caractérisation des communautés procaryotes (16S), protistes (18S V4) et métazoaires (18S V1-V2, COI), en comparant différentes stratégies d'échantillonnage des sédiments et de l'eau superficielle déployées simultanément au niveau d'un site sur le plateau continental.

En ce qui concerne le sédiment, pour tous les marqueurs étudiés, le tri de tailles par tamisage n'a pas eu d'effet significatif sur la diversité alpha détectée ou les profils taxonomiques au niveau du phylum. Il a en revanche permis d'augmenter la détection des phyla de la méiofaune.

Pour l'eau superficielle, les larges volumes obtenus avec la pompe *in situ* (~6000 L) ont généré une détection plus importante de la diversité des métazoaires que les boîtes d'échantillonnage de 7.5 L. Cependant, cette pompe étant restreinte à de grandes tailles de filtres (> 20 µm), elle n'a permis de capturer qu'une fraction de la diversité microbienne, tandis que les boîtes d'échantillonnage donnent accès au pico- et nanoplancton. En outre, les communautés de l'eau affleurant le sédiment diffèrent significativement des communautés sédimentaires, quel

CHAPTER 1

que soit le volume filtré, et seuls 3 à 8% des unités moléculaires sont partagés entre les deux types d'échantillons.

Dans l'ensemble, ces résultats démontrent que le tamisage des sédiments marins peut permettre une meilleure caractérisation des communautés métazoaires benthiques, et que l'eau superficielle n'est pas une alternative à l'échantillonnage des sédiments pour l'inventaire de la diversité benthique.



OPEN

Evaluating sediment and water sampling methods for the estimation of deep-sea biodiversity using environmental DNA

Miriam I. Brandt¹✉, Florence Pradillon², Blandine Trouche³, Nicolas Henry⁴, Cathy Liautard-Haag¹, Marie-Anne Cambon-Bonavita³, Valérie Cueff-Gauchard³, Patrick Wincker⁵, Caroline Belser⁵, Julie Poulain⁵, Sophie Arnaud-Haond¹✉ & Daniela Zeppilli²

Despite representing one of the largest biomes on earth, biodiversity of the deep seafloor is still poorly known. Environmental DNA metabarcoding offers prospects for fast inventories and surveys, yet requires standardized sampling approaches and careful choice of environmental substrate. Here, we aimed to optimize the genetic assessment of prokaryote (16S), protistan (18S V4), and metazoan (18S V1–V2, COI) communities, by evaluating sampling strategies for sediment and aboveground water, deployed simultaneously at one deep-sea site. For sediment, while size-class sorting through sieving had no significant effect on total detected alpha diversity and resolved similar taxonomic compositions at the phylum level for all markers studied, it effectively increased the detection of meiofauna phyla. For water, large volumes obtained from an in situ pump (~6000 L) detected significantly more metazoan diversity than 7.5 L collected in sampling boxes. However, the pump being limited by larger mesh sizes (>20 µm), only captured a fraction of microbial diversity, while sampling boxes allowed access to the pico- and nanoplankton. More importantly, communities characterized by aboveground water samples significantly differed from those characterized by sediment, whatever volume used, and both sample types only shared between 3 and 8% of molecular units. Together, these results underline that sediment sieving may be recommended when targeting metazoans, and aboveground water does not represent an alternative to sediment sampling for inventories of benthic diversity.

Environmental DNA (eDNA) metabarcoding is an increasingly used tool for non-invasive and rapid biodiversity surveys and impact assessments. Using high-throughput sequencing (HTS) and bioinformatic processing, target organisms are detected using their DNA directly extracted from soil, water, or air samples¹. Covering more than 50% of Planet Earth, the deep seafloor is mostly comprised of sedimentary habitats, characterised by a predominance of small organisms^{2,3} difficult to identify based on morphological features⁴, and by high local and regional diversity^{5–7}. Given its increased time-efficiency and its wide taxonomic applicability, eDNA metabarcoding is thus a good candidate for large-scale biodiversity surveys and Environmental Impact Assessments in the deep-sea biome.

Size-class sorting such as sieving or elutriation is usually performed on sediment samples in order to split the organisms by size and facilitate morphological characterization of meiofauna and macrofauna. For metabarcoding approaches, it also has the advantage of limiting the over dominance of large organisms, which may produce higher amounts of DNA template, resulting in an incomplete detection of small and abundant taxa. However,

¹MARBEQ, IFREMER, IRD, CNRS, Univ Montpellier, Sète, France. ²Centre Brest, Laboratoire Environnement Profond (REM/EEP/LEP), IFREMER, CS10070, 29280 Plouzané, France. ³IFREMER, CNRS, Laboratoire de Microbiologie Des Environnements Extrêmes (LM2E), Univ Brest, Plouzané, France. ⁴CNRS, Station Biologique de Roscoff, AD2M, UMR 7144, Sorbonne University, 29680 Roscoff, France. ⁵Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ of Evry, Paris-Saclay University, 91057 Evry, France. ✉email: miriam.isabelle.brandt@gmail.com; sophie.arnaud@ifremer.fr

sieving requires large volumes of sediment, is very time-consuming, and previous studies have found that the use of non-sieved material does not significantly alter metazoan diversity patterns⁸, suggesting that dominance of large (and often rare) taxa in the DNA extract does not result in important biases. Besides, for logistic reasons, the use of non-sieved sediment samples is preferable to (1) minimize on-board processing time, (2) minimize risks of contamination, and (3) facilitate other future applications (e.g., avoid RNA degradation, avoid losing the extracellular DNA compartment).

Finally, studies from various marine habitats have reported that benthic taxa could be found in aboveground water (overlying water layer to 6.5 m above seafloor), possibly due to sediment resuspension and transport, but also to active dispersal^{9–11}. Application of eDNA metabarcoding on deep-sea aboveground water could thus be a convenient alternative to surface sediment collection, as it involves simplified sample processing and shipping, while additionally allowing investigating benthopelagic diversity and dispersal capacities of benthic organisms. However, distance above seafloor has been variable (0.5–6.5 m) among studies^{9–11}, and so has the water volume sampled (12–1000 L). As the latter is a crucial aspect for efficient species detection¹², it remains unclear whether small volumes (< 10 L) are sufficient to obtain comprehensive species inventories in the deep-sea.

To evaluate the effect of sampling strategy on eDNA metabarcoding inventories targeting prokaryotes (16S V4–V5), unicellular eukaryotes (18S V4), and metazoans (18S V1–V2, COI) from deep-sea sediment and aboveground water, we compared biodiversity inventories resulting from (1) sieved versus unsieved sediment and (2) on-board filtration of 7.5 L of water collected with a sterile sampling box versus in situ filtration of large volumes (~6000 L) using a newly-developed pump.

Results

High-throughput sequencing results. A total of 26 million COI reads, 19 million raw 18S V1–V2 reads, 14 million 18S V4 reads, and 17 million 16S V4–V5 reads were obtained from three Illumina HiSeq runs of amplicon libraries built from pooled triplicate PCRs of 22 environmental samples, 2 extraction blanks, and 4–6 PCR blanks (Supplementary Table S4 online). The in situ pump yielded less raw reads for COI and 16S (Supplementary Fig. S1 online, $F = 4.02–14.4$, $p = 0.0003–0.03$), while more raw reads were recovered from both water sampling methods with 18S V4 ($F = 6.5$, $p = 0.007$). Water samples generally yielded fewer raw clusters ($F = 5.1–35.1$, $p = 3.2 \times 10^{-6}–0.02$), except for 18S V4 where numbers were comparable across sample types (Supplementary Fig. S1 online).

Bioinformatic processing (quality filtering, error correction, chimera removal, and clustering for metazoans) reduced read numbers to 20 million for COI, 12 million for 18S V1–V2, 11 million for 18S V4, and 10 million for 16S V4–V5, resulting in 10,351 and 17,608 raw OTUs for COI and 18S V1–V2 respectively; 35,538 raw 18S V4 ASVs, and 62,646 raw 16S ASVs (Supplementary Table S4 online). For eukaryote markers, 17–55% of the raw reads remained in PCR blanks after bioinformatic processing, while 50–75% remained in extraction blanks and 52–87% in true samples. In contrast, with 16S, these values were at 87% for PCR blanks, 67% for extraction blanks, and 29–73% for true samples. Thus, negative control samples accounted for 7–13% of bioinformatically processed reads with eukaryotes, compared to 27% with prokaryotes. The vast majority of 16S reads generated by negative controls belonged to a common contaminant of *Phusion* polymerase kits, which is well amplified in low concentration samples such as negative controls. These reads however accounted for < 1% of 16S ASVs. After data refining (decontamination, removal of all control samples and of all unassigned or non-target clusters), rarefaction curves showed a plateau was reached for all samples except in situ pump samples with microbial loci and sediment samples with 18S V4, suggesting that not all protist and prokaryote diversity was captured at this sequencing depth in these samples (Supplementary Fig. S1 online). Refined datasets contained 7 million reads for prokaryotes and between 4.8 and 8.5 million target reads for eukaryotes, delivering 38,816 prokaryote ASVs (16S V4–V5), 8031 protist ASVs (18S V4), and 2,319 (COI) and 1,460 (18S V1–V2) LULU curated metazoan OTUs (Table S4). For COI, while only 1.2% of metazoan OTUs were unassigned at phylum-level, 57% had BLAST hit identities < 80%, i.e. unreliable at phylum-level. For 18S, 7% (18S V4) to 16% (18S V1–V2) of final ASVs/OTUs were unassigned at phylum-level, but only 12% (18S V4) and 13% (18S V1–V2) had BLAST hit identities < 86%. For 16S, 0.9% and 3% of ASVs had no or unreliable phylum-level assignments, respectively.

Alpha diversity between sampling methods. The number of recovered molecular clusters significantly differed with sampling method, water samples detecting less diversity than sediment samples for metazoans and prokaryotes (COI: $F = 20.1$, $p = 4.4 \times 10^{-5}$, 18S V1–V2: $F = 6.5$, $p = 0.01$, 16S: $F = 56.0$, $p = 3 \times 10^{-7}$), but both sample types recovering similar levels of protist diversity (18S V4: $F = 2.9$, $p = 0.07$, Fig. 1). Sieved and unsieved sediment resulted in statistically comparable total cluster numbers in all loci investigated (Table 1) although, for metazoans detected with 18S V1–V2, this lack of significance was likely due to the very low yield observed in one sieved sample (PL11), as the other sieved samples detected considerably more OTUs (Fig. 1). For metazoans detected with COI, sieved samples tended to detect less OTUs although being based on larger sediment volumes (pool of 5 DNA extracts). The number of total recovered OTUs did not differ significantly between the water sampling box and the in situ pump for metazoans (COI, 18S V1–V2), but differences were observed for unicellular eukaryotes (18S V4) and prokaryotes (16S V4–V5) depending on the sampling box size fraction (Table 1), with the smallest fraction (0.2–2 μm) detecting more ASVs (Fig. 1).

Differences in total recovered diversity were not solely a result of differences in sample volume. Indeed, sieved samples, based on 3–6 times more sediment, did not consistently detect more diversity (Fig. 1). Similarly, the in situ pump, although sampling ~800 times more water than the sampling box, did not consistently detect more diversity than any size fraction of the sampling box (Fig. 1).

Recovered diversity among sample types strongly differed depending on phylum (Fig. 2). For metazoans, water samples led to the detection of significantly higher numbers of OTUs than sediment samples for Arthropoda,

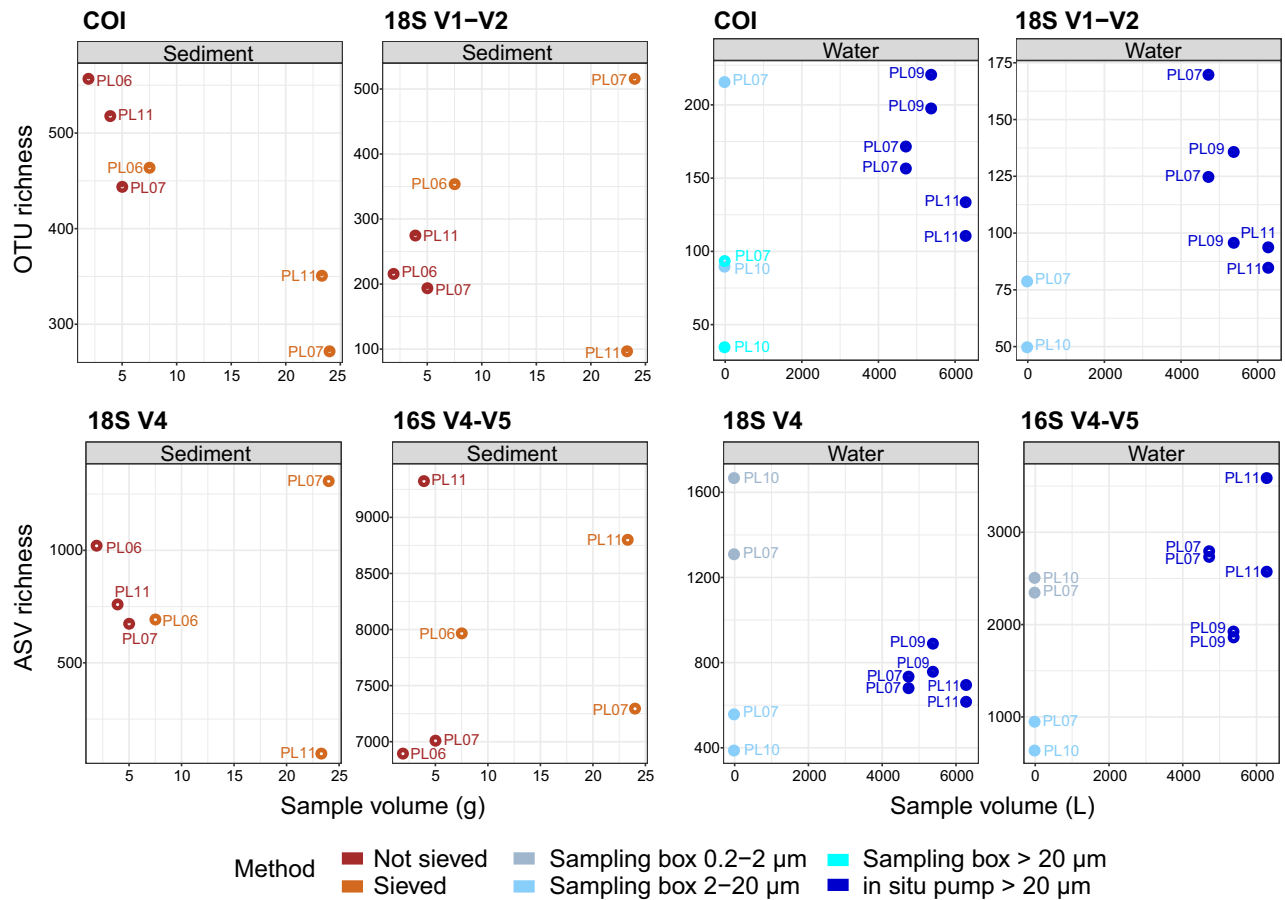


Figure 1. Numbers of metazoan OTUs (COI, 18S V1–V2), unicellular eukaryote (18S V4) and prokaryote (16S V4–V5) ASVs recovered by deep-sea sediment (brown) and aboveground water (blue), using for each sample type two sampling methods based on varying amounts of starting material. Sediment was either sieved through 5 mesh sizes to size-sort organisms prior DNA extraction, or DNA was extracted directly from crude sediment samples. Water was collected with a 7.5 L sampling box, allowing recovery of up to two size classes per taxonomic compartment, or sampled in large volumes with an in situ pump. Cluster abundances were calculated on rarefied datasets.

Chordata (COI, t -tests, $p = 0.006$ – 0.01) and Mollusca (18S V1–V2, t -test, $p = 0.03$), and some phyla like Brachiopoda, Ctenophora, Echinodermata, or Gastrotricha, were only detected in water samples (Fig. 2). In contrast, phyla such as Platyhelminthes, Porifera (COI, 18S V1–V2, t -tests, $p = 0.001$ – 0.04), Kinorhyncha, Nematoda, Tardigrada, or Xenacoelomorpha (18S V1–V2, t -tests, $p = 0.001$ – 0.04) produced significantly more OTUs in sediment than water samples (Fig. 2). Similarly, some protistan groups, such as the Acantharea, Chlorophyta, Dinophyceae, or Syndiniales (t -tests, $p = 0.002$ – 0.03) were predominant in water samples (Supplementary Fig. S2 online), while others were significantly more diverse in sediment, e.g., Apicomplexa, Filosa groups, Ciliophora, Labyrinthulea, RAD-B (t -tests, $p = 0.001$ – 0.04). For prokaryotes, most lineages were predominant in sediment (t -tests, $p = 2.2 \times 10^{-7}$ – 0.02 , e.g., Archaea, Acidobacteria, Actinobacteria, Bacteroidetes, Chloroflexi, Delta-, Gamma-proteobacteria, Gemmatimonadetes, Latescibacteria, Hydrogenedetes, Nitrospirae, Planctomycetes), and only Cyanobacteria were significantly more diverse in water samples (t -test, $p = 0.001$).

For sediment, recovered levels of alpha diversity among sampling methods also varied by phyla and organism size class (Fig. 2). For meiofauna phyla, best detected with 18S V1–V2, more OTUs were detected from sieved than from unsieved sediment (Kinorhyncha, Nematoda, Platyhelminthes, Rotifera, Tardigrada, Xenacoelomorpha), although this difference was not significant, likely due to the low sample size. However, differences in alpha diversity among sampling methods were not only a result of differences in sample volume, as some unsieved samples yielded similar or greater numbers of OTUs than sieved samples for many phyla (Supplementary Fig. S3 online). Sieved and unsieved sediment detected comparable ASV numbers in most microbial groups, except the Chrysophyceae, Actinobacteria, Cyanobacteria, Gammaproteobacteria, Nanoarchaeaeota (Supplementary Fig. S2 online, paired t -tests, $p = 0.02$ – 0.04).

Water sampling methods strongly differed in terms of recovered alpha diversity depending on taxonomic compartment. The in situ pump generally detected more metazoan diversity than the sampling box, and phyla such as Brachiopoda, Ctenophora, and Echinodermata were only detected by the pump (Fig. 2). However, the in situ pump detected significantly more ASVs than the sampling box only in some taxonomic groups for protists (Bacillariophyta, Oomycota, Phaeodarea) and prokaryotes (e.g., Bacteroidetes, Chlamydiae, Firmicutes, Tenericutes,

Pairwise comparison	Cluster richness				Community differentiation			
	COI	18S V1–V2	18S V4	16S V4–V5	COI	18S V1–V2	18S V4	16S V4–V5
Not sieved/ Sieved*	0.26	0.64	0.99	1.0	0.4, R ² =0.22	0.1, R ² =0.22	0.1, R ² =0.26	0.1, R ² =0.42
Sampling box 0.2–2 µm/ not sieved*	na	na	0.26	<0.0001	na	na	0.1, R ² =0.64	0.1, R ² =0.87
Sampling box 2–20 µm/ Not sieved*	<0.0001	0.10	0.63	<0.0001	0.1, R ² =0.51	0.1, R ² =0.46	0.1, R ² =0.57	0.1, R ² =0.89
Sampling box > 20 µm/not sieved*	<0.0001	na	na	na	0.1, R ² =0.50	na	na	na
Sampling box 0.2–2 µm/ sieved*	na	na	0.10	<0.0001	na	na	0.1, R ² =0.56	0.1, R ² =0.88
Sampling box 2–20 µm/ sieved*	0.01	0.02	0.86	<0.0001	0.1, R ² =0.44	0.1, R ² =0.43	0.1, R ² =0.49	0.1, R ² =0.89
Sampling box > 20 µm/ sieved*	0.0001	na	na	na	0.1, R ² =0.41	na	na	
In situ pump/ not sieved	<0.0001	0.13	0.99	<0.0001	0.01 , R ² =0.32	0.01 , R ² =0.29	0.02 , R ² =0.45	0.01 , R ² =0.78
In situ pump/ sieved	0.0001	0.002	1.00	<0.0001	0.01 , R ² =0.28	0.01 , R ² =0.28	0.01 , R ² =0.38	0.007 , R ² =0.78
Sampling box 0.2–2 µm/in situ pump	na	na	0.04	1.0	na	na	0.03 , R ² =0.49	0.04 , R ² =0.74
Sampling box 2–20 µm/in situ pump	0.99	0.70	0.79	0.001	0.03 , R ² =0.25	0.03 , R ² =0.25	0.04 , R ² =0.43	0.04 , R ² =0.59
Sampling box > 20 µm/in situ pump	0.10	na	na	na	0.04 , R ² =0.21	na	na	na
Sampling box 0.2–2 µm/ Sampling box 2–20 µm*	na	na	0.03	0.007	na	na	0.3, R ² =0.47	0.3, R ² =0.89
Sampling box 0.2–2 µm/ Sampling box > 20 µm*	na	na	na	na	na	na	na	na
Sampling box 2–20 µm/ Sampling box > 20 µm*	0.26	na	na	na	0.3, R ² =0.45	na	na	na

Table 1. Effect of sampling method on cluster richness and community structure for the 4 studied genes. Pairwise comparisons of ANOVAs of ASV/OTU richness performed using rarefied datasets. Pairwise PERMANOVAs were performed on rarefied datasets using Jaccard distances for metazoans and Bray–Curtis distances for 18S V4 and 16S V4–V5. Significance was evaluated by permuting 999 times when possible, and comparisons where this was not possible are marked by *. Significant *p* values are in bold. R²: R-squared.

Lentisphaerae, and Delta-, Gammaproteobacteria, see Supplementary Fig. S2 online, *t*-tests, $p=9 \times 10^{-5}$ –0.003). Other clades were significantly more diverse in the sampling box (e.g., the protist groups Haptophyta, Picozoa, Telonemia, and the Cyanobacteria, *t*-tests, $p=0.002$ –0.02). With the sampling box, the smallest size fraction (0.2–2 µm) allowed recovering more alpha diversity in all microbial groups than the larger size fraction (2–20 µm). This difference was significant only for Chlorophyta, Labyrinthulea, Chloroflexi, and Verrucomicrobia (paired *t*-tests, $p=0.01$ –0.03), although non-significant comparisons may result from the limited number of samples available. The two size fractions available with the sampling box for COI (2–20 µm, > 20 µm) did not reveal differences in diversity recovery with size class, as most phyla were detected equally well in both (Fig. 2).

Effect of sampling method on community structures. Community compositions significantly differed among sampling methods for all investigated loci (COI: pseudo-F=2.3, $p=0.001$; 18S V1–V2: pseudo-F=2.3, $p=0.001$, 18S V4: pseudo-F=4.1, $p=0.001$, 16S: pseudo-F=18.3, $p=0.001$) and sampling method accounted for 41–45% of variation among samples for metazoans (COI, 18S V1–V2), 60% for protists (18S V4), and 87% for prokaryotes (16S).

Pairwise PERMANOVAs showed that community structures differed most strongly among sample types (water or sediment, R²=0.28–0.89), although not all pairwise comparisons were significant, likely due to the limited number of samples available for the sampling box (Table 1). Relative taxonomic compositions revealed by

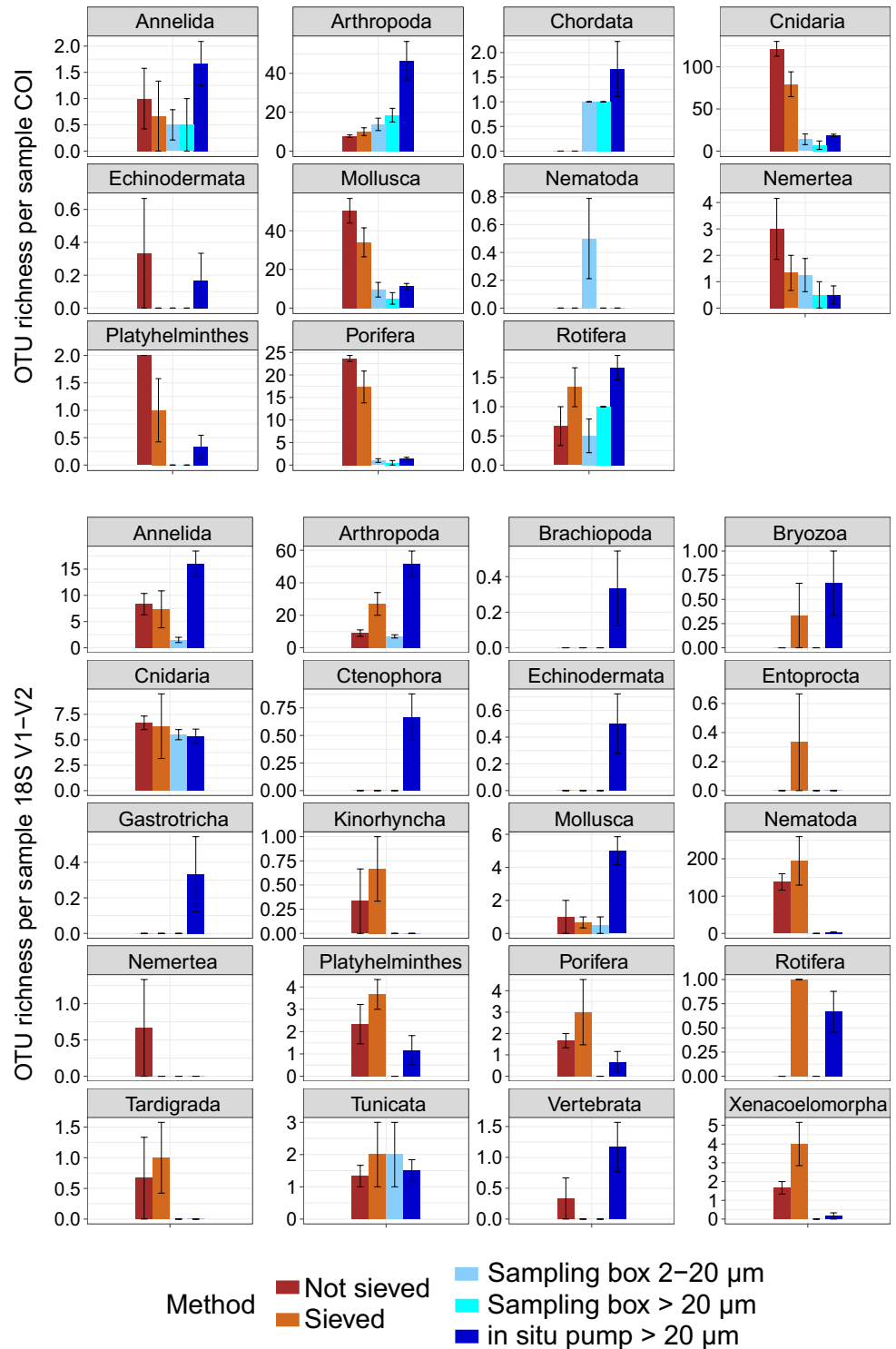


Figure 2. Mean numbers (\pm SE) of metazoan COI and 18S V1–V2 OTUs detected in target phyla for sediment (brown) and water (blue), using two sampling methods for both sample types. Sediment was either sieved to size-sort organisms prior DNA extraction, or DNA was extracted directly from crude sediment samples. Water was collected with a 7.5 L sampling box, allowing recovery of two size classes, or sampled in large volumes with an in situ pump. OTU numbers were calculated on rarefied datasets.

aboveground water samples differed from sediment samples, with higher proportions of arthropods, chordates, annelids, and tunicates in the water samples, while nematodes, poriferans, platyhelminths, and xenacoelomorphs

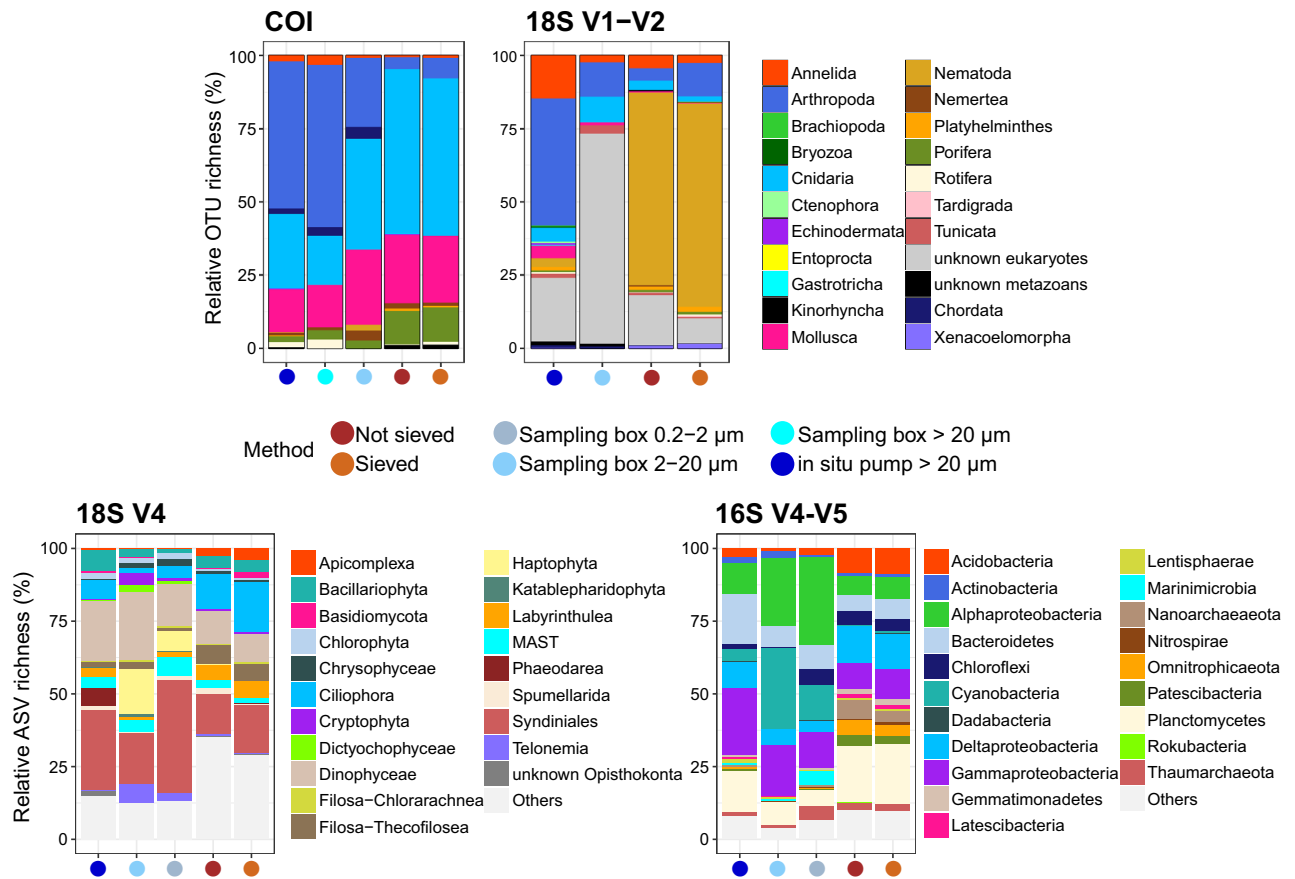


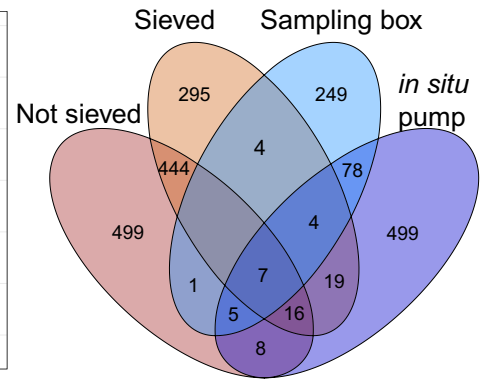
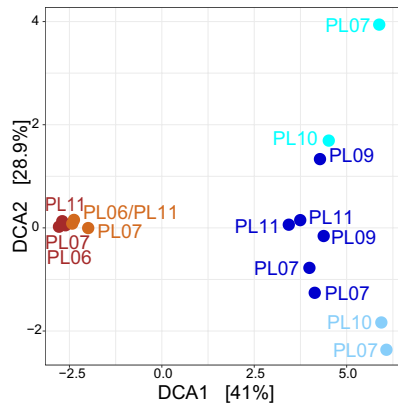
Figure 3. Patterns of relative cluster abundance resolved by eDNA metabarcoding of deep-sea sediment (brown) and aboveground water (blue), using two sampling methods for both sample types, and using four barcode markers targeting metazoans (COI, 18S V1–V2), micro-eukaryotes (18S V4), and prokaryotes (16S V4–V5). Sediment was either sieved to size-sort organisms prior DNA extraction, or DNA was extracted directly from crude sediment samples. Water was collected with a 7.5 L sampling box, allowing recovery of up to two size classes per taxonomic compartment, or sampled in large volumes with an in situ pump. Top 20 most abundant taxa are displayed for microbial groups.

were predominant in sediment samples (Fig. 3 COI and 18S V1–V2). Similarly, protist diversity in aboveground water samples was dominated by Dinophyceae, Haptophyta, Phaeodarea, Syndiniales, and to a lesser extent Bacillariophyta and Telonemia, while apicomplexans, ciliates, filosaurs, and labyrinthuleans represented higher proportions of diversity in sediment samples (Fig. 3 18S V4). For prokaryotes, aboveground water communities were characterised by Alphaproteobacteria, Cyanobacteria, and Gammaproteobacteria, while Acidobacteria, Deltaproteobacteria, Archaea, Latescibacteria, and Planctomycetes showed higher diversity in sediment (Fig. 3 16S V4–V5).

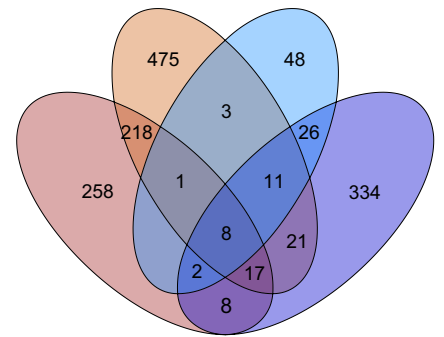
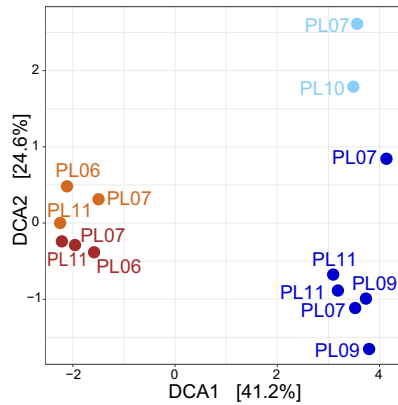
Only 3% (COI), 5% (18S V1–V2), 8% (18S V4), and 5% (16S) of clusters were shared between sediment and water samples, resulting in strong segregation in ordinations (Fig. 4). For metazoans, taxa shared among water and sediment samples were mostly assigned to hydrozoans (COI, 28%, 18S, 7%), copepods (COI, 6%, 18S, 20%), gastropods (COI, 31%), demosponges (COI, 6%), or polychaetes (18S, 10%), and chromadorean nematodes (18S, 11%). For protists, ASVs shared among sample types primarily belonged to the Syndiniales (39%), but other taxa included dinophyceans (11%), filosaurs (9%), labyrinthuleans (5%), and bacillariophytes (6%). For prokaryotes, ASVs shared across sample types were predominantly belonging to the Proteobacteria (Gamma, 19%, Alpha, 10%, Delta, 8%), Bacteroidetes (15%), or Planctomycetes (16%).

Sediment processing did not significantly affect detected community structures (Table 1), and sieved and unsieved sediment resolved comparable communities at the phylum-level (Fig. 3), although community segregation was observed in ordinations of metazoans resolved with 18S V1–V2 and protists resolved with 18S V4 (Fig. 4). Between 21 and 36% of sediment OTUs/ASVs were shared among sieved and unsieved sediment samples. Shared metazoan OTUs primarily belonged to Hydrozoa (18S, 2.5%, COI, 32%, Siphonophorae, Anthoathecata, Leptothecata), Demospongiae (COI, 9%), Gastropoda (COI, 32%), Nematoda (18S, 61% Chromadorea, 11% Enoplea), Polychaeta (18S, 2.5%), or Copepoda (18S, 4.5%). Microbial ASVs shared among sieved and unsieved sediment mostly belonged to Syndiniales (17%), Filosa (19%), Ciliophora (11%), Dinophyceae (9%), Planctomycetes (22%), Acidobacteria (10%), or Proteobacteria (Gamma, 9%, Alpha, 8%, Delta, 11%).

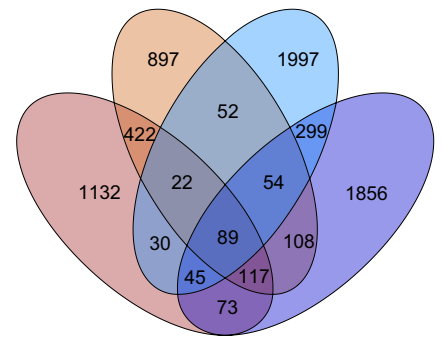
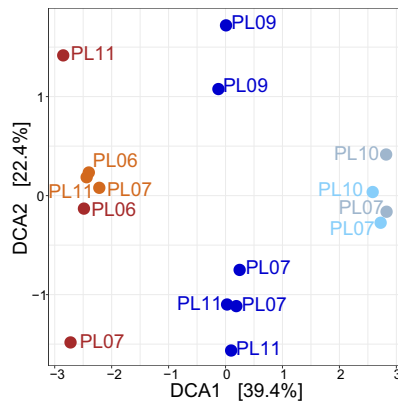
COI



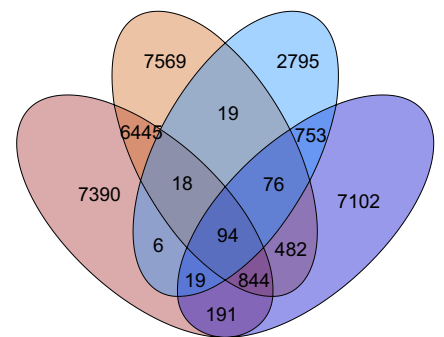
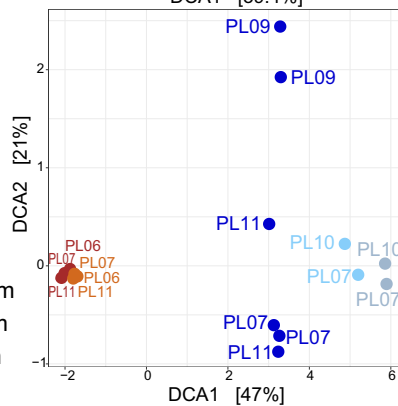
18S V1-V2



18S V4



16S V4-V5



- Not sieved
- Sieved
- Sampling box 0.2–2 µm
- Sampling box 2–20 µm
- Sampling box > 20 µm
- in situ pump > 20 µm

Figure 4. Detrended Correspondence Analyses (DCA) ordinations (left) and Venn diagrams (right), showing differences in community compositions detected by deep-sea sediment (brown) and aboveground water (blue) for metazoans (COI and 18S V1–V2), micro-eukaryotes (18S V4), and prokaryotes (16S V4–V5). Community segregation is strongest between sample types, but also among target size class in the water samples. Sediment was either sieved to size-sort organisms prior DNA extraction, or DNA was extracted directly from crude sediment samples. Water was collected with a 7.5 L sampling box, allowing recovery of two size classes in each taxonomic compartment, or sampled in large volumes with an in situ pump.

In contrast, sampling method significantly affected resolved community structure for water, as for all size fractions, the sampling box detected significantly different communities than the in situ pump (Table 1). Both sampling methods resolved different communities at the phylum-level (Fig. 3), and water samples always clustered apart in ordinations (Fig. 4) for all taxonomic compartments investigated. Between 8 and 11% of ASVs/OTUs detected in water were shared between the in situ pump and the sampling box. Taxonomic structures resolved by both sampling methods clearly changed due to targeted size fraction. The sampling box's 2–20 µm size fraction did not detect the same metazoan community assemblage as the > 20 µm assemblage detected by the pump (Fig. 3 COI and 18S V1–V2). Similarly, for microbial data, the in situ pump targeting the > 20 µm size class, and the sampling box targeting both the 2–20 µm and the 0.2–2 µm size classes, detected different community assemblages. For protists, the in situ pump detected higher proportions of ASVs for Bacillariophyta, Ciliophora, Labyrinthulea, or Phaeodarea, while the sampling box detected more cryptophytes, haptophytes, MAST, and telonemians (Fig. 3 18S V4). For prokaryotes, the sampling box detected more diversity in the Alphaproteobacteria, Chloroflexi, or Marinimicrobia (Fig. 3 16S V4–V5).

Discussion

Importance of substrate nature. Sediment samples, whether sieved or unsieved, led to the detection of higher numbers of metazoan OTUs and prokaryote ASVs than water samples (Fig. 1), indicating that more diversity could be found in the benthos compared to the pelagos at this Mediterranean site for those groups. For unicellular eukaryotes, the difference in diversity between sediment and aboveground water was not significant. However, this may primarily be due to the fact that more benthic protists (e.g., filicosans, labyrinthuleans, and ciliates) were well detected by water samples (Supplementary Fig. S2 online). Indeed, 19% of protist sediment ASVs were also detected in the water samples, while for other loci this percentage was at 5–8%. These findings are congruent with other studies in the marine realm that reported notably higher diversity in sediments compared to seawater^{13–15} for microbial communities, and show that higher diversity can also be expected for metazoans.

Community compositions differed markedly between sediment and aboveground water samples for all life compartments investigated (Figs. 3, 4), and only 3–8% of total molecular clusters were shared between substrate types, a range congruent with previous findings^{11,15,16}. Metazoan infauna taxa (e.g., nematodes, platyhelminths, kinorhynchans, tardigrades, and xenacoelomorphs) were specifically detected by sediment samples, while other epibenthic, benthopelagic, and pelagic metazoans were more prevalent in water samples (e.g., echinoderms, chordates, ctenophores). Similarly, with protists and prokaryotes, sediment samples detected lineages typically reported in the deep seafloor, with prokaryotic communities mostly comprised of Proteobacteria, Acidobacteria, Planctomycetes, Archaea, Bacteroidetes, and Chloroflexi^{17–20}, and protist communities characterized by benthic heterotrophic groups such as ciliates, labyrinthuleans, and filicosans^{21,22}. Water samples instead recovered taxa commonly reported in pelagic studies, with unicellular eukaryotes such as dinoflagellates (Dinophyceae, Syndiniales), radiolarians (Acantharea, Phaeodarea, Spumellarida), or MAST (incl. diatoms, Chlorophyta, Chrysophyceae)^{11,23,24}, and bacterial groups such as Alpha- and Gamma-proteobacteria, Bacteroidetes and Cyanobacteria^{25–27}.

Most of the metazoans shared among sediment and water samples displayed benthopelagic life cycles with a benthic adult and a pelagic larva (hydrozoans, gastropods, demosponges, polychaetes), indicating that the detection of benthic taxa in water samples may predominantly reflect the occurrence of dispersal phases of those organisms¹⁰. Other metazoan OTUs shared across sediment and water belonged to Copepoda (incl. Cyclopoida and Harpacticoida), Polychaeta, and Nematoda, confirming that active dispersal and/or resuspension of benthic taxa can also occur⁹. Similarly, benthic protists (e.g. filicosans, labyrinthuleans), and bacteria known to occur at the sediment–water interface (e.g., Bacteroidetes, Planctomycetes)^{14,28}, were also predominant in this shared fraction, supporting the existence of sediment resuspension dynamics. Finally, the presence of pelagic taxa such as Actinopteri, scyphozoans, cephalopods, diatoms (Bacillariophyta), and dinoflagellates (Dinophyceae and parasitic Syndiniales) in both sediment and water samples supports the fact that dead material, detritus, or faecal pellets can sink to the deep seafloor²⁹.

Overall, our results confirm previous findings showing that sample nature strongly affects the type of organisms targeted by eDNA metabarcoding^{30,31}, and underline that eDNA from water samples cannot be used to comprehensively survey benthic communities^{16,32,33}, even when large volumes of aboveground water are collected.

Sieving sediment is not essential for comprehensive benthic biodiversity surveys. Studies investigating the effect of size-sorting in macroinvertebrates showed that sorting organisms by size and pooling them proportionately according to their abundance led to a more equal amplification of taxa, the sorted samples recovering 30% more taxa than the unsorted samples at the same sequencing depth³⁴. The size fractions used in this study were specifically aiming to concentrate the meiofauna (32 µm–1 mm) compartment, which is known to be important in deep-sea sediments, both in terms of abundance and biomass^{2,35,36}. Meiofauna taxa, best captured by 18S V1–V2, were more numerous in sieved than unsieved sediment samples, and total recovered OTU numbers were higher in sieved than in unsieved samples for two cores with this marker (Fig. 1). These differences were however not significant. It could be that the equimolar pooling performed with DNA extracts from each different size fraction maintained biases in abundance, as larger organisms contributed more DNA molecules within each size fraction. Alternatively, some individual size fractions having yielded low DNA concentrations, their equimolar pool effectively diluted the size fractions that generated most of the DNA. Indeed, highest DNA recovery was observed in the 20–40 µm and the 40–250 µm size fractions for all cores, while the larger size fractions had DNA concentrations < 1 ng/µL. This problem was particularly severe for PL11, explaining why this core performed so poorly in the sieved treatment for both 18S markers (Fig. 1). Proportional pooling may be a better approach, but is feasible only if relative abundance of organisms in each size class can be calculated (e.g.,

using dry sample and specimen weights). A more accurate approach would be to sequence each size fraction separately, which would likely result in many more ASVs/OTUs due to increased sequencing depth. This however would also increase sequencing costs five-fold. An alternative to sequencing each size fraction separately could be the pooling of the size fractions after PCR, which would reduce size related biomass biases during PCR amplification, without increasing sequencing costs. However, the fact that more diversity was detected when sieving than when not sieving at the same sequencing depth for the 18S marker (Supplementary Fig. S4 online), indicates that sieving effectively reduces biomass biases, thus allowing the detection of more diversity at the same sampling depth. Alternatively, new technologies affording much higher sequencing depths³⁷ might circumvent the need for size-class sorting in the future.

The advantage provided by sieving observed in this study for some phyla may also result from the fact that sieved samples were based on more starting material, as five DNA extractions were performed for the sieved treatment (one for each size fraction), when only one was performed for non-sieved sediment. Sample volume can however not fully explain differences in recovered diversity, as the latter varied considerably within each method, and samples based on larger sediment volumes did not consistently yield more OTUs/ASVs (Fig. 1, Supplementary Fig. S3 online).

Elutriation (i.e. resuspension of organisms and pouring of supernatant on a 32- μ m sieve) or density extraction techniques are other methods traditionally used to study meiofauna^{38,39}. These allow to process whole sediment layers more rapidly than sieving, and effectively concentrate metazoan organisms³⁸. However, if the retention of organisms is achieved using only a single mesh size marking the lower size boundary of meiofauna, this also maintains size-abundance biases. Thus, whether sieving, elutriating, or density extracting, mesh sizes for size-class sorting have to be carefully chosen in order to reach the best compromise between processing time and biomass biases. As underlined by Elbrecht et al. (2017)³⁴, sorting is most useful when samples contain specimens with biomasses spanning several orders of magnitude. Given that deep-sea sediments contain large numbers of small organisms, and given the high detection capacity of metabarcoding, implementing five mesh sizes for sorting metazoans may be excessive. Instead, separating organisms into small, medium, and large size categories, as performed by Elbrecht et al. (2017)³⁴ for freshwater macroinvertebrates and by Leray & Knowlton (2015)⁴⁰ for coastal benthic communities may be sufficient to maximize metazoan species detection.

However, the rationale behind size sorting should be carefully considered when implementing an eDNA metabarcoding study on the deep seafloor. Indeed, for general biodiversity studies not targeting rare or invasive species, the proportion of abundant taxa is most relevant to reach accurate conclusions, and it may thus not be necessary to detect all small and rare taxa in such studies. Moreover, effects of size sorting on other taxonomic compartments have to be taken into consideration. For microbial organisms, sieving down to a 20- μ m mesh size is very likely to result in the loss of most small and/or free-living taxa. This idea is supported by the fact that metazoan OTUs shared between sieved and unsieved sediment were mainly assigned to macrofauna (> 1 mm), indicating that small taxa predominantly explain the differences obtained between both methods. For protists and prokaryotes, although sieved and unsieved sediment uncovered comparable alpha diversity levels (Fig. 1), and resolved similar taxonomic compositions at phylum level (Fig. 3), ordinations indicated that communities segregated with processing method for protists (Fig. 4 18S V4). Many sediment microorganisms are living within biofilms (e.g., Bacteroidetes, Archaeae), attached to sediment particles (e.g., Planctomycetes) or as symbionts of larger taxa (e.g., Syndiniales, some Dinophyceae and Proteobacteria), making their retention on a 20- μ m sieve possible. Our results support this idea, as microbial ASVs shared among sieved and unsieved sediment were mostly belonging to those groups or to taxa larger than 20 μ m (e.g. ciliates), possibly explaining the non-significant difference we obtained in PERMANOVA (Table 1).

Finally, sieving is associated to higher contamination risks, as sieves need to be carefully washed between samples and water used for sieving (or elutriation) needs to be ultra-filtered (which can be problematic for the large volumes needed). Considering the limited improvement gained by sieving on metazoan communities, the logistic inconvenience, and the risk of bias for other taxonomic compartments, DNA extractions performed directly on 10 g of sediment appear as a satisfactory approach for large-scale biodiversity surveys targeting multiple life compartments.

Adjusting water sample volume and filter mesh size to target organisms. Numerous aquatic metabarcoding studies have highlighted that sampled water volume is a key factor affecting species detection rates with eDNA, and has to be adapted to the target ecosystem⁴¹. Positive relationships between increased water volume and increased detection rate have been reported for macroinvertebrates and amphibians^{42,43}, and studies in freshwater ecosystems have shown that 20 l to 30–68 l of water are necessary to detect entire metazoan communities^{12,44,45}. While 1 L may be appropriate for macroinvertebrate detection in rivers⁴² or marine surface waters⁴⁶, the results presented here clearly show that 7.5 L of deep-sea water are not sufficient to accurately detect metazoan fauna. The sampling boxes detected less metazoan diversity than the in situ pump (Fig. 1), and failed to detect many phyla with 18S V1–V2 (Fig. 2). Overall, 21% (COI) to 39% (18S V1–V2) of metazoan diversity detected in water was recovered by the sampling box, compared to 89% (COI) and 71% (18S V1–V2) by the in situ pump. This reflects the low abundance and biomass of large organisms in deep waters, combined with the very limited lifetime of extracellular DNA in seawater^{47–50}.

Water sampling methods for eDNA metabarcoding relying on on-board filtration or precipitation are intrinsically limited by the amount of water that can be processed. Although purpose-built sampling equipment has been developed for increased efficiency and standardization, filtration flow rates rarely exceed 1 L/min⁵¹. New developments allowing the processing of thousands of litres of water, such as the SALSA in situ pump presented here, its equivalent single-sample version⁵², or tow net methods developed for lentic ecosystems⁵³, improve the

detection sensitivity for metazoan taxa in low biomass environments and will allow for more comprehensive and reliable surveys.

With protists and bacteria, taxonomic structures recovered by each sampling method clearly changed with targeted size class (Fig. 3, Supplementary Fig. S2 online). Most protistan micro- to mesoplankton were better detected by the in situ pump (e.g., diatoms, phaeodareans, Acantharea, Ciliophora), while pico- to nanoplankton were preferentially targeted by the sampling box (e.g., Haptophyta, Telonemia), with many groups identified mostly by the smallest size fraction (0.2–2 μm , Chlorophyta, Choanoflagellida, Picozoa, Chrysophyceae, MAST). For bacteria, groups known to occur in aggregates, on larger particles, or as symbionts of larger organisms were better recovered by the in situ pump (e.g., Actinobacteria, Bacteroidetes, Delta-, Gammaproteobacteria, Lentisphaerae, Firmicutes, Tenericutes), while other, likely free-living, bacterioplankton were predominant in the sampling box samples (e.g., Cyanobacteria, Marinimicrobia). This differential taxon recovery of water collection methods has already been reported in shallower studies²⁴, and highlights the importance of targeting the 0.2–2 μm for accurately surveying microbial diversity.

Although the SALSA prototype presented here (Supplementary Fig. S1 online) has since been improved to pump through a 5- μm nylon mesh, in situ filtration techniques are inherently limited by mesh size in order to filter large volumes of water. Thus, although targeting large volumes such as the ones allowed by SALSA represents the most suitable strategy for assessing metazoan diversity in deep-sea waters, its limitation in terms of mesh size leads to the detection of only a fraction of microbial diversity, i.e. mostly larger planktonic groups or taxa fixed on larger faunal specimens or mineral particles. On board filtration of smaller volumes of water remains necessary to access the pico- and nanoplankton, highlighting that both sampling methods are complementary and should be deployed in parallel for integrative biodiversity surveys across the tree of life.

Overall, this comparative study contributes to more comprehensive and more reliable assessments of metazoan and microbial deep-sea communities based on eDNA metabarcoding. First, only sediment samples can allow the characterization of benthic taxa and aboveground water samples do not provide a good alternative. Second, sieving sediment leads to an improvement in taxon detection for metazoans, but as expected, also modifies the retrieved community composition for protists and prokaryotes. Thus, for studies targeting only metazoans, it is advisable to first separate the organisms from the sediment particles using sieving, elutriation, or density extraction techniques as recommended by Brannock & Halanych³⁸. If both metazoan and microbial communities are targeted, and provided sample volume is large enough, an ideal sampling design would be to use multiple sub-samples for microbial taxa and size-sort the remaining sediment for detecting metazoans, as suggested by Nascimento et al.⁵⁴. Alternatively, as shown here, using sufficient volumes of unsorted sediment seems to be satisfactory for integrative biodiversity studies across taxonomic compartments. Finally, water sample volume and mesh size need to be carefully chosen depending on taxa of interest, and while volumes collected by sampling boxes (or Niskin bottles) allow the surveying of microbial diversity, much larger volumes are needed to detect deep-sea metazoans.

Methods

Sample collection. Sediment cores and water samples were collected from a continental slope site during the EssNaut16 cruise in the Mediterranean in April 2016 (Supplementary Table S1 online). Sampling was carried out with a human operated vehicle (Nautile, Ifremer). Triplicate tube cores were collected at the sampling site, and the upper first centimetre sediment layer was used to compare two sediment sampling methods. The sediment samples were either (1) transferred into zip-lock bags and frozen at $-80\text{ }^{\circ}\text{C}$ on board or (2) sieved through five different mesh sizes (1000 μm , 500 μm , 250 μm , 40 μm , and 20 μm) in order to concentrate organisms and separate them by size-class. Sieving was performed with cold surface water filtered at 0.2 μm . Each mesh concentrate was subsequently stored in a separate zip-lock bag and frozen at $-80\text{ }^{\circ}\text{C}$. All samples were shipped on dry ice to the laboratory.

Two different aboveground water-sampling methods were evaluated during EssNaut16 to target microbial and metazoan taxa. All water samples were collected at most 1 m above the seafloor. Water was collected with a newly developed in situ pump, the Serial Autonomous Larval Sampler (SALSA), i.e. a McLane WTS-LV sampler adapted by Ifremer, Brest, France to allow replicated sampling. SALSA has a rosette holding five 2.8 L sampling bowls mounted underneath a rotator plate that allows the alignment of each sampling bowl with the water intake, in a pre-programmed sequential fashion (Supplementary Fig. S1 online). The pump is placed downstream the sampling bowls and the outlet of each bowl is equipped with a 20- μm nylon mesh, retaining particles larger than 20 μm within the bowl while the water passes through the outlet. SALSA thus allows to obtain a time-series of five samples, each resulting from a 4 h filtration event that concentrates particles from ~ 6000 L of water (depending on the filtration rate applied) at the exact same location. Here, 2 samples of each time series were used while the remaining samples were sorted for classical morphological diversity assessments. Two SALSA deployments were performed at the study site (PL07, PL11) and one deployment within the same habitat but at shallower depth due to technical reasons impeding deployment at the original site (PL09). Analyses were performed with and without PL09, and as results were comparable, PL09 was included in the study. For each SALSA deployment, particles retained in every sample (i.e. 2.8 L sampling bowl containing $> 20\text{ }\mu\text{m}$ particles retained during in situ filtration) were concentrated on board on a polycarbonate filter membrane with 2- μm mesh size (Millipore, Burlington, MA, USA, ref. TTTP04700). Water was also collected using two 7.5 L Nautile-deployed sterile and watertight sampling boxes³⁰. These samples were filtered on board successively through membrane filters with 20 μm , 2 μm , and 0.2 μm mesh size (Millipore, Burlington, MA, USA, refs. NY2004700, TTTP04700, GTTP04700), generating three size fractions ($> 20\text{ }\mu\text{m}$, 2–20 μm , and 0.2–2 μm). Each water filter was stored in an individual Petri dish, frozen at $-80\text{ }^{\circ}\text{C}$, and shipped on dry ice to the laboratory.

Nucleic acid extractions. For sediment, DNA extractions were performed using 2–10 g of sediment (Supplementary Table S1 online) with the PowerMax Soil DNA Isolation Kit (MOBIO Laboratories Inc.; Qiagen, Hilden, Germany). All DNA extracts were stored at -80°C . For the non-sieved method, DNA was extracted directly from 2 to 5 g of sediment, the volume varying with the amount of sediment available. For the sieved method, DNA was extracted from each size fraction separately (from 1 to 10 g of sediment per size fraction), and for each of the three cores, an equimolar pool of the DNA extracts of each size fraction was prepared for PCR and sequencing. Thus, for the sieved method, the samples, were based on 3–6 times more sediment than the unsieved raw extracts, however, as some size fractions yielded low DNA concentrations in each core, the sieved samples were at a lower DNA concentration than the unsieved samples (Supplementary Table S1 online). Water DNA extractions were carried out by Genoscope (Évry, France) using the same protocol as described by Alberti et al. (2017)⁵⁵ for Tara Oceans water samples. The protocol is based on cryogenic grinding of membrane filters, followed by nucleic acid extraction with NucleoSpin RNA kits combined with the NucleoSpin DNA buffer set (Macherey–Nagel, Düren, Germany). A negative extraction control was performed alongside sample extractions for both water and sediment samples, adding nothing in the place of sample in the first extraction step.

PCR amplification and sequencing. DNA extracts were normalised to $0.25\text{ ng}/\mu\text{L}$ and $10\ \mu\text{L}$ of standardized sample were used for PCR. Four primer pairs were used to amplify one mitochondrial and three ribosomal RNA (rRNA) barcode loci. The cytochrome c oxidase I (COI)^{56,57} and 18S V1–V2 rRNA⁸ barcodes were used to target metazoans, while 18S V4⁵⁸ was used for unicellular eukaryotes, and 16S V4–V5⁵⁹ for prokaryotes (Supplementary Table S2 online). PCR amplifications for each locus (see supplemental information for amplification and purification details) were carried out in triplicate in order to level-off intra-sample variance while obtaining sufficient amounts of amplicons for Illumina sequencing. PCR triplicates were pooled, purified, quality checked and quantified, and 100 ng of amplicons were directly end-repaired, A-tailed, and ligated to Illumina adapters on a Biomek FX Laboratory Automation Workstation (Beckman Coulter, Brea, CA, USA). Library amplification was performed using a Kapa HiFi HotStart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA, USA), with the same cycling conditions applied for all metagenomic libraries, and libraries were purified using 1X AMPure XP beads (Beckman Coulter, Brea, CA, USA). They were then quantified by Quant-iT dsDNA HS assay kits using a Fluoroskan Ascent microplate fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and sample amplicon libraries were pooled equimolarly. The pools were then quantified 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). Library profiles were assessed using a high-throughput microfluidic capillary electrophoresis system (LabChip GX, Perkin Elmer, Waltham, MA, USA). Libraries normalised to $8\text{--}9\text{ pM}$ and containing a 20% PhiX spike-in were sequenced individually on HiSeq2500 (System User Guide Part # 15035786) instruments in a 250 bp paired-end mode. For sediment, this procedure was carried out on two DNA aliquots of each sample (for each core, 2 aliquots of not sieved extract and 2 aliquots of “sieved pool”), leading to two amplicon libraries per sample. For water collected with the sampling box, we targeted and sequenced microbial loci in the $0.2\text{--}2$ and $2\text{--}20\ \mu\text{m}$ fractions, and targeted metazoans in the $2\text{--}20$ and $>20\ \mu\text{m}$ fractions. The size fractions were processed separately but sequencing failed for the $>20\ \mu\text{m}$ fractions with microbial loci, possibly due to the low DNA concentrations of these samples.

Bioinformatic analyses. All bioinformatic analyses were performed using a Unix shell script, available on Gitlab (<https://gitlab.ifremer.fr/abyss-project/>), on a home-based cluster (DATARMOR, Ifremer), and the samples of the present study were analysed in parallel with 12 to 28 other deep-sea water samples for more accurate error correction and LULU filtering. The details of the pipeline, along with specific parameters used for all metabarcoding markers, are given in Supplementary Table S3 online.

Illumina read pairs were corrected with DADA2 v.1.10⁶⁰, following the online tutorial for paired-end data (<https://benjjneb.github.io/dada2/tutorial.html>), delivering inventories of Amplicon Sequence Variants (ASVs). We chose to evaluate unicellular eukaryote and prokaryote diversity at the ASV level due to their lower intraspecific diversity, making ASVs appropriate to study species-level biodiversity patterns in these microbial taxa. Intraspecific diversity being much more pronounced in metazoans than unicellular organisms due to the extremely varying numbers of cells, organelles, pseudogenes (e.g. numts for COI), or ribosomal repeats in their genomes, ASVs reflect metazoan diversity at the intra-species level, which is dependent on the level of intraspecific variation in the genome, known to vary widely among taxa^{61,62}. As we were interested in species-level diversity, we chose to cluster metazoan data. ASVs from COI and 18S V1–V2 were clustered into Operational Taxonomic Units (OTUs) with swarm v2⁶³ using the FROGS pipeline⁶⁴. Swarm v2 is a single-linkage clustering algorithm that aggregates sequences iteratively and locally around seed sequences based on d , the number of nucleotide differences, to determine coherent groups of sequences. This avoids a universal clustering threshold, which is particularly useful in highly biodiverse samples such as those analysed in this study. Metazoan ASVs were swarm-clustered at $d = 3$ for 18S V1–V2 ($\sim 99\%$) and $d = 6$ for COI ($\sim 98\%$), which has been shown to be appropriate for evaluating species diversity in samples⁶⁵.

Clusters were taxonomically assigned with BLAST+ (v2.6.0) based on minimum similarity (70%) and minimum coverage (80%). For ASVs, sequences obtained with DADA2 were subsequently assigned with *blastn*. For OTUs, BLAST assignment was performed in FROGS using the *affiliation_OTU.py* command. The Silva132 reference database was used for taxonomic assignment of the 16S V4–V5 and 18S V1–V2 rRNA marker genes⁶⁶, PR2 v4.11⁶⁷ was used for 18S V4, and MIDORI-UNIQUE⁶⁸ reduced to marine taxa only was used for COI. An initial test implementing BLAST+ to assign taxonomy only to the COI dataset using a 96% percent identity threshold led to the exclusion of the majority of the clusters. Indeed, it is not uncommon for deep-sea taxa to have closest relatives in databases (even congeners) exhibiting nucleotide divergences of 20%^{69,70}. Considering our interest in

diverse and poorly characterized communities, more stringent BLAST thresholds were not implemented at this stage. However, additional filters were performed during downstream bioinformatic processing described below.

Molecular inventories were refined in R v.3.5.1⁷¹. A blank correction was made using the *decontam* package v.1.2.1⁷², removing all clusters that were more prevalent in negative control samples (PCR and extraction controls) than in true samples. After comparison, results from the technical duplicates produced for sediment samples were merged and read counts were summed for identical OTUs. Fully unassigned clusters were removed (COI: 63%, 18S V1–V2: 20%, 18S V4: 17%, 16S: 5%). When present, non-target clusters were removed (protists or fungi in 18S V1–V2: 60%; metazoans or plants in 18S V4: 10.5%). Additionally, for COI and 18S V1–V2, all metazoan OTUs with a terrestrial assignment (groups known to be terrestrial-only) were removed (COI: 1.5%, 18S: 0.15%). Samples were checked to ensure they had more than 10,000 target reads. Metazoan OTU tables were further curated with LULU v.0.1⁷³ to limit bias due to intraspecific variation and pseudogenes, using a minimum co-occurrence of 0.95, a minimum match at 84%, and a minimum ratio at 1000, which is more appropriate for sample-poor datasets⁶².

Statistical analyses. Data were analysed using R with the packages *phyloseq* v1.22.3⁷⁴, following guidelines in online tutorials (<http://joey711.github.io/phyloseq/tutorials-index.html>), and *vegan* v2.5.2⁷⁵. Read and cluster abundances were evaluated via analyses of variance (ANOVA) on generalised linear models using quasi-poisson distributions. Pairwise post-hoc comparisons were performed via Tukey HSD tests using the *emmeans* package. Alpha and beta diversity were compared among sampling methods using datasets rarefied to the minimum sequencing depth (COI: 62,660; 18S V1: 127,044; 18S V4: 37,000; 16S: 100,952). For comparisons by phylum, paired Welch's t-tests were performed for comparing both sediment methods, and unpaired t-tests were performed for other comparisons. If normality was not verified (Shapiro–Wilk normality test), Wilcoxon (paired) rank tests were performed. Differences in community composition among methods were assessed with Venn diagrams (computed using the *venn* function in the *venn* package) and with permutational multivariate analysis of variance (PERMANOVA), using the *adonis2* function (*vegan*) with significance evaluated using 1000 permutations. Incidence-based Jaccard dissimilarities were used for metazoans, while Bray–Curtis dissimilarities were used for prokaryotes and unicellular eukaryotes. The rationale behind this choice is that metazoans are multicellular organisms of extremely varying numbers of cells, organelles, or ribosomal repeats in their genomes, and can also be detected through a diversity of remains. The number of reads can thus not be expected to reliably reflect the abundance of detected OTUs. Pairwise PERMANOVAs among sampling methods were performed with the *pairwiseAdonis* package. Differences in community structures among samples were visualized via detrended correspondence analyses on rarefied incidence datasets. Finally, taxonomic compositions in terms of cluster abundance were compared among processing methods only using clusters reliably assigned at phylum-level. Phylum-level reliability thresholds were chosen based on Stefanni et al.⁷⁶ and were set at minimum hit identity of 86% for rRNA loci and 80% for COI.

Data availability

The raw data for this work can be accessed in the European Nucleotide Archive database (Study accession numbers: PRJEB37673 for water, PRJEB33873 for sediment). Please refer to the metadata excel file for ENA file names. The dataset, including raw sequences, databases, as well as raw and refined ASV/OTU tables are available on <https://doi.org/10.12770/2deb785a-74c5-4b9d-84d6-82a81e0dda6d>. Bioinformatic scripts can be accessed following the Gitlab link.

Received: 18 November 2020; Accepted: 15 March 2021

Published online: 12 April 2021

References

1. Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L. H. Environmental DNA. *Mol. Ecol.* **21**, 1789–1793 (2012).
2. Rex, M. A. et al. Global bathymetric patterns of standing stock and body size in the deep-sea benthos. *Mar. Ecol. Prog. Ser.* **317**, 1–8 (2006).
3. Snelgrove, P. V. R. Getting to the bottom of Marine biodiversity: sedimentary habitats. *Bioscience* **49**, 129 (1999).
4. Carugati, L., Corinaldesi, C., Dell'Anno, A. & Danovaro, R. Metagenetic tools for the census of marine meiofaunal biodiversity: an overview. *Mar. Genom.* **24**, 11–20 (2015).
5. Grassle, J. F. & Maciulek, N. J. Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. *Am. Nat.* **139**, 313–341 (1992).
6. Smith, C. R. & Snelgrove, P. V. R. *A Riot of Species in an Environmental Calm in 311–342* (CRC Press, 2002). <https://doi.org/10.1201/9780203180594.ch6>.
7. Hauquier, F. et al. Distribution of free-living marine nematodes in the clarion-clipperton zone: implications for future deep-sea mining scenarios. *Biogeosciences* **16**, 3475–3489 (2019).
8. Sinniger, F. et al. Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic knowledge of deep-sea benthos. *Front. Mar. Sci.* **3**, 92 (2016).
9. Boeckner, M. J., Sharma, J. & Proctor, H. C. Revisiting the meiofauna paradox: dispersal and colonization of nematodes and other meiofaunal organisms in low- and high-energy environments. *Hydrobiologia* **624**, 91–106 (2009).
10. Klunder, L. et al. A molecular approach to explore the background Benthic Fauna around a hydrothermal vent and their Larvae: implications for future mining of deep-sea SMS deposits. *Front. Mar. Sci.* **7**, 1–12 (2020).
11. Zhao, F., Filker, S., Xu, K., Huang, P. & Zheng, S. Microeukaryote communities exhibit phyla-specific distance-decay patterns and an intimate link between seawater and sediment habitats in the Western Pacific Ocean. *Deep Res. Part I Oceanogr. Res. Pap.* **160**, 103279 (2020).
12. Cantera, I. et al. Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. *Sci. Rep.* **9**, 1–11 (2019).
13. Forster, D. et al. Benthic protists: the under-charted majority. *FEMS Microbiol. Ecol.* <https://doi.org/10.1093/femsec/fiw120> (2016).

14. Probandt, D. *et al.* Permeability shapes bacterial communities in sublittoral surface sediments. *Environ. Microbiol.* **19**, 1584–1599 (2017).
15. Zinger, L. *et al.* Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* **6**, e24570 (2011).
16. Antich, A. *et al.* Marine biomonitoring with eDNA: can metabarcoding of water samples cut it as a tool for surveying benthic communities?. *Mol. Ecol.* <https://doi.org/10.1111/mec.15641> (2020).
17. Liao, L. *et al.* Microbial diversity in deep-sea sediment from the cobalt-rich crust deposit region in the Pacific Ocean. *FEMS Microbiol. Ecol.* **78**, 565–585 (2011).
18. Bienhold, C., Zinger, L., Boetius, A. & Ramette, A. Diversity and biogeography of bathyal and abyssal seafloor bacteria. *PLoS ONE* **11**, e0148016 (2016).
19. Zhang, J., Sun, Q. L., Zeng, Z. G., Chen, S. & Sun, L. Microbial diversity in the deep-sea sediments of Iheya North and Iheya Ridge, Okinawa Trough. *Microbiol. Res.* **177**, 43–52 (2015).
20. Zhang, L. *et al.* Bacterial and archaeal communities in the deep-sea sediments of inactive hydrothermal vents in the Southwest India Ridge. *Sci. Rep.* **6**, 1–11 (2016).
21. Zhao, F., Filker, S., Stoeck, T. & Xu, K. Ciliate diversity and distribution patterns in the sediments of a seamount and adjacent abyssal plains in the tropical Western Pacific Ocean. *BMC Microbiol.* **17**, 192 (2017).
22. Rodríguez-Martínez, R. *et al.* Controlled sampling of ribosomally active protistan diversity in sediment-surface layers identifies putative players in the marine carbon sink. *ISME J.* **14**, 984–998 (2020).
23. Pernice, M. C. *et al.* Global abundance of planktonic heterotrophic protists in the deep ocean. *ISME J.* **9**, 782–792 (2015).
24. Massana, R. R. *et al.* Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ. Microbiol.* **17**, 4035–4049 (2015).
25. Salazar, G. *et al.* Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* **10**, 596–608 (2016).
26. Díez-Vives, C. *et al.* Delineation of ecologically distinct units of marine Bacteroidetes in the Northwestern Mediterranean Sea. *Mol. Ecol.* **28**, 2846–2859 (2019).
27. Lochte, K. & Turley, C. M. Bacteria and cyanobacteria associated with phytodetritus in the deep sea. *Nature* **333**, 67–69 (1988).
28. Stokke, R. *et al.* Functional interactions among filamentous epsilonproteobacteria and bacteroidetes in a deep-sea hydrothermal vent biofilm. *Environ. Microbiol.* **17**, 4063–4077 (2015).
29. Agusti, S. *et al.* Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nat. Commun.* **6**, 1–8 (2015).
30. Roussel, E. *et al.* Comparison of microbial communities associated with three Atlantic ultramafic hydrothermal systems. *FEMS Microbiol. Ecol.* **77**, 647–665 (2011).
31. Koziol, A. *et al.* Environmental DNA metabarcoding studies are critically affected by substrate selection. *Mol. Ecol. Resour.* **19**, 366–376 (2019).
32. Hajibabaei, M. *et al.* Watered-down biodiversity? A comparison of metabarcoding results from DNA extracted from matched water and bulk tissue biomonitoring samples. *PLoS ONE* **14**, 1–16 (2019).
33. Gleason, J. E., Elbrecht, V., Braukmann, T. W. A., Hanner, R. H. & Cottenie, K. Assessment of stream macroinvertebrate communities with eDNA is not congruent with tissue-based metabarcoding. *Mol. Ecol.* <https://doi.org/10.1111/mec.15597> (2020).
34. Elbrecht, V., Peinert, B. & Leese, F. Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecol. Evol.* **7**, 6918–6926 (2017).
35. Thiel, H. Meiobenthos and nanobenthos of the deep-sea. In *The Sea 8* (ed. Rowe, G. T.) 167–230 (Wiley, 1983).
36. Zeppilli, D. *et al.* Characteristics of meiofauna in extreme marine ecosystems: a review. *Mar. Biodivers.* **48**, 35–71 (2018).
37. Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A. & Hajibabaei, M. Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci. Rep.* **9**, 1–12 (2019).
38. Brannock, P. M. & Halanych, K. M. Meiofaunal community analysis by high-throughput sequencing: comparison of extraction, quality filtering, and clustering methods. *Mar. Genom.* **23**, 67–75 (2015).
39. Burgess, R. An improved protocol for separating meiofauna from sediments using colloidal silica sols. *Mar. Ecol. Prog. Ser.* **214**, 161–165 (2001).
40. Leray, M. & Knowlton, N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proc. Natl. Acad. Sci. USA* **112**, 2076–2081 (2015).
41. Goldberg, C. S. *et al.* Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol. Evol.* **7**, 1299–1307 (2016).
42. Mächler, E., Deiner, K., Spahn, F. & Altermatt, F. Fishing in the water: effect of sampled water volume on environmental DNA-based detection of macroinvertebrates. *Environ. Sci. Technol.* **50**, 305–312 (2016).
43. Lopes, C. M. *et al.* eDNA metabarcoding: a promising method for anuran surveys in highly diverse tropical forests. *Mol. Ecol. Resour.* **17**, 904–914 (2017).
44. Hänfling, B. *et al.* Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Mol. Ecol.* **25**, 3101–3119 (2016).
45. Evans, N. T. *et al.* Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. *Can. J. Fish. Aquat. Sci.* **74**, 1362–1374 (2017).
46. Grey, E. K. *et al.* Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys. *Sci. Rep.* **8**, 8843 (2018).
47. Andruszkiewicz, E. A., Sassoubre, L. M. & Boehm, A. B. Persistence of marine fish environmental DNA and the influence of sunlight. *PLoS ONE* **12**, e0185043 (2017).
48. Dejean, T. *et al.* Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE* **6**, e23398 (2011).
49. Collins, R. A. *et al.* Persistence of environmental DNA in marine systems. *Commun. Biol.* **1**, 185 (2018).
50. Sassoubre, L. M., Yamahara, K. M., Gardner, L. D., Block, B. A. & Boehm, A. B. Quantification of environmental DNA (eDNA) shedding and decay rates for three marine fish. *Environ. Sci. Technol.* **50**, 10456–10464 (2016).
51. Thomas, A. C., Howard, J., Nguyen, P. L., Seimon, T. A. & Goldberg, C. S. ANDe™: a fully integrated environmental DNA sampling system. *Methods Ecol. Evol.* **9**, 1379–1385 (2018).
52. Kersten, O., Vetter, E. W., Jungbluth, M. J., Smith, C. R. & Goetze, E. Larval assemblages over the abyssal plain in the Pacific are highly diverse and spatially patchy. *PeerJ* **2019**, e7691 (2019).
53. Schabacker, J. C. *et al.* Increased eDNA detection sensitivity using a novel high-volume water sampling method. *Environ. DNA* edn3.63 (2020). <https://doi.org/10.1002/edn3.63>.
54. Nascimento, F. J. A., Lallias, D., Bik, H. M. & Creer, S. Sample size effects on the assessment of eukaryotic diversity and community structure in aquatic sediments using high-throughput sequencing. *Sci. Rep.* **8**, 1–12 (2018).
55. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 1–20 (2017).
56. Geller, J., Meyer, C., Parker, M. & Hawk, H. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol. Ecol. Resour.* **13**, 851–861 (2013).
57. Leray, M. *et al.* A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front Zool* **10**, 34 (2013).

58. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **19**, 21–31 (2010).
59. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* **18**, 1403–1414 (2016).
60. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
61. Turon, X., Antich, A., Palacin, C., Præbel, K. & Wangenstein, O. S. From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecol. Appl.* **30**, e02036 (2020).
62. Brandt, M. I. *et al.* Bioinformatic pipelines combining correction and clustering tools allow flexible and comprehensive prokaryotic and eukaryotic metabarcoding. *Rev. Mol. Ecol. Resour.* (2021).
63. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**, e1420 (2015).
64. Escudé, F. *et al.* FROGS: find, rapidly, OTUs with galaxy solution. *Bioinformatics* **34**, 1287–1294 (2018).
65. Brandt, M. I. *et al.* A flexible pipeline combining bioinformatic correction tools for prokaryotic and eukaryotic metabarcoding. *bioRxiv* 717355, ver. 3 peer-reviewed *Recomm. by PCI Ecol.* (2020). <https://doi.org/10.1101/717355>.
66. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.* **41**, D590–D596 (2012).
67. Guillou, L. *et al.* The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucl. Acids Res.* **41**, D597–604 (2013).
68. Machida, R. J., Leray, M., Ho, S. L. & Knowlton, N. Data descriptor: metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Sci. Data* **4**, 1–7 (2017).
69. Shank, T. M., Black, M. B., Halanych, K. M., Lutz, R. A. & Vrijenhoek, R. C. Miocene radiation of deep-sea hydrothermal vent shrimp (Caridea: Bresiliidae): evidence from mitochondrial cytochrome oxidase subunit I. *Mol. Phylogenet. Evol.* **13**, 244–254 (1999).
70. Herrera, S., Watanabe, H. & Shank, T. M. Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Mol. Ecol.* **24**, 673–689 (2015).
71. R Core Team. R: A Language and Environment for Statistical Computing (2018).
72. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
73. Froslev, T. G. *et al.* Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.* **8**, 1–11 (2017).
74. McMurdie, P. J. & Holmes, S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
75. Oksanen, J. *et al.* vegan: Community Ecology Package (2018).
76. Stefanni, S. *et al.* Multi-marker metabarcoding approach to study mesozooplankton at basin scale. *Sci. Rep.* **8**, 12085 (2018).

Acknowledgements

This work is part of the “*Pourquoi Pas les Abysses?*” project funded by Ifremer, and the Project eDNAbys (AP2016-228) funded by France Génomique (ANR-10-INBS-09) and Genoscope-CEA. This work also received funding from the ‘Investissements d’Avenir’ program OCEANOMICS (ANR-11-BTBR-0008; NH). We wish to thank Laure Quintric and Patrick Durand for bioinformatic support and Stéphane Pesant for help in data management. We also wish to express our gratitude to the crew of the R/V Atalante, to the Nautile crew and operators, and to the participants of the EssNaut16 cruise.

Author contributions

M.B., D.Z., and S.A.-H. designed the study, F.P., M.A.C.-B., V.C.-G. carried out the sampling, M.B., J.P., and C.L.-H. carried out the laboratory work. M.B., B.T., and C.B. performed the bioinformatic analyses. M.B., B.T., and N.H. performed the statistical analyses. M.B., D.Z., and S.A.H. wrote the manuscript. All authors contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-86396-8>.

Correspondence and requests for materials should be addressed to M.I.B. or S.A.-H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

2. An assessment of environmental metabarcoding protocols aiming at favoring contemporary biodiversity in inventories of deep-sea communities

Résumé de l'article en français

Les plaines abyssales représentent plus de 50% de la surface de la Terre et sont un réservoir de biodiversité largement méconnue. Elles sont la cible de pressions croissantes liées aux industries d'extraction de ressources malgré ce manque de connaissances qui rend difficile l'appréhension des conséquences de cette exploitation. Dans ces écosystèmes difficiles d'accès, le métabarcoding à partir d'ADN environnemental est un outil particulièrement utile pour l'étude de la biodiversité et l'évaluation des impacts environnementaux. Cependant, ce type d'analyse pourrait être biaisé par la persistance d'ADN ancien dans les sédiments marins, conduisant à la caractérisation des communautés passées plutôt que contemporaines.

A l'aide de kits d'extraction disponibles dans le commerce, nous nous sommes intéressés à l'effet de cinq méthodes de traitement moléculaire sur les inventaires par métabarcoding environnemental de la biodiversité procaryote (16S), eucaryote unicellulaire (18S-V4) et métazoaire (18S-V1, COI). La taille des fragments archivés d'ADN ancien étant généralement petite, nous avons évalué l'impact du retrait des fragments courts d'ADN par sélection de taille ou re-concentration à l'éthanol sur l'ADN extrait à partir de 10g de sédiments provenant de 5 sites marins profonds. Nous avons également comparé les résultats obtenus en extrayant l'ADN et l'ARN de 2g de sédiments des mêmes sites.

Le retrait des fragments courts d'ADN n'influence les estimations de diversité alpha et bêta dans aucun des compartiments biologiques considérés. Les résultats obtenus pour la co-extraction ADN/ARN confirment également les doutes quant à la possibilité de mieux décrire les communautés vivantes à l'aide d'ARN environnemental. Pour les marqueurs ribosomiaux, l'ARN montre la même ségrégation spatiale des échantillons que l'ADN co-extrait, mais donne

CHAPTER 1

des estimations de la richesse spécifique significativement plus élevées. Ces observations vont dans le sens de l'hypothèse d'une persistance accrue de l'ARN ribosomal dans l'environnement par rapport à l'ARN messager, ainsi que de biais non mesurés liés à la variabilité de la sécrétion d'ARN par les organismes en fonction de leur activité métabolique. Pour le marqueur mitochondrial, l'ARN environnemental détecte une diversité plus faible de métazoaires et décrit moins précisément les relations entre échantillons, ce qui reflète l'instabilité accrue de l'ARN messager. Les résultats indiquent aussi l'importance de l'extraction d'ADN à partir de plus grandes quantités de sédiments (> 10g) pour caractériser efficacement la diversité eucaryote.

Ces comparaisons nous conduisent à préférer l'extraction d'ADN plutôt que d'ARN pour réaliser une étude réaliste, répétable et fiable des communautés benthiques, et confirment que l'utilisation de quantités plus importantes de sédiments lors de l'extraction permet un accès plus complet à la diversité eucaryote benthique. Enfin, elles illustrent l'importance des réplicats biologiques plutôt que techniques pour assurer une description fiable des réalités écologiques.



An Assessment of Environmental Metabarcoding Protocols Aiming at Favoring Contemporary Biodiversity in Inventories of Deep-Sea Communities

Miriam I. Brandt^{1*}, Blandine Trouche², Nicolas Henry^{3,4}, Cathy Liautard-Haag¹, Lois Maignien², Colombar de Vargas^{3,4}, Patrick Wincker^{4,5}, Julie Poulain^{4,5}, Daniela Zeppilli⁶ and Sophie Arnaud-Haond^{1*}

¹ MARBEC, Univ Montpellier, Ifremer, IRD, CNRS, Sète, France, ² Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Univ Brest, CNRS, Ifremer, Plouzané, France, ³ Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M, UMR 7144, Roscoff, France, ⁴ Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, Paris, France, ⁵ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, Evry, France, ⁶ Ifremer, Centre Brest, REM/EEP/LEP, ZI de la Pointe du Diable, CS10070, Plouzané, France

OPEN ACCESS

Edited by:

J. Murray Roberts,
University of Edinburgh,
United Kingdom

Reviewed by:

Olivier Laroche,
University of Hawai'i at Mānoa,
United States
S. Kim Juniper,
University of Victoria, Canada

*Correspondence:

Miriam I. Brandt
miriam.isabelle.brandt@gmail.com
Sophie Arnaud-Haond
sophie.arnaud@ifremer.fr

Specialty section:

This article was submitted to
Deep-Sea Environments and Ecology,
a section of the journal
Frontiers in Marine Science

Received: 25 November 2019

Accepted: 25 March 2020

Published: 08 May 2020

Citation:

Brandt MI, Trouche B, Henry N, Liautard-Haag C, Maignien L, de Vargas C, Wincker P, Poulain J, Zeppilli D and Arnaud-Haond S (2020) An Assessment of Environmental Metabarcoding Protocols Aiming at Favoring Contemporary Biodiversity in Inventories of Deep-Sea Communities. *Front. Mar. Sci.* 7:234. doi: 10.3389/fmars.2020.00234

The abyssal seafloor covers more than 50% of planet Earth and is a large reservoir of still mostly undescribed biodiversity. It is increasingly targeted by resource-extraction industries and yet is drastically understudied. In such remote and hard-to-access ecosystems, environmental DNA (eDNA) metabarcoding is a useful and efficient tool for studying biodiversity and implementing environmental impact assessments. Yet, eDNA analysis outcomes may be biased toward describing past rather than present communities as sediments contain both contemporary and ancient DNA. Using commercially available kits, we investigated the impacts of five molecular processing methods on eDNA metabarcoding biodiversity inventories targeting prokaryotes (16S), unicellular eukaryotes (18S-V4), and metazoans (18S-V1, COI). As the size distribution of ancient DNA is skewed toward small fragments, we evaluated the effect of removing short DNA fragments *via* size selection and ethanol reconcentration using eDNA extracted from 10 g of sediment at five deep-sea sites. We also compare communities revealed by eDNA and environmental RNA (eRNA) co-extracted from ~2 g of sediment at the same sites. Results show that removing short DNA fragments does not affect alpha and beta diversity estimates in any of the biological compartments investigated. Results also confirm doubts regarding the possibility to better describe *live* communities using eRNA. With ribosomal loci, eRNA, while resolving similar spatial patterns than co-extracted eDNA, resulted in significantly higher richness estimates, supporting hypotheses of increased persistence of ribosomal RNA (rRNA) in the environment and unmeasured bias due to overabundance of rRNA and RNA release. With the mitochondrial locus, eRNA detected lower metazoan richness and resolved fewer spatial patterns than co-extracted eDNA, reflecting high messenger RNA lability. Results also highlight the importance of using large amounts of sediment (≥ 10 g) for accurately surveying eukaryotic diversity. We conclude that eDNA should be favored over eRNA

for logistically realistic, repeatable, and reliable surveys and confirm that large sediment samples (≥ 10 g) deliver more complete and accurate assessments of benthic eukaryotic biodiversity and that increasing the number of biological rather than technical replicates is important to infer robust ecological patterns.

Keywords: environmental metabarcoding, RNA versus DNA, extracellular DNA, deep-sea biodiversity, benthic ecology, biomonitoring, method testing

INTRODUCTION

Environmental DNA (eDNA) metabarcoding is an increasingly used tool for biodiversity inventories and ecological surveys. Using high-throughput sequencing (HTS) and bioinformatic processing, it allows the detection or the inventory of target organisms using their DNA directly extracted from soil, water, or air samples (Taberlet et al., 2012a). As it does not require specimen isolation, it represents a practical and efficient tool in large and hard-to-access ecosystems, such as the marine realm. Besides allowing studying various biological compartments simultaneously, metabarcoding is also very effective for detecting diversity of small organisms (microorganisms, meiofauna) largely disregarded in visual biodiversity inventories due to the difficulty of their identification based on morphological features (Carugati et al., 2015).

The deep sea, covering more than 50% of Planet Earth, remains critically understudied, despite being increasingly impacted by anthropogenic activities and targeted by resource-extraction industries (Ramirez-Llodra et al., 2011). The abyssal seafloor is mostly composed of sedimentary habitats containing high numbers of small (< 1 mm) organisms, and characterized by high local and regional diversity (Grassle and Maciolek, 1992; Smith and Snelgrove, 2002). Given the increased time efficiency offered by eDNA metabarcoding and its wide taxonomic applicability, this tool is a good candidate for large-scale biodiversity surveys and environmental impact assessments (EIAs) in the deep-sea biome.

eDNA is a complex mixture of genomic DNA present in living cells, extra-organismal DNA, and extracellular DNA originating from the degradation of organic material and biological secretions (Torti et al., 2015). Extracellular DNA has been shown to be very abundant in marine sediments, representing 50–90% of the total DNA pool (Dell'Anno and Danovaro, 2005; Corinaldesi et al., 2018). However, this extracellular DNA compartment may not only contain DNA from contemporary communities. Indeed, nucleic acids can persist in marine sediments as their degradation rate decreases due to adsorption onto the sediment matrix (Corinaldesi et al., 2008; Torti et al., 2015). Low temperatures, high salt concentrations, and the absence of UV light are additional factors enhancing long-term archiving of DNA in deep-sea sediments (Torti et al., 2015; Nagler et al., 2018). Decreased rates of abiotic DNA decay can thus allow DNA persistence over millennial timescales. Indeed, up to 125,000-year-old ancient DNA (aDNA) has been reported in oxic and anoxic marine sediments at various depths (Boere et al., 2011; Coolen et al., 2013; Lejzerowicz et al., 2013a). As extracellular DNA fragment size depends on its state of

degradation (Nagler et al., 2018 report overall size ranges from 80 to over 20,000 bp), aDNA fragments have generally been reported to be $< 1,000$ bp long (Boere et al., 2011; Coolen et al., 2013; Lejzerowicz et al., 2013a; Lennon et al., 2018). Restricting molecular biodiversity assessments to large DNA fragments may thus allow avoiding the bias of aDNA in biodiversity assessments aiming at describing contemporary communities using eDNA metabarcoding.

Environmental RNA (eRNA) has been viewed as a way to avoid the problem of aDNA in eDNA biodiversity inventories because RNA is only produced by living organisms and quickly degrades when released in the environment due to spontaneous hydrolysis and the abundance of RNases (Torti et al., 2015). Few studies have investigated this in the deep-sea, with contrasting results. Investigating foraminiferal assemblages, Lejzerowicz et al. (2013b) found similar taxonomic compositions with DNA and RNA, although highlighting that RNA is more appropriate for targeting the active community component. Contrastingly, Guardiola et al. (2016) detected marked differences between RNA and DNA inventories for most eukaryotic groups but found that both biomolecules detected similar patterns of ecological differentiation, concluding that “dead” DNA did not blur patterns of community structure. Laroche et al. (2018, 2017) found stronger responses to environmental impact in alpha diversity measured with eRNA, while eDNA was better at detecting effects on community composition. Finally, long-term archived and even fossil RNA were also reported in sediment and soil (Orsi et al., 2013; Cristescu, 2019), casting doubts as to its advantage over DNA to inventory contemporary biodiversity.

The design of a sound environmental metabarcoding protocol to inventory biodiversity on the deep seafloor relies on a better understanding of the potential influence of aDNA on the different taxonomic compartments targeted. Using commercially available kits based on 2 and 10 g of sediment, we studied samples from five deep-sea sites encompassing three different habitats and spanning wide geographic ranges in order to select an optimal protocol to survey contemporary benthic deep-sea communities spanning the tree of life. We analyze eDNA and eRNA extracts *via* metabarcoding, targeting the V4–V5 regions of the 16S ribosomal RNA (rRNA) barcode (Parada et al., 2016) for prokaryotes, the 18S-V4 rRNA barcode region for micro-eukaryotes (Stoeck et al., 2010), and the 18S-V1V2 rRNA (thereafter 18S-V1) and Cytochrome c Oxidase I (COI) barcode markers for metazoans (Leray et al., 2013; Sinniger et al., 2016).

Our objectives were threefold:

- (1) Evaluate the effect of removing short DNA fragments from DNA extracts obtained using a 10-g extraction kit,

- (2) Compare eDNA and eRNA inventories resulting from the same samples *via* a 2-g joint extraction kit, and
- (3) Assess the aforementioned kits in terms of repeatability and suitability for different taxonomic compartments.

MATERIALS AND METHODS

Collection of Samples

Sediment cores were collected from five deep-sea sites from various habitats (mud volcano, seamounts, and an area close to hydrothermal vents; **Supplementary Table S1**). Triplicate tube cores were collected with a multicorer or with a remotely operated vehicle at each sampling site. The sediment cores were sliced into layers, which were transferred into zip-lock bags, homogenized, and frozen at -80°C onboard before being shipped on dry ice to the laboratory. The first layer (0–1 cm) was used for the present analysis. In each sampling series, an empty bag was kept as a field control processed through DNA extraction and sequencing.

Nucleic Acid Extractions and Molecular Treatments

eDNA With the 10-g PowerMax Kit

DNA extractions were performed using ~ 10 g of sediment with the PowerMax Soil DNA Isolation Kit (MO BIO Laboratories, Inc., Qiagen, Hilden, Germany). To increase the DNA yield, the elution buffer was left on the spin filter membrane for 10 min at room temperature before centrifugation. For field controls, the first solution of the kit was poured into the control zip lock before following the usual extraction steps. DNA extracts were stored at -80°C .

Size Selection of eDNA Extracts

Size selection of total eDNA extracted as detailed above from ~ 10 g of sediment was carried out to remove small DNA fragments. NucleoMag NGS Clean-up and Size Select beads (Macherey-Nagel, Düren, Germany) were used at a ratio of $0.5 \times$ for removing DNA fragments $< 1,000$ bp from $500 \mu\text{l}$ of extracted eDNA. The target fragments were eluted from the beads with $100 \mu\text{l}$ elution buffer, and successful size selection was verified by electrophoresis on an Agilent TapeStation using the Genomic DNA High ScreenTape kit (Agilent Technologies, Santa Clara, CA, United States).

Ethanol Reconcentration of eDNA Extracts

A 3.5 ml aliquot of eDNA extracted from ~ 10 g of sediment was reconcentrated with 7 ml of 96% ethanol (EtOH) and $200 \mu\text{l}$ of 5 M sodium chloride (NaCl), according to the guidelines in the *Hints and Troubleshooting Guide* of the PowerMax Soil DNA Isolation Kit. As this protocol does not include any incubation time, it favors large DNA fragments. The DNA pellet was washed with 1 ml 70% EtOH, centrifuged again for 15 min at $2,500 \times g$, and air-dried before being resuspended in $450 \mu\text{l}$ elution buffer.

Joint Environmental DNA/RNA With the 2-g RNeasy PowerSoil Kit

Joint RNA/DNA extractions were performed with the RNA PowerSoil Total RNA Isolation Kit combined with the RNeasy PowerSoil DNA elution kit (MO BIO Laboratories, Inc., Qiagen, Hilden, Germany). Between 3 and 5 g of wet and frozen sediment were used, following the manufacturer's suggestions for marine sediments (**Supplementary Table S2**). Extraction controls were performed alongside sample extractions. The RNA pellet was resuspended in $60 \mu\text{l}$ of RNase/DNase-free water. Extracted RNA was then transcribed to first-strand complementary DNA (cDNA) using the iScript Select cDNA synthesis kit (Bio-Rad Laboratories, CA, United States) with its proprietary random primer mix. Quality control 16S-V4V5 , 18S-V1 , and COI PCRs were performed on the RNA extracts to test for potential DNA contamination.

PCR Amplification and Sequencing

Nucleic acid extracts were normalized to 0.25 ng/ μl , and $10 \mu\text{l}$ of standardized samples were used in PCR. Four primer pairs were used to amplify one mitochondrial and three rRNA barcode loci targeting metazoans (COI, 18S-V1), micro-eukaryotes (18S-V4), and prokaryotes (16S-V4V5 for homogeneity; **Supplementary Table S3**). Two metazoan mock communities (detailed in Brandt et al., 2020) were included for 18S-V1 and COI. For each sample and marker, triplicate amplicon libraries (see Supporting Information for amplification details) were prepared by ligation of Illumina adapters on 100 ng of amplicons following the Kapa HiFi HotStart NGS Library Amplification Kit (Kapa Biosystems, Wilmington, MA, United States). After quantification and quality control, library concentrations were normalized to 10 nM, and 8 – 9 pM of each library containing a 20% PhiX spike-in were sequenced on a HiSeq2500 (System User Guide Part #15035786) instruments in a 250 bp paired-end mode.

Bioinformatic Analyses

All bioinformatic analyses were performed using a Unix shell script (Brandt et al., 2020), available on Gitlab¹, on a home-based cluster (DATARMOR, Ifremer). The details of the pipeline, along with specific parameters used for all metabarcoding markers, are given in **Supplementary Table S4** and in Brandt et al. (2020). Pairs of Illumina reads were corrected with DADA2 v.1.10 (Callahan et al., 2016) following the online tutorial for paired-end data² and delivered inventories of amplicon sequence variants (ASVs). Metazoan data were further clustered into operational taxonomic units (OTUs) with swarm v2, a single-linkage clustering algorithm (Mahé et al., 2015) that aggregates sequences iteratively and locally around seed sequences based on d , the number of nucleotide differences, to determine coherent groups of sequences, independent of amplicon input order, allowing highly scalable and fine-scale clustering. ASVs were swarm clustered at d -values of 4 for 18S-V1 and 6 for COI, using the FROGS pipeline (Escudé et al., 2018).

¹<https://gitlab.ifremer.fr/abyss-project/>

²<https://benjjneb.github.io/dada2/tutorial.html>

We chose to evaluate micro-eukaryote and prokaryote diversity at the ASV level due to its increasing use in the literature (Callahan et al., 2017). Although the use of OTUs may also be justified for microbial diversity depending on study objectives (Brandt et al., 2020), we did not expect an alteration of alpha and beta diversity patterns between ASV and OTU levels for the different molecular treatments investigated. ASVs and OTUs were taxonomically assigned *via* BLAST + (v2.6.0) based on minimum similarity and minimum coverage (-perc_identity 70 and -qcov_hsp 80). For ASVs, sequences obtained with DADA2 were subsequently assigned with *blastn*. For OTUs, BLAST assignment in FROGS was performed using the *affiliation_OTU.py* command. It is not uncommon for deep-sea taxa to have closest relatives in databases (even congenics) exhibiting nucleotide divergence exceeding 20% (Shank et al., 1999; Herrera et al., 2015). Considering our interest in diverse and poorly characterized communities, more stringent BLAST thresholds were thus not implemented at this stage. However, additional filters were performed during downstream bioinformatic processing described below, and taxonomic information was used at phylum level, only when the assignment was deemed reliable at this taxonomic level. The Silva132 reference database was used for taxonomic assignment of rRNA marker genes (Quast et al., 2012), and MIDORI-UNIQUE (Machida et al., 2017) was used for COI.

Molecular inventories were refined in R v.3.5.1 (R Core Team, 2018). A blank correction was made using the *decontam* package v.1.2.1 (Davis et al., 2018), removing all clusters that were more prevalent in negative control samples than in true or mock samples. Unassigned and non-target clusters were removed. Additionally, for metazoan loci, all clusters with a terrestrial assignment (groups known to be terrestrial-only) were removed. Samples with fewer than 10,000 target reads were discarded. We performed an abundance renormalization to remove spurious ASVs/OTUs due to random tag switching (Wangensteen and Turon, 2016). The COI OTU table was further curated with LULU v.0.1 (Frøslev et al., 2017) to limit the bias due to pseudogenes, using a minimum co-occurrence of 0.93 and a minimum similarity threshold of 84%.

Statistical Analyses

Sequence tables were analyzed using R with the packages phyloseq v1.22.3 (McMurdie and Holmes, 2013), following guidelines in online tutorials³, and vegan v2.5.2 (Oksanen et al., 2018). Alpha diversity between molecular processing methods was estimated with the number of observed target clusters in rarefied datasets. Cluster abundances were compared *via* analyses of deviances (ANODEV) on generalized linear mixed models using negative binomial distributions, as the data were overdispersed. Pairwise *post hoc* comparisons were performed *via* Tukey honestly significant difference (HSD) tests using the *emmeans* package.

Homogeneity of multivariate dispersions was evaluated with the *betapart* package v.1.5.1 (Baselga and Orme, 2012), and statistical tests performed on balanced datasets for COI

as dispersions were different between 2- and 10-g datasets (**Supplementary Table S5**). Data were rarefied for metazoans and Hellinger-normalized for microbial data.

Differences in community compositions resulting from molecular processing were evaluated with Mantel tests (Jaccard and Bray-Curtis dissimilarities for metazoan and microbial taxa, respectively; Pearson's product-moment correlation; 1,000 permutations). Permutational multivariate analysis of variance (PERMANOVA) was performed on normalized datasets to evaluate the effect of molecular processing and site on community compositions using the function *adonis2* (vegan) with Jaccard dissimilarities (presence/absence) for metazoan and Bray-Curtis dissimilarities for prokaryotes and micro-eukaryotes. The rationale behind this choice is that metazoans are multicellular organisms of extremely varying numbers of cells, organelles, or ribosomal repeats in their genomes, and can also be detected through a diversity of remains. The number of reads can thus not be expected to reflect the abundance of detected OTUs. Significance was evaluated *via* marginal effects of terms using 10,000 permutations with site as a blocking factor. Pairwise *post hoc* comparisons were performed *via* the *pairwiseAdonis* package, with site as a blocking factor. Differences between samples were visualized *via* principal coordinates analysis (PCoA) based on the abovementioned dissimilarities.

Finally, taxonomic compositions in terms of cluster and read abundance were compared between molecular processing methods. In order to compare accurately phylum-level taxonomic compositions, datasets were subsampled to clusters having a minimum hit identity of 86% for rRNA loci and 80% for COI. These values were chosen as they represent approximate minimum identity for reliable phylum assignment (Stefanni et al., 2018).

RESULTS

High-Throughput Sequencing Results

A total of 70 million 18S-V1 reads, 61 million COI reads, 30 million 18S-V4 reads, and 45 million 16S-V4V5 reads were obtained from four Illumina HiSeq runs of pooled amplicon libraries built from triplicate PCR replicates of 75 sediment samples, two mock communities (for 18S-V1 and COI), three extraction blanks, and two to four PCR negative controls (**Supplementary Table S6**). One to seven sediment samples failed amplification in each dataset. These were always coming from the same sampling sites (MDW-ST117 and MDW-ST38) and predominantly comprised RNA samples (**Supplementary Table S6**). After bioinformatic processing, read numbers were reduced to 44 million for 18S-V1, 45 million for COI, 16 million for 18S-V4, and 24 million for 16S-V4V5 (**Supplementary Table S6**). For eukaryote markers, fewer reads were retained in negative controls (2–64%) than in true or mock samples (49–83%), while the opposite was observed for prokaryotes with 16S-V4V5 (62% of reads retained in control samples against 49–57% in true samples). Negative control samples (extraction and PCR blanks) contained 0.001–0.6% of total processed reads compared to 1.3–1.5% in true samples.

³<http://joey711.github.io/phyloseq/tutorials-index.html>

DNA extracts obtained from the joint DNA/RNA protocol based on the 2-g kit produced fewer eukaryotic reads than DNA extracts from the 10-g kit, while similar yields were obtained for prokaryotes. RNA extracts produced more reads than DNA extracts with the ribosomal loci, while they produced fewer reads with the mitochondrial COI locus (**Supplementary Table S6**).

After data refining, abundance renormalization (Wangsten and Turon, 2016), and LULU curation for COI, the final datasets comprised between 8.6 and 16.2 million target reads for eukaryotes and 21.7 million prokaryote reads. Target reads delivered 4,333 and 6,031 metazoan OTUs for COI and 18S-V1 respectively, 40,868 micro-eukaryote 18S-V4 ASVs, and 138,478 prokaryote 16S-V4V5 ASVs (**Supplementary Table S6**).

Alpha Diversity Between Processing Methods

Rarefaction curves showed that a plateau was reached for all samples, suggesting an overall sequencing depth adequate to capture the diversity present (**Supplementary Figure S1**). Processing methods significantly affected the number of recovered eukaryote and prokaryote clusters, and significant variability among sites was detected for 18S-V1 for homogeneity and 18S-V4 (**Table 1** and **Supplementary Figure S2**).

Molecular processing designed to remove small DNA fragments (i.e., size selection of DNA to remove fragment <1,000 bp and EtOH reconcentration) did not significantly affect recovered cluster numbers obtained from eDNA extracted from 10 g of sediment for any of the loci investigated (**Figure 1** and **Table 1**; Tukey's HSD multiple comparisons tests, $p > 0.9$).

Extracts based on the 2-g kit resulted in more variability, reflected by greater standard errors in mean recovered cluster numbers (15–26% of the mean for eukaryotes, 7–9% for prokaryotes) than in DNA extracts based on 10 g of sediment (8–11% for eukaryotes, 3–6% for prokaryotes).

DNA extracted using the 2-g kit recovered significantly fewer eukaryotic clusters than extracts based on ~10 g of sediment (**Figure 1** and **Table 1**), a trend consistent across most taxa (**Figure 2**). DNA 2-g extracts recovered an average of 110 ± 16 18S-V1 and 113 ± 27 COI metazoan OTUs per sample compared to 264 ± 26 (18S-V1) and 222 ± 23 (COI) in the DNA 10-g extracts. Similarly, DNA 10-g extracts recovered on average $1,117 \pm 100$ protistan 18S-V4 ASVs per sample compared to 595 ± 109 detected in DNA from the 2-g kit. Contrastingly to eukaryotes, all DNA methods, whether based on ~2 or ~10 g of sediment, resulted in comparable prokaryote ASV numbers detected (**Figures 1, 2** and **Table 1**; $p > 0.8$), ranging from $5,330 \pm 199$ to $5,810 \pm 170$ per sample on average.

The joint RNA/DNA extracts shared 15% (COI) to 25% (18S-V1) of metazoan OTUs, 14% of protistan 18S-V4 ASVs, and 25% of prokaryotic 16S ASVs (**Supplementary Figure S3**). With COI, most unique OTUs were present in DNA extracts (74%) and RNA detected significantly fewer metazoan OTUs than co-extracted DNA (**Figure 1**, 44 ± 12 versus 113 ± 27

respectively), a trend observed in most detected metazoan phyla (**Figure 2**). Contrastingly, with ribosomal loci, most clusters were unique to RNA (56% for 18S-V1, 63% for 18S-V4, 45% for 16S; **Supplementary Figure S3**), which recovered significantly more clusters than co-extracted DNA (**Figure 1** and **Table 1**). For prokaryotes, RNA extracts even detected significantly more ASVs than DNA extracts based on 10 g of sediment (**Table 1** and **Figure 1**), a pattern observed in most prokaryotic clades, except for the Actinobacteria, Nanoarchaeaeota, Omnitrphicaeota, and Thaumarchaeota (**Figure 2**). For 18S-V4 and 18S-V1, RNA detected a cluster richness comparable to DNA 10-g extracts (Tukey's HSD multiple comparisons tests, $p > 0.16$), yet, average cluster numbers per sample were higher in RNA than in DNA 10-g extracts in numerous groups (**Figure 2**).

Effect of Molecular Processing Methods on Beta-Diversity Patterns

PERMANOVA showed that although site was the main source of variation among samples (accounting for 20–57% of variability), significant differences existed among molecular methods in terms of community structure for all loci investigated over and above any variation due to site (**Table 1**). Pairwise comparisons indicated no significant effect of small DNA fragment removal on revealed community composition (**Table 1**), and high and significant correlations in Mantel tests ($r: 0.92\text{--}1.0$, $p = 0.001$) confirmed the minor effect of size selection and EtOH reconcentration. Based on these results, the size-selected and EtOH-reconcentrated DNA data were removed from further analyses, and community structures of the DNA 10-g extracts were compared with those derived from co-extracted DNA/RNA using the 2-g kit.

Pairwise comparisons showed significant differences in community structures between RNA and DNA for all markers analyzed (**Table 1**). Ordinations confirmed the predominant effect of site as the first two PCoA axes mostly resolved spatial effects (**Supplementary Figure S4**) but also revealed that communities detected by RNA differed from those detected by DNA (co-extracted DNA and DNA 10-g), the level of differentiation varying among sites (**Figure 3**).

Pairwise comparisons also indicated significant differences in community structure between DNA extracts from the 2-g and 10-g kits (**Table 1**) possibly due to higher variability among replicate cores in the DNA 2-g method as seen in ordinations (**Figure 3**).

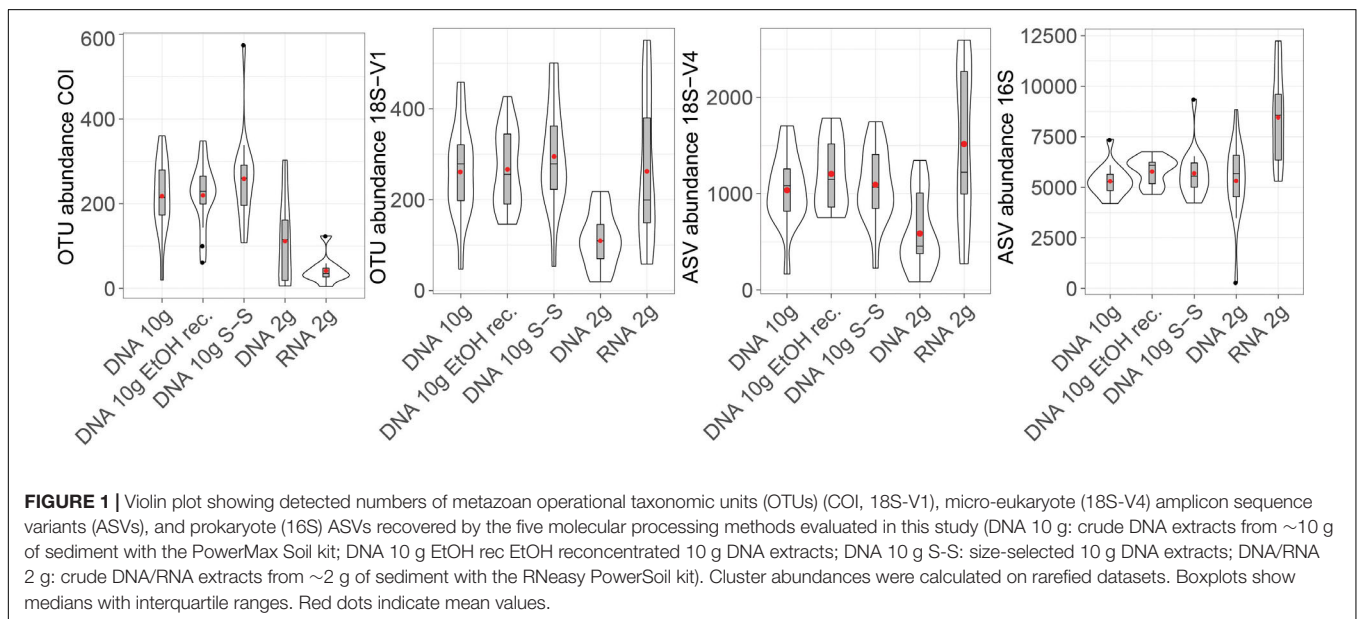
Extraction Kit Versus Nature of Nucleic Acid

PERMANOVA of the dataset containing DNA 10-g, DNA 2-g, and RNA 2-g extracts confirmed that site was the predominant effect, explaining ~20% of variation for metazoans, 33% of variation for micro-eukaryotes, and 54% of variation for prokaryotes. The analysis also indicated that the differences observed between processing methods were predominantly due to the type of nucleic acid rather than the kit used for extraction. Nucleic acid

TABLE 1 | Changes in cluster richness and community structures with molecular processing method (DNA 10 g: DNA extracts from ~10 g of sediment with the PowerMax Soil kit; DNA/RNA 2 g: DNA/RNA extracts from ~2 g of sediment with the RNeasy PowerSoil kit) and site for the four studied genes.

Locus	Cluster richness			Community differentiation			
	Chi-square	<i>p</i> -value	Significant pairwise comparisons		R ²	<i>p</i> -value	Significant pairwise comparisons
18S-V1							
Molecular processing	50.3	<0.001	DNA2 g < RNA 2 g***	Molecular processing	0.06	<0.001	DNA2 g/RNA 2 g***
Site	16.2	<0.001	DNA 10 g > DNA 2 g***	Site	0.23	<0.001	DNA 10 g/DNA 2 g***
				Molecular processing × Site	0.19	0.16	DNA 10 g/RNA 2 g***
COI							
Molecular processing	57.3	<0.001	DNA2 g > RNA 2 g**	Molecular processing	0.09	<0.001	DNA2 g/RNA 2 g**
Site	2.2	0.14	DNA 10 g > DNA 2 g* DNA 10 g > RNA 2 g***	Site	0.20	<0.001	DNA 10 g/DNA 2 g*
				Molecular processing × Site	0.17	0.0013	DNA 10 g/RNA 2 g**
18S-V4							
Molecular processing	38.3	<0.001	DNA2 g < RNA 2 g***	Molecular processing	0.08	<0.001	DNA2 g/RNA 2 g**
Site	15.9	<0.001	DNA 10 g > DNA 2 g**	Site	0.35	<0.001	DNA 10 g/DNA 2 g**
				Molecular processing × Site	0.20	<0.001	DNA 10 g/RNA 2 g**
16S							
Molecular processing	55.0	<0.001	DNA2 g < RNA 2 g***	Molecular processing	0.06	<0.001	DNA2 g/RNA 2 g***
Site	3.4	0.07	DNA 10 g < RNA 2 g***	Site	0.57	<0.001	DNA 10 g/DNA 2 g**
				Molecular processing × Site	0.14	<0.001	DNA 10 g/RNA 2 g***

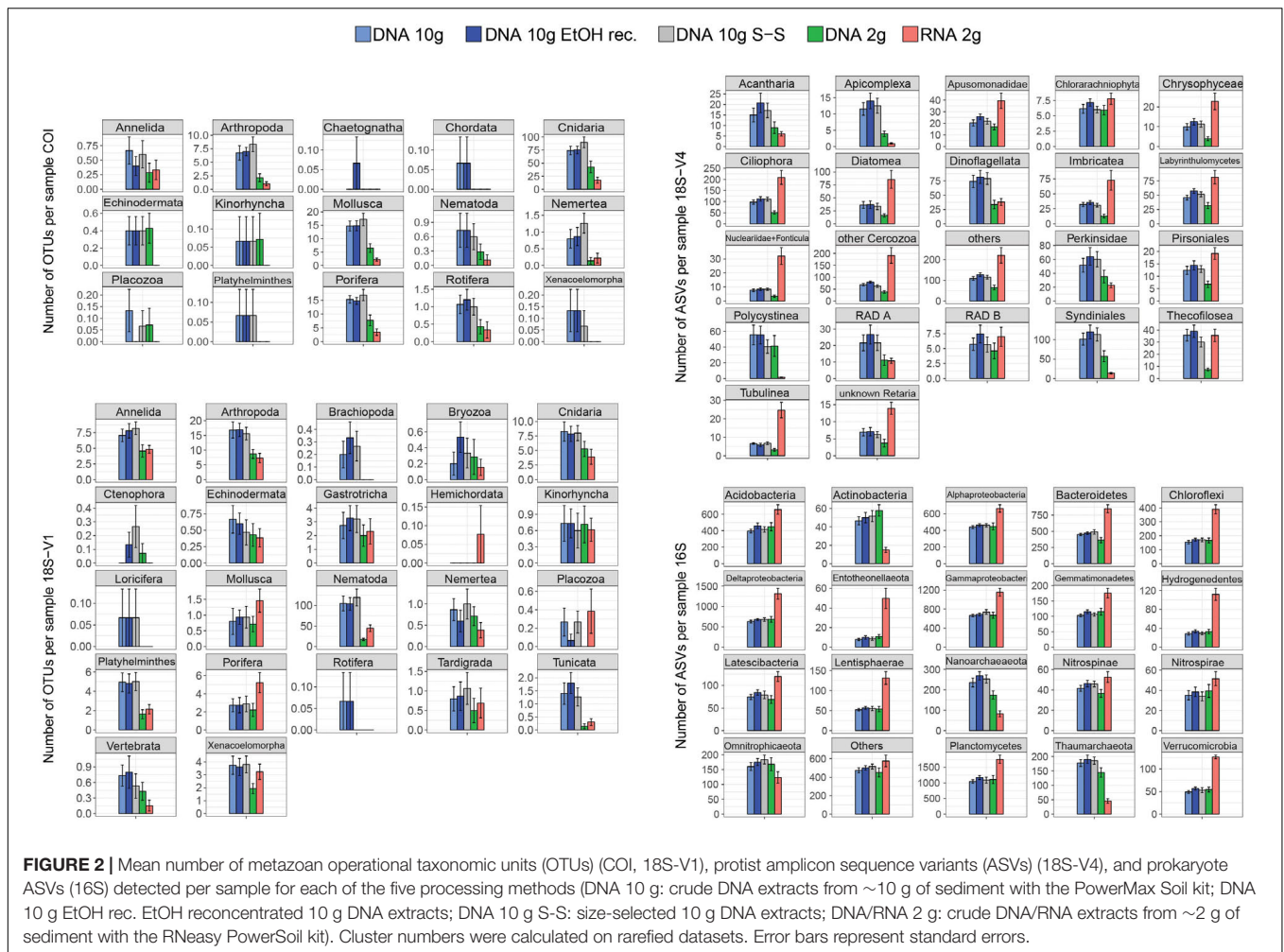
ANODEVs were performed on mixed models with negative binomial distributions using rarefied datasets. PERMANOVAs were calculated on normalized datasets by permuting 10,000 times with Site as a blocking factor using Jaccard dissimilarities for 18S-V1 and COI and Bray-Curtis dissimilarities for 18S-V4 and 16S. Significant *p*-values are in bold. For pairwise comparisons, DNA 10 g comprises all processing methods based on DNA extracted from ~10 g of sediment, and significance codes are ****p* < 0.001; ***p* < 0.01; **p* < 0.05.



nature (DNA versus RNA) led to significant differences among assemblages for all loci, while DNA extraction kit resulted in significant differences only for 18S-V1 and 18S-V4 (Supplementary Table S7).

This supported observations in relative taxonomic compositions, which were more similar between samples based on DNA (Figure 4), a pattern consistent across cores within each site (Supplementary Figure S5). Expectedly, when looking

at read numbers, resolved taxonomic structures were also more similar among DNA-based methods (Supplementary Figure S6). Comparing read and cluster abundances revealed that relative taxonomic compositions based on read numbers (Supplementary Figure S6) were comparable to those based on cluster numbers (Figure 4) for micro-eukaryotes and prokaryotes and confirmed that this was not the case for metazoans.



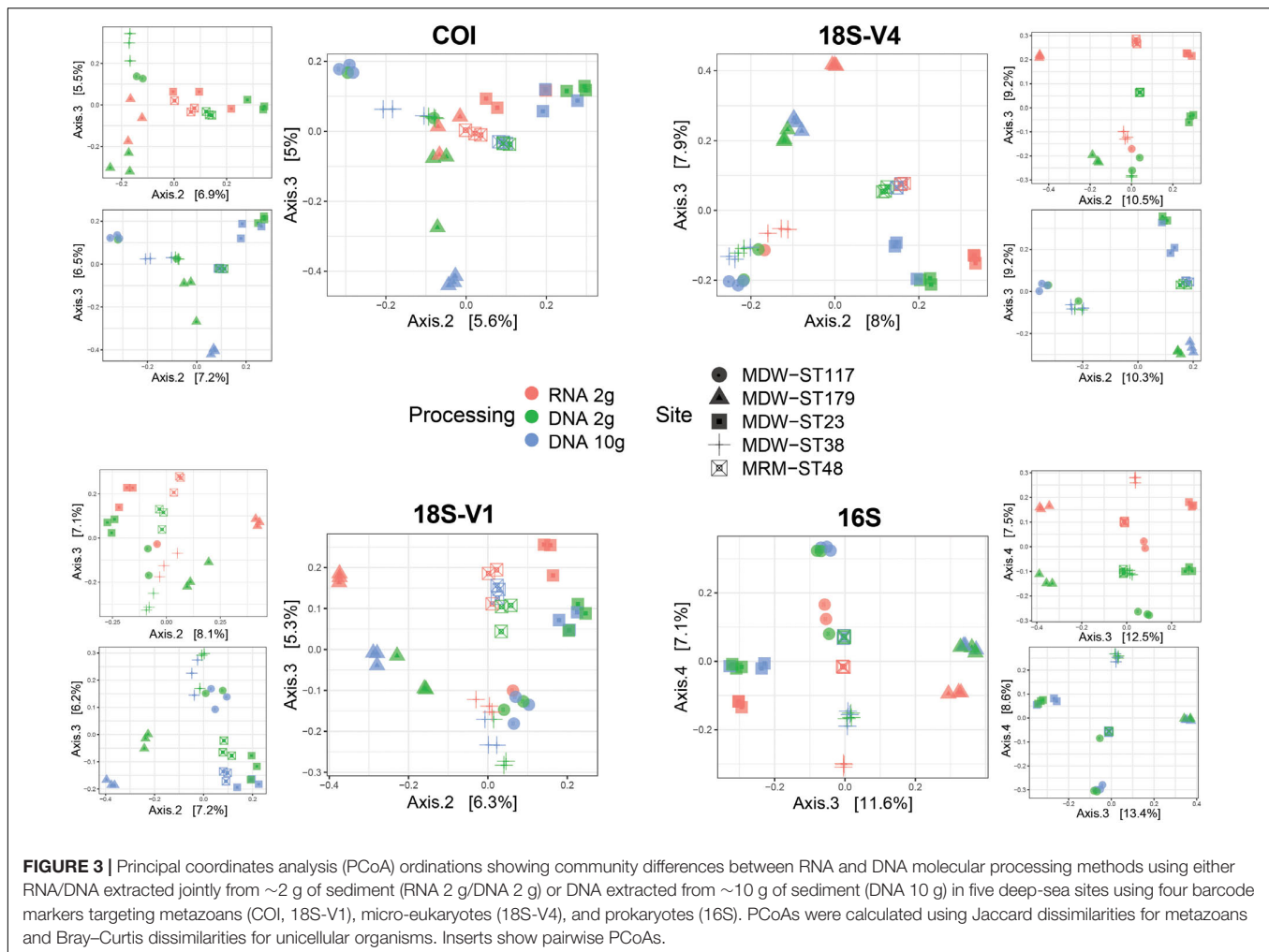
DISCUSSION

The aim of this study was to evaluate different molecular methods in order to select the most appropriate eDNA metabarcoding protocol to inventory contemporary deep-sea communities, with the lowest possible bias due to aDNA.

Using RNA rather than DNA to inventory contemporaneous communities has been suggested as a means of avoiding the bias due to long-term persistence of DNA in marine sediments. Indeed, RNA is only produced by living organisms and is thought to quickly degrade when released in the environment due to spontaneous hydrolysis and the abundance of RNases (Torti et al., 2015). Expectedly, in our COI dataset, RNA resulted in fewer OTUs (Figure 1) and detected fewer phyla (Figure 2) than co-extracted DNA. Contrastingly, for ribosomal loci, RNA detected higher cluster numbers than co-extracted DNA (Figure 1), resulting in more clusters per sample for most of the taxonomic groups detected (Figure 2). In these joint datasets, 45–63% of clusters were unique to RNA (Supplementary Figure S2). These unique clusters were not singleton clusters as only up to 2.2% of them had fewer than three reads, even if 5–28% had fewer than 10 reads

(data not shown). Although proportions vary strongly among investigations, other studies using ribosomal loci have also reported increased recovery of OTUs in RNA datasets as well as considerable amounts of unshared OTUs between joint RNA and DNA data (Guardiola et al., 2016; Laroche et al., 2017, and references therein).

This difference observed here between COI and ribosomal loci is likely related to the nature of the targeted RNA molecule. The rapid hydrolysis of RNA mostly applies to random coils (like messenger RNA), while helical conformations (including most types of RNA, such as ribosomal RNA, transfer RNA, viral genomic RNA, or ribozymes) are less prone to hydrolysis by water molecules (Torti et al., 2015). The degradation of rRNA is thus likely to be much slower than that of messenger RNA (mRNA), which, combined with decreased digestion by RNases due to adsorption onto sediment particles (Torti et al., 2015), makes long-term persistence of rRNA possible and observed in sediments and even in fossils (Orsi et al., 2013; Cristescu, 2019). Finally, the great abundance of RNA over DNA in living organisms (e.g., 20.5% versus 3.1% in *Escherichia coli*) may also favor its persistence in the environment. This is especially true for rRNA, which is represented in a cell's RNA pool as many times



as there are ribosomes, while only being present in a few copies (10–150) in the genome (Torti et al., 2015).

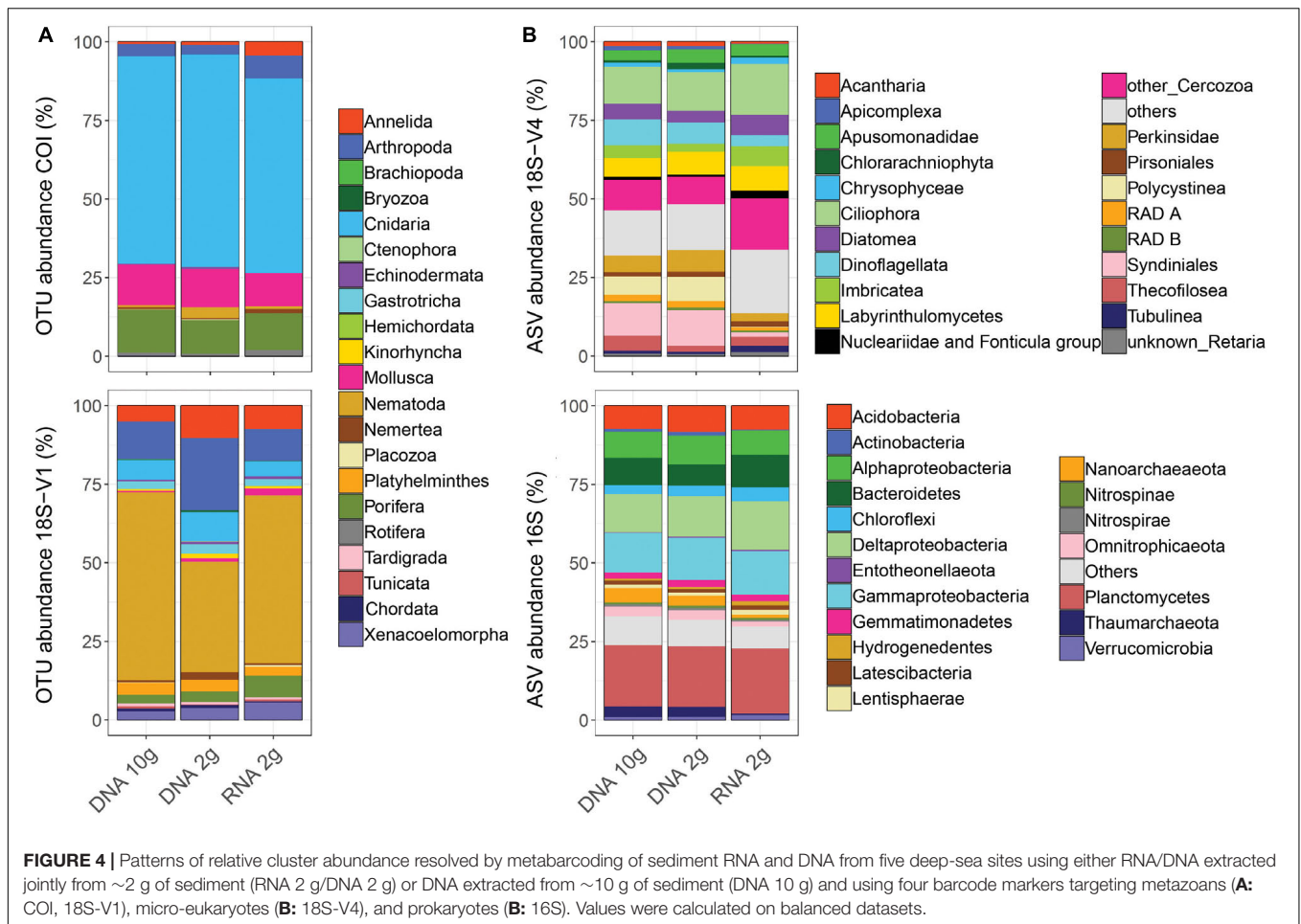
While RNA has been reported as an effective way to depict the active community compartment (Baldrian et al., 2012; Lejzerowicz et al., 2013b; Pawlowski et al., 2014), variation in activity levels between taxonomic groups as well as differences in life histories, life strategies, and non-growth activities may confound this interpretation and generate taxonomic bias (Blazewicz et al., 2013). Instead, DNA/RNA ratios might reflect different genomic architectures (variation in rDNA copy number) among taxonomic groups rather than different relative activities (Massana et al., 2015). Thus, eRNA data need to be interpreted with caution, as some molecular clusters could be overrepresented due to increased cellular activities (Pochon et al., 2017). This could explain the higher cluster numbers detected here for ribosomal loci with eRNA compared to eDNA for several taxa (Figure 2).

Moreover, many of the unique RNA ASVs/OTUs may be artifacts from the reverse transcription of RNA to cDNA, a process known to generate errors that are difficult to measure and detect in bioinformatic analyses (Laroche et al., 2017) but highlighted by the greater amounts of chimeras detected in RNA

extracts with ribosomal loci (Supplementary Table S6). This overestimation of RNA-based data will affect non-clustered data more than clustered datasets, in line with the results observed here for microbial ASVs and metazoan OTUs.

In terms of beta diversity patterns, although RNA and DNA detected significantly different communities (Table 1), DNA and RNA samples resolved similar spatial configurations, with samples clustering by site (Figure 3). This is consistent with Guardiola et al. (2016), who also reported similar patterns of ecological differentiation between DNA and RNA in deep-sea sites, although both datasets resolved different communities. Although the comparative study performed here targeted only the first 1 cm layer of sediment, the comparable results obtained by Guardiola et al. (2016) on 5 cm suggest that these findings may be expanded to deeper layers of sediments. However, spatial variation was more pronounced with DNA samples for eukaryotes, which is congruent with Laroche et al. (2017), who suggested that eDNA may be more reliable for assessing differences in community composition.

Thus, due to its suspected persistence in the environment and the unknown but potentially additional sources of bias suspected here, using eRNA for metabarcoding of deep-sea sediments does



not seem to effectively address the problem of aDNA, and even less so for ribosomal loci. Other studies suggested that a more efficient way to deal with aDNA may be to use joint RNA and DNA datasets and trim for shared OTUs (Laroche et al., 2017; Pochon et al., 2017). This is however particularly stringent (given the low shared OTU proportions observed in this and other studies) and may result in a substantial number of false negatives. With COI, while mRNA may be more effectively targeting living organisms, the approach remains confronted with the taxonomic bias mentioned above, combined with higher *in vitro* lability of mRNA, making it more challenging to work with (highlighted by the increased failure of RNA extracts in this study; **Supplementary Table S6**).

Removing small DNA fragments *via* size selection (removing fragments < 1,000 bp) or EtOH reconcentration did not affect recovered cluster numbers in any of the biological compartments investigated (**Figure 1**). The methods also did not result in any significant difference in community structures (**Table 1**), suggesting that small, likely ancient, DNA fragments have a negligible impact on biodiversity inventories produced through eDNA metabarcoding. This finding is in line with results from the deep-sea (Guardiola et al., 2016; Ramírez et al., 2018) and various other habitats (Lennon et al., 2018), which showed no evidence

that spatial patterns were blurred by “dead” DNA persistence, and suggested a minimal effect of extracellular DNA on estimates of taxonomic and phylogenetic diversity.

None of the methods evaluated in the present study removes DNA not enclosed in *living* cells (e.g., DNA in organelles, DNA from dead cells. . .). It is still unclear how long DNA can remain intracellular after cell death or within organelles. Future research quantifying the rate at which “dead” intracellular DNA becomes extracellular and degraded, and investigation of deeper layers of sediment, will be valuable to estimate the potential bias of archived intracellular DNA in eDNA metabarcoding inventories of extant communities. However, there is increasing evidence that DNA from non-living cells is mostly contemporary (Lennon et al., 2018). This ability to detect extant taxa that were not present in the sample at the time of collection highlights the capacity of eDNA metabarcoding to detect local presence of organisms even from their remains or excretions, and even with a small amount of environmental material.

It remains to be elucidated whether more cost- and time-effective extraction protocols specifically targeting extracellular DNA offer similar ecological resolution as total DNA kits. This is suggested to be the case for terrestrials soils (Taberlet et al., 2012b; Zinger et al., 2016), although authors have highlighted

that conclusions from these studies should be interpreted with caution as results might be influenced by actively released and ancient DNA (Nagler et al., 2018). The only available study testing this in the deep-sea showed that richness patterns were strikingly different in several metazoan phyla between extracellular DNA and total DNA. The authors suggested this to be the result of activity bias: sponges and cnidarians were overrepresented in the extracellular DNA pool because they continuously expel DNA, while nematodes were underrepresented as their cuticles shield DNA (Guardiola et al., 2016). As this comparison was performed on samples collected in two consecutive years, differences observed may partly result from temporal variation. However, another study of shallow and mesobenthic macroinvertebrates showed that targeting solely the extracellular eDNA compartment of marine sediments led to the detection of more than 100 taxa fewer than bulk metabarcoding or morphology, suggesting that extracellular DNA may not be adequate for marine sediments (Aylagas et al., 2016).

Larger amounts of sediment (≥ 10 g) allowed detecting significantly more eukaryotic clusters. This was not true for prokaryotes, for which both ~ 2 and ~ 10 g of sediment detected similar numbers of ASVs (Table 1 and Figure 1). It may be suggested that in the joint RNA/DNA kit, DNA elution occurring after RNA elution induces partial DNA loss. However, such effect would be expected to equally affect eu- and prokaryotes, which was not the case here, supporting the fact that the quantity of the starting material significantly affects results for eukaryotes. The importance of adjusting the amount of starting material to the biological compartment investigated has already been documented (Creer et al., 2016; Dopheide et al., 2019), and this study confirms that while 2–5 g of deep-sea sediment may be enough to capture prokaryote diversity, microbial eukaryotes and metazoans are more effectively surveyed with larger sediment volumes.

Finally, the ~ 2 -g protocols were generally associated with higher variability among replicate cores for all loci investigated (Figures 1, 3). This variability increases confidence intervals, reduces statistical power, and increases the risk of not identifying differences among communities, and thus impacts in EIA studies (Type II errors). Small-scale (centimeters to meters) patchiness has often been reported in the deep-sea (Grassle and Maciolek, 1992; Smith and Snelgrove, 2002; Lejzerowicz et al., 2014). While technical (PCR) replicates allow increasing taxon detection probability (decrease false positives), this within-site variability can only be mitigated by collecting more biological replicates per sampling station and using a sufficiently high amount of starting material to extract nucleic acids.

DATA AVAILABILITY STATEMENT

The data for this work has been submitted to the European Nucleotide Archive (ENA) under the following project: PRJEB33873. Please refer to the sample metadata in **Supplementary Material** to download samples from ENA.

Additionally, the full dataset, including raw sequences, databases, ASV/OTU tables, and scripts (Gitlab link) are available through <https://doi.org/10.12770/cf00aa7b-67e7-49c4-8939-038c4a9d887f>.

AUTHOR CONTRIBUTIONS

MB, CL-H, and SA-H designed the study. MB, JP, and CL-H carried out the laboratory work. MB and BT performed the bioinformatic analyses. MB, BT, and NH performed the statistical analyses. MB and SA-H wrote the manuscript. All authors contributed to the final manuscript.

FUNDING

This work was part of the “*Pourquoi Pas les Abysses?*” project funded by Ifremer and the project eDNabyss (AP2016-228) funded by France Génomique (ANR-10-INBS-09) and Genoscope-CEA. This work also received funding from the European Union’s Horizon 2020 Research and Innovation Program under grant agreement no. 678760 (ATLAS), and the ‘Investissements d’Avenir’ program OCEANOMICS (ANR-11-BTBR-0008; NH and CV). This output reflects only the authors’ view, and the European Union cannot be held responsible for any use that may be made of the information contained therein.

ACKNOWLEDGMENTS

We wish to thank Laure Quintric, Patrick Durand, Caroline Belser, and Stéphane Pesant for bioinformatic and archiving support, as well as Jan Pawlowski and Eva Ramirez Llodra for useful advice on this work. We also wish to express our gratitude to the crew, participants, and mission chiefs of the MarMine cruise (Eva Ramirez Llodra, project 247626/O30, funded by the Research Council of Norway and associated industrial partners) and the MEDWAVES cruise supported by the ATLAS project and the Spanish Ministry of Economy, Industry, and Competitiveness (Covadonga Orejas and all the crew from the Sarmiento de Gamboa), as well as to all the people who helped collecting samples (Perregrino Cambeiro, Juancho Movilla, Maria Rakka, Joana Boavida, Anna Addamo). This manuscript has been released as a Pre-Print at <https://doi.org/10.1101/836080>.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmars.2020.00234/full#supplementary-material>

DATA SHEET S1 | Supplementary material tables (S1 to S7) and figures (S1 to S6).

DATA SHEET S2 | Sample metadata.

REFERENCES

- Aylagas, E., Borja, A., Irigoien, X., and Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. *Front. Mar. Sci.* 3:96. doi: 10.3389/fmars.2016.00096
- Baldrian, P., Kolář, M., Štursová, M., Kopecký, J., Valášková, V., Větrovský, T., et al. (2012). Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *ISME J.* 6, 248–258. doi: 10.1038/ismej.2011.95
- Baselga, A., and Orme, C. D. L. (2012). Betapart: an R package for the study of beta diversity. *Methods Ecol. Evol.* 3, 808–812. doi: 10.1111/j.2041-210X.2012.00224.x
- Blazewicz, S. J., Barnard, R. L., Daly, R. A., and Firestone, M. K. (2013). Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *ISME J.* 7, 2061–2068. doi: 10.1038/ismej.2013.102
- Boere, A. C., Rijpstra, W. I. C., De Lange, G. J., Sinninghe Damsté, J. S., and Coolen, M. J. L. (2011). Preservation potential of ancient plankton DNA in pleistocene marine sediments. *Geobiology* 9, 377–393. doi: 10.1111/j.1472-4669.2011.00290.x
- Brandt, M. I., Trouche, B., Quintric, L., Wincker, P., Poulain, J., and Arnaud-Haond, S. (2020). A flexible pipeline combining clustering and correction tools for prokaryotic and eukaryotic metabarcoding. *bioRxiv*. [Preprint]. doi: 10.1101/717355
- Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869
- Carugati, L., Corinaldesi, C., Dell'Anno, A., and Danovaro, R. (2015). Metagenetic tools for the census of marine meiofaunal biodiversity: an overview. *Mar. Genom.* 24, 11–20. doi: 10.1016/j.margen.2015.04.010
- Coolen, M. J. L., Orsi, W. D., Balkema, C., Quince, C., Harris, K., Sylva, S. P., et al. (2013). Evolution of the plankton paleome in the Black sea from the deglacial to Anthropocene. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8609–8614. doi: 10.1073/pnas.1219283110
- Corinaldesi, C., Beolchini, F., and Dell'Anno, A. (2008). Damage and degradation rates of extracellular DNA in marine sediments: implications for the preservation of gene sequences. *Mol. Ecol.* 17, 3939–3951. doi: 10.1111/j.1365-294X.2008.03880.x
- Corinaldesi, C., Tangherlini, M., Manea, E., and Dell'Anno, A. (2018). Extracellular DNA as a genetic recorder of microbial diversity in benthic deep-sea ecosystems. *Sci. Rep.* 8:1839. doi: 10.1038/s41598-018-20302-7
- Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., et al. (2016). The Ecologist's field guide to sequence-based identification of biodiversity." Edited by Freckleton, R. *Methods Ecol. Evol.* 7, 1008–1018. doi: 10.1111/2041-210X.12574
- Cristescu, M. E. (2019). Can environmental RNA revolutionize biodiversity science? *Trends Ecol. Evol.* 34, 694–697. doi: 10.1016/j.tree.2019.05.003
- Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6:226. doi: 10.1186/s40168-018-0605-2
- Dell'Anno, A., and Danovaro, R. (2005). Ecology: extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science* 309:2179. doi: 10.1126/science.1117475
- Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., and Newcomb, R. D. (2019). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity." Edited by Bunce, M. *Methods Ecol. Evol.* 10, 120–133. doi: 10.1111/2041-210X.13086
- Escudé, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., et al. (2018). FROGS: find, rapidly, OTUs with galaxy solution. Edited by Berger, B. *Bioinformatics* 34, 1287–1294. doi: 10.1093/bioinformatics/btx791
- Froslev, T. G., Kjoller, R., Bruun, H. H., Ejrnaes, R., Brunbjerg, A. K., Pietroni, C., et al. (2017). Algorithm for post-clustering curation of DNA Amplicon data yields reliable biodiversity estimates. *Nat. Commun.* 8:1188. doi: 10.1038/s41467-017-01312-x
- Grassle, J. F., and Maciolek, N. J. (1992). Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. *Am. Nat.* 139, 313–341. doi: 10.1086/285329
- Guardiola, M., Wangensteen, O. S., Taberlet, P., Coissac, E., Uriz, M. J., and Turon, X. (2016). Spatio-temporal monitoring of deep-sea communities using metabarcoding of Sediment DNA and RNA. *PeerJ* 4:e2807. doi: 10.7717/peerj.2807
- Herrera, S., Watanabe, H., and Shank, T. M. (2015). Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Mol. Ecol.* 24, 673–689. doi: 10.1111/mec.13054
- Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lear, G., and Pochon, X. (2018). A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Mar. Pollut. Bull.* 127, 97–107. doi: 10.1016/j.marpolbul.2017.11.042
- Laroche, O., Wood, S. A., Tremblay, L. A., Lear, G., Ellis, J. I., and Pochon, X. (2017). Metabarcoding monitoring analysis: the pros and cons of using co-extracted environmental DNA and RNA data to assess Offshore oil production impacts on benthic communities. *PeerJ* 2017:e3347. doi: 10.7717/peerj.3347
- Lejzerowicz, F., Esling, P., Majewski, W., Szczuciński, W., Decelle, J., Obadia, C., et al. (2013a). Ancient DNA complements microfossil record in deep-sea subsurface sediments. *Biol. Lett.* 9:20130283. doi: 10.1098/rsbl.2013.0283
- Lejzerowicz, F., Voltsky, I., and Pawlowski, J. W. (2013b). Identifying active Foraminifera in the Sea of Japan using metatranscriptomic approach. *Deep Sea Res. Part II Top. Stud. Oceanogr.* 86–87, 214–220. doi: 10.1016/j.dsr2.2012.08.008
- Lejzerowicz, F., Esling, P., and Pawlowski, J. W. (2014). Patchiness of deep-sea benthic foraminifera across the Southern Ocean: insights from high-throughput DNA sequencing. *Deep Sea Res. Part II Top. Stud. Oceanogr.* 108, 17–26. doi: 10.1016/j.dsr2.2014.07.018
- Lennon, J. T., Muscarella, M. E., Placella, S. A., and Lehmkuhl, B. K. (2018). How, when, and where relic DNA affects microbial diversity. *mBio* 9:e00637-18. doi: 10.1128/mBio.00637-18
- Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., et al. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front. Zool.* 10:34. doi: 10.1186/1742-9994-10-34
- Machida, R. J., Leray, M., Ho, S. L., and Knowlton, N. (2017). Data descriptor: metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Sci. Data* 4:170027. doi: 10.1038/sdata.2017.27
- Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution Amplicon clustering. *PeerJ* 3:e1420. doi: 10.7717/peerj.1420
- Massana, R. R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., et al. (2015). Marine protist diversity in European Coastal waters and sediments as revealed by high-throughput sequencing. *Environ. Microbiol.* 17, 4035–4049. doi: 10.1111/1462-2920.12955
- McMurdie, P. J., and Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. Edited by Watson, M. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217
- Nagler, M., Insam, H., Pietramellara, G., and Ascher-Jenull, J. (2018). Extracellular DNA in natural environments: features, relevance and applications. *Appl. Microbiol. Biotechnol.* 102, 6343–6356. doi: 10.1007/s00253-018-9120-4
- Oksanen, J., Blanchet, M., Friendly, G. F., Kindt, R., Legendre, P., McGlenn, D., et al. (2018). *Vegan: Community Ecology Package.* <https://cran.r-project.org/package=vegan>.
- Orsi, W., Biddle, J. F., and Edgcomb, V. (2013). Deep sequencing of seafloor eukaryotic rRNA reveals active fungi across marine subsurface provinces. Edited by López-García, P. *PLoS One* 8:e56335. doi: 10.1371/journal.pone.0056335
- 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. *Environ Microbiol* 18, 1403–1414. doi: 10.1111/1462-2920.13023
- Pawlowski, J. W., Esling, P., Lejzerowicz, F., Cedhagen, T., and Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing

- metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Mol. Ecol. Resour.* 14, 1129–1140. doi: 10.1111/1755-0998.12261
- Pochon, X., Zaiko, A., Fletcher, L. M., Laroche, O., and Wood, S. A. (2017). Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. Edited by Doi, H. *PLoS One* 12:e0187636. doi: 10.1371/journal.pone.0187636
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Ramírez, G. A., Jørgensen, S. L., Zhao, R., and D'Hondt, S. (2018). Minimal influence of extracellular DNA on molecular surveys of marine sedimentary communities. *Front. Microbiol.* 9:2969. doi: 10.3389/fmicb.2018.02969
- Ramírez-Llodra, E., Tyler, P. A., Baker, M. C., Bergstad, O. A., Clark, M. R., Escobar, E., et al. (2011). Man and the last great wilderness: human impact on the Deep sea." Edited by Roopnarine, P. *PLoS One* 6:e22588. doi: 10.1371/journal.pone.0022588
- Shank, T. M., Black, M. B., Halanych, K. M., Lutz, R. A., and Vrijenhoek, R. C. (1999). Miocene radiation of Deep-Sea hydrothermal vent shrimp (Caridea: Bresiliidae): evidence from Mitochondrial Cytochrome Oxidase Subunit I. *Mol. Phylogenet. Evol.* 13, 244–254. doi: 10.1006/mpev.1999.0642
- Sinniger, F., Pawlowski, J. W., Harii, S., Gooday, A. J., Yamamoto, H., Chevaldonné, P., et al. (2016). Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic knowledge of deep-sea benthos. *Front. Mar. Sci.* 3:92. doi: 10.3389/fmars.2016.00092
- Smith, C. R., and Snelgrove, P. (2002). *A Riot of Species in an Environmental Calm*. Boca Raton, FL: CRC Press, 311–342. doi: 10.1201/9780203180594.ch6
- Stefanni, S., Stanković, D., Borme, D., de Olazabal, A., Juretić, T., Pallavicini, A., et al. (2018). Multi-Marker Metabarcoding approach to study Mesozooplankton at basin scale. *Sci. Rep.* 8:12085. doi: 10.1038/s41598-018-30157-7
- Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H. W., et al. (2010). Multiple Marker Parallel tag environmental DNA sequencing reveals a highly complex Eukaryotic community in marine anoxic water. *Mol. Ecol.* 19(Suppl. 1), 21–31. doi: 10.1111/j.1365-294X.2009.04480.x
- Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. (2012a). Environmental DNA. *Mol. Ecol.* 21, 1789–1793. doi: 10.1111/j.1365-294X.2012.05542.x
- Taberlet, P., Prud'Homme, S. M., Campione, E., Roy, J., Miquel, C., Shehzad, W., et al. (2012b). Soil sampling and isolation of Extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Mol. Ecol.* 21, 1816–1820. doi: 10.1111/j.1365-294X.2011.05317.x
- Torti, A., Lever, M. A., and Jørgensen, B. B. (2015). Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Mar. Genom.* 24(Pt 3), 185–196. doi: 10.1016/j.margen.2015.08.007
- Wangenstein, O. S., and Turon, X. (2016). "Metabarcoding techniques for assessing biodiversity of marine animal forests," in *Marine Animal Forests*, eds S. Rossi, L. Bramanti, A. Gori, and C. Orejas Saco del Valle (Cham: Springer International Publishing), 1–29. doi: 10.1007/978-3-319-17001-5_53-1
- Zinger, L., Chave, J., Coissac, E., Iribar, A., Louisanna, E., Manzi, S., et al. (2016). Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biol. Biochem.* 96, 16–19. doi: 10.1016/j.soilbio.2016.01.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Brandt, Trouche, Henry, Liautard-Haag, Maignien, de Vargas, Wincker, Poulain, Zeppilli and Arnaud-Haond. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

3. Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding

Résumé de l'article en français

Le metabarcoding à partir d'ADN environnemental est un outil important pour l'étude de la biodiversité. Cependant, le traitement bioinformatique appliqué doit être adapté à la diversité des compartiments taxonomiques visés ainsi qu'aux spécificités des gènes marqueurs choisis. Nous avons implémenté et testé un pipeline basé sur la correction de séquences avec DADA2 dans le but d'analyser des données de metabarcoding procaryotes (16S) et eucaryotes (18S, COI). Nous avons inclus la possibilité d'agrèger les variants de séquences d'amplicon (ASVs) en unités taxonomiques opérationnelles (OTUs) avec Swarm, algorithme de clustering basé sur l'analyse de réseaux, et l'option de filtrer les ASVs/OTUs avec LULU. Enfin, l'assignation taxonomique est possible en utilisant le classifieur bayésien du Ribosomal Database Project (RDP) et/ou BLAST. Nous avons évalué ce pipeline pour deux marqueurs ribosomaux et un marqueur mitochondrial en utilisant des communautés synthétiques métazoaires et 42 échantillons de sédiments abyssaux.


Les résultats montrent qu'ASVs et OTUs décrivent des niveaux différents de diversité, à choisir suivant les questions scientifiques considérées. Par ailleurs, la complémentarité du clustering et de la filtration avec LULU permet de produire des inventaires de biodiversité métazoaire proches de ceux obtenus en utilisant des critères morphologiques. En effet, le regroupement des séquences efface la variabilité intraspécifique, et d'autre part, LULU retire efficacement les groupes fallacieux issus d'erreurs ou de variabilité intragénomique. L'utilisation de Swarm a un impact différent sur les diversités alpha et bêta mesurées en fonction du gène marqueur. Plus précisément, les valeurs de d supérieures à 1 semblent moins adaptées lors de l'utilisation du 18S pour caractériser les communautés métazoaires.

CHAPTER 1

De même, ajuster le *minimum ratio* de LULU s'est avéré crucial pour éviter de perdre des espèces dans les jeux de données comprenant peu d'échantillons. Enfin, la comparaison des résultats d'assignation de BLAST et RDP a démontré la fiabilité de RDP pour les espèces de l'océan profond, mais a souligné la nécessité d'un effort concerté pour créer des bases de données complètes et spécifiques des écosystèmes considérés.

RESOURCE ARTICLE

Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding

Miriam I. Brandt¹  | Blandine Trouche²  | Laure Quintric³ | Babett Günther¹  | Patrick Wincker^{4,5} | Julie Poulain^{4,5} | Sophie Arnaud-Haond¹ 

¹MARBEC, University of Montpellier, Ifremer, IRD, CNRS, Sète, France

²Laboratoire de Microbiologie des Environnements Extrêmes, University of Brest, Ifremer, CNRS, Plouzané, France

³Cellule Bioinformatique, Ifremer, Plouzané, France

⁴Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université of Evry, Université Paris-Saclay, Evry, France

⁵Research Federation for the study of Global Ocean Systems Ecology and Evolution, Paris, France

Correspondence

Sophie Arnaud-Haond and Miriam I. Brandt, MARBEC, Ifremer, University of Montpellier, IRD, CNRS, Sète, France.
Email: sarnaud@ifremer.fr; miriam.isabelle.brandt@gmail.com

Funding information

Ifremer, Grant/Award Number: AP2016-228; France Génomique, Grant/Award Number: ANR-10-INBS-09; Genoscope-CEA

Abstract

Environmental DNA metabarcoding is a powerful tool for studying biodiversity. However, bioinformatic approaches need to adjust to the diversity of taxonomic compartments targeted as well as to each barcode gene specificities. We built and tested a pipeline based on read correction with DADA2 allowing analysing metabarcoding data from prokaryotic (16S) and eukaryotic (18S, COI) life compartments. We implemented the option to cluster amplicon sequence variants (ASVs) into operational taxonomic units (OTUs) with swarm, a network-based clustering algorithm, and the option to curate ASVs/OTUs using LULU. Finally, taxonomic assignment was implemented via the Ribosomal Database Project Bayesian classifier (RDP) and BLAST. We validated this pipeline with ribosomal and mitochondrial markers using metazoan mock communities and 42 deep-sea sediment samples. The results show that ASVs and OTUs describe different levels of biotic diversity, the choice of which depends on the research questions. They underline the advantages and complementarity of clustering and LULU-curation for producing metazoan biodiversity inventories at a level approaching the one obtained using morphological criteria. While clustering removes intraspecific variation, LULU effectively removes spurious clusters, originating from errors or intragenomic variability. Swarm clustering affected alpha and beta diversity differently depending on genetic marker. Specifically, *d*-values > 1 appeared to be less appropriate with 18S for metazoans. Similarly, increasing LULU's minimum ratio level proved essential to avoid losing species in sample-poor data sets. Comparing BLAST and RDP underlined that accurate assignments of deep-sea species can be obtained with RDP, but highlighted the need for a concerted effort to build comprehensive, ecosystem-specific databases.

KEYWORDS

DADA2, deep-sea biodiversity, LULU, mock communities, multimer metabarcoding, swarm

1 | INTRODUCTION

High-throughput sequencing (HTS) technologies are revolutionizing the way we assess biodiversity. By producing millions of DNA sequences per sample, HTS allows broad taxonomic biodiversity surveys through metabarcoding of bulk DNA from complex communities or from environmental DNA (eDNA) directly extracted from soil, water, and air samples. First developed to unravel cryptic and uncultured prokaryotic diversity, metabarcoding methods have been extended to eukaryotes as powerful, noninvasive tools, allowing detection of a wide range of taxa in a rapid, cost-effective way using a variety of sample types (Creer et al., 2016; Stat et al., 2017; Taberlet et al., 2012; Valentini et al., 2009). In the last decade, these tools have been used to describe past and present biodiversity in terrestrial (Ji et al., 2013; Pansu et al., 2015; Slon et al., 2017; Yoccoz et al., 2012; Yu et al., 2012), freshwater (Bista et al., 2015; Deiner et al., 2016; Dejean et al., 2011; Evans et al., 2016; Valentini et al., 2016), and marine (Bik et al., 2012; Boussarie et al., 2018; Fonseca et al., 2010; Massana et al., 2015; Pawlowski et al., 2011; Salazar et al., 2016; Sinniger et al., 2016; Vargas et al., 2015) environments.

As every new technique brings on new challenges, a number of studies have put considerable effort into delineating critical aspects of metabarcoding protocols to ensure robust and reproducible results (see Figure 1 in Fonseca, 2018). Recent studies have addressed many issues regarding sampling methods (Dickie et al., 2018), contamination risks (Goldberg et al., 2016), DNA extraction protocols (Brannock & Halanych, 2015; Deiner et al., 2015; Zinger et al., 2016), amplification biases and required PCR replication levels for improved detection probability (Alberdi et al., 2017; Ficetola et al., 2015; Nichols et al., 2018). Similarly, computational pipelines, through which molecular data are transformed into ecological inventories of putative taxa, have also been in constant improvement. PCR-generated errors and sequencing errors are major bioinformatic challenges for metabarcoding pipelines, as they can strongly bias biodiversity estimates (Bokulich et al., 2013; Coissac et al., 2012). A variety of tools have thus been developed for quality-filtering amplicon data to remove erroneous reads and improve the reliability of Illumina-sequenced metabarcoding inventories (Bokulich et al., 2013; Eren et al., 2013; Minoche et al., 2011). Studies that evaluated bioinformatic processing steps have generally found that sequence quality-filtering parameters and clustering thresholds most strongly affect molecular biodiversity inventories, resulting in considerable variation during data analysis (Brannock & Halanych, 2015; Brown et al., 2015; Clare et al., 2016; Xiong & Zhan, 2018).

There were historically two main reasons for clustering sequences into operational taxonomic units (OTUs). The first was to limit the bias due to PCR, sequencing errors, and intragenomic variability (e.g., pseudogenes) by clustering erroneous sequences with error-free target sequences. The second was to delineate OTUs as clusters of homologous sequences (by grouping the alleles/haplotypes of a same locus) that would best fit a “morphospecies level”, that is, the OTUs defined using a classical phenetic proxy (Sokal & Crovello, 1970). Recent bioinformatic algorithms alleviate the influence of

errors and intraspecific variability in metabarcoding data sets. First, amplicon-specific error correction methods, commonly used to correct sequences produced by pyrosequencing (Coissac et al., 2012), have now become available for Illumina-sequenced data. Introduced in 2016, DADA2 effectively corrects Illumina sequencing errors and has quickly become a widely used tool, particularly in the microbial world, producing more accurate biodiversity inventories and resolving finescale genetic variation by defining amplicon sequence variants (ASVs) (Callahan et al., 2016; Nearing et al., 2018). Second, LULU is a recently developed curation algorithm designed to filter out spurious clusters, originating from PCR and sequencing errors, or intraindividual variability (pseudogenes, heteroplasmy), based on their similarity (minimum match) and co-occurrence rate (minimum relative cooccurrence) with more abundant clusters, allowing the acquisition of curated data sets while avoiding arbitrary abundance filters (Frøslev et al., 2017). The authors validated their approach on metabarcoding of plants using ITS2 (nuclear ribosomal internal transcribed spacer region 2) and evaluated it on several pipelines. Their results show that ASV definition with DADA2, subsequent clustering to address intraspecific variation, and final curation with LULU is the safest pathway for producing reliable and accurate metabarcoding data. The authors concluded that their validation on plants is relevant to other organism groups and other markers, while recommending future validation of LULU on mock communities as LULU's minimum match parameter may need to be adjusted to less variable marker genes.

The impact of errors being strongly decreased by correction algorithms such as DADA2 and LULU, the relevance of clustering sequences into OTUs is now being debated. Indeed, after presenting their new algorithm on prokaryotic communities, the authors of DADA2 proposed that the reproducibility and comparability of ASVs across studies challenge the need for clustering sequences, as OTUs have the disadvantage of being study-specific and defined using arbitrary thresholds (Callahan et al., 2017). Yet, clustering sequences may still be necessary in metazoan data sets, where very distinct levels of intraspecific polymorphism can exist in the same gene region among taxa, due to both evolutionary and biological specificity (Bucklin et al., 2011; Phillips et al., 2019). ASV-based inventories will thus be biased in favour of taxa with high levels of intraspecific diversity, even though these are not necessarily the most abundant ones (Bazin et al., 2006). Such bias is magnified with presence-absence data, commonly used for metazoan metabarcoding (Ji et al., 2013). However, as intraspecific polymorphism and interspecific divergence are phylum-specific, imposing a universal clustering threshold on metabarcoding data sets is also introducing bias, penalizing groups with lower polymorphism or divergence levels, while overestimating species diversity in groups with higher interspecific divergence. Universal clustering thresholds can be avoided with tools such as swarm v2, a single-linkage clustering algorithm (Mahé et al., 2015), implemented in recent bioinformatic pipelines, such as FROGS (Escudé et al., 2018) or SLIM (Dufresne et al., 2019). Based on network theory, swarm v2 aggregates sequences iteratively and locally around seed sequences, based on d , the number of nucleotide

differences, to determine coherent groups of sequences, independent of amplicon input order, allowing highly scalable and finescale clustering. Finally, it is widely recognized that homogeneous entities sharing a set of evolutionary and ecological properties, namely, *species* (Mayr, 1942; de Queiroz, 2005), sometimes also referred to as “ecotypes” for prokaryotes (Cohan, 2001; Gevers et al., 2005), represent a fundamental category of biological organization that is the cornerstone of most ecological and evolutionary theories and empirical studies. However, maintaining ASV information for feeding databases and cross-comparing studies is not incompatible with their clustering into OTUs, and this choice likely depends on the purpose of the study (for example, providing a census of the extent and distribution of genetic polymorphism for a given gene, or a census of biodiversity to be used and manipulated in ecological or evolutionary studies).

Here, we evaluated DADA2 and LULU, using them alone and in combination with swarm v2, to assess the performance of these new tools for metabarcoding of metazoan communities. Using both mitochondrial COI and the V1–V2 region of the 18S ribosomal RNA (rRNA) gene, we evaluated the need for clustering and the effectiveness of LULU curation to select pipeline parameters delivering the most accurate resolution of two deep-sea mock communities. We then tested the different bioinformatic tools on a deep-sea sediment data set in order to select an optimal trade-off between inflating biodiversity estimates and losing rare biodiversity. As a baseline for comparison, and in the perspective of the joint study of metazoan and microbial taxa, we also analysed the 16S V4–V5 rRNA barcode on these environmental samples.

Our objectives were to (a) discuss the use of ASV versus OTU-centred data sets depending on taxonomic compartment and study objectives, and (b) determine the most adequate swarm-clustering and LULU curation thresholds that avoid inflating biodiversity estimates while retaining rare biodiversity.

2 | MATERIALS AND METHODS

2.1 | Preparation of samples

2.1.1 | Mock communities

Two genomic-DNA mass-balanced metazoan mock communities (5 ng/μl) were prepared using standardized 10 ng/μl DNA extracts of 10 deep-sea specimens belonging to five taxonomic groups (Polychaeta, Crustacea, Anthozoa, Bivalvia, Gastropoda; Table S1). Specimen DNA was extracted using a CTAB extraction protocol, from muscle tissue or from whole polyps in the case of cnidarians. The mock communities differed in terms of ratios of total genomic DNA from each species, with increased dominance of three species and secondary species DNA input decreasing from 3% to 0.7%. We individually barcoded the species present in the mock communities: PCRs of both target genes were performed using the same primers as the ones used in metabarcoding (see below). The PCR reactions

(25 μl final volume) contained 2 μl DNA template with 0.5 μM concentration of each primer, 1x *Phusion* Master Mix, and an additional 1 mM MgCl₂ for COI. PCR amplifications (98°C for 30 s; 40 cycles of 10 s at 98°C, 45 s at 48°C (COI) or 57°C (18S), 30 s at 72°C; and 72°C for 5 min) were cleaned up with ExoSAP (Thermo Fisher Scientific) and sent to Eurofins (Eurofins Scientific) for Sanger sequencing. The barcode sequences obtained for all mock specimens were added to the databases used for taxonomic assignments of metabarcoding data sets, and were submitted on Genbank under accession numbers MN826120–MN826130 and MN844176–MN844185.

2.1.2 | Environmental DNA

Sediment cores were collected from fourteen deep-sea sites ranging from the Arctic to the Mediterranean during various cruises (Table S2). Sampling was carried out with a multicorer or with a remotely operated vehicle. Three tube cores were taken at each sampling station (GPS coordinates in Table S2). The latter were sliced into depth layers that were transferred into zip-lock bags, homogenised, and frozen at –80°C on board before being shipped on dry ice to the laboratory. The first layer (0–1 cm) was used in the present study. DNA extractions were performed using approximately 10 g of sediment with the PowerMax Soil DNA Isolation Kit (Qiagen). To increase the DNA yield, the elution buffer was left on the spin filter membrane for 10 min at room temperature before centrifugation. The ~5 ml extract was then split into three parts, one of which was kept in screw-cap tubes for archiving purposes and stored at –80°C. For the four field controls, the first solution of the kit was poured into the control zip-lock bag, before following the usual extraction steps. For the two negative extraction controls, a blank extraction (adding nothing to the bead tube) was performed alongside sample extractions.

2.2 | Amplicon library construction and high-throughput sequencing

Two primer pairs were used to amplify the mitochondrial COI and the 18S V1–V2 rRNA barcode genes specifically targeting metazoans, and one pair of primer was used to amplify the prokaryote 16S V4–V5 region. PCR amplifications, library preparation, and sequencing were carried out at Genoscope (Evry, France) as part of the eDNAbyss project. Four (16S), eight (18S), and 10 (COI) control PCRs were performed alongside sample PCRs, depending on the amount of trials needed to achieve successful amplification.

2.2.1 | Eukaryotic 18S V1–V2 rRNA gene amplicon generation

Amplifications were performed with the *Phusion* High Fidelity PCR Master Mix with GC buffer (Thermo Fisher Scientific) and the

SSUF04 (5'- GCTTGTCTCAAAGATTAAGCC-3') and SSUR22*mod* (5'- CCTGCTGCCTTCCTTRGA-3') primers (Sinniger et al., 2016), preferentially targeting metazoans, the primary focus of this study. The PCR reactions (25 µl final volume) contained 2.5 ng or less of DNA template with 0.4 µM concentration of each primer, 3% of DMSO, and 1X Phusion Master Mix. Three PCR replicates (98°C for 30 s; 25 cycles of 10 s at 98°C, 30 s at 45°C, 30 s at 72°C; and 72°C for 10 min) were performed in order to smooth the intrasample variance while obtaining sufficient amounts of amplicons for Illumina sequencing.

2.2.2 | Eukaryotic COI gene amplicon generation

Metazoan COI barcodes were generated using the mCOLintF (5'-GGWACWGGWTGAACWGTWTAYCCYCC-3') and jgHCO2198 (5'- TAIACYTCIGGRTGICCRARAAYCA-3') primers (Leray et al., 2013). Triplicate PCR reactions (20 µl final volume) contained 2.5 ng or less of total DNA template with 0.5 µM final concentration of each primer, 3% of DMSO, 0.175 mM final concentration of dNTPs, and 1x Advantage 2 Polymerase Mix (Takara Bio). Cycling conditions included a 10 min denaturation step followed by 16 cycles of 95°C for 10 s, 30 s at 62°C (-1°C per cycle), 68°C for 60 s, followed by 15 cycles of 95°C for 10 s, 30 s at 46°C, 68°C for 60 s and a final extension of 68°C for 7 min.

2.2.3 | Prokaryotic 16S rRNA gene amplicon generation

Prokaryotic barcodes were generated using 515F-Y (5'- GTGYCAGC MGCCGCGGTAA-3') and 926R (5'- CCGYCAATTYMTTTRAGTTT-3') 16S V4-V5 primers (Parada et al., 2016). Triplicate PCR reactions were prepared as described above for 18S V1-V2, but cycling conditions included a 30 s denaturation step followed by 25 cycles of 98°C for 10 s, 53°C for 30 s, 72°C for 30 s, and a final extension of 72°C for 10 min.

2.2.4 | Amplicon library preparation

PCR triplicates were pooled and PCR products purified using 1X AMPure XP beads (Beckman Coulter) clean up. Aliquots of purified amplicons were run on an Agilent Bioanalyser using the DNA High Sensitivity LabChip kit (Agilent Technologies) to check their lengths and quantified with a Qubit fluorimeter (Invitrogen). One hundred nanograms of pooled amplicon triplicates were directly end-repaired, A-tailed and ligated to Illumina adapters on a Biomek FX Laboratory Automation Workstation (Beckman Coulter). Library amplification was performed using a Kapa Hifi HotStart NGS library Amplification kit (Kapa Biosystems) with the same cycling conditions applied for all libraries and purified using 1X AMPure XP beads.

2.2.5 | Sequencing library quality control

Amplicon libraries were quantified by Quant-iT dsDNA HS assay kits using a Fluoroskan Ascent microplate fluorometer (Thermo Fisher Scientific) and then by qPCR with the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems) on an MxPro instrument (Agilent Technologies). Library profiles were assessed using a high-throughput microfluidic capillary electrophoresis system (LabChip GX).

2.2.6 | Sequencing procedures

Amplicon libraries are characterized by low diversity sequences at the beginning of the reads due to the presence of the primer sequence. Low-diversity libraries can interfere in correct cluster identification, resulting in a drastic loss of data output. Therefore, loading concentrations of libraries were decreased to 8–9 pM (instead of 12–14 pM for standard libraries) and PhiX DNA spike-in was set to 20% in order to minimize impacts on run quality. Libraries were sequenced on HiSeq2500 (System User Guide Part # 15035786) instruments (Illumina) in a 250 bp paired-end mode.

2.3 | Bioinformatic analyses

All bioinformatic analyses were performed using a Unix shell script run on a home-based cluster (DATARMOR, Ifremer). The script is available on Gitlab (<https://gitlab.ifremer.fr/abyss-project/>) and is based on DADA2 v.1.10 (Callahan et al., 2016) and FROGS (Escudie et al., 2018) as core processing tools. It allows the use of sequence data obtained from libraries produced by double PCR or adaptor ligation methods, as well as having built-in options for using six commonly used metabarcoding primers.

For all analyses, the mock communities were analysed alongside all environmental samples, and used to validate the metabarcoding pipeline in terms of detection of correct species and presence of false-positives. The details of the pipeline, along with specific parameters used for all three metabarcoding markers are listed in Table S3.

2.3.1 | Reads preprocessing

Our multiplexing strategy relies on ligation of adapters to amplicon pools, meaning that contrary to libraries produced by double PCR, the reads in each paired sequencing run can be forward or reverse. DADA2 correction is based on error distribution differing between R1 and R2 reads. We thus developed a custom script (abyss-preprocessing in abyss-pipeline) allowing separating forward and reverse reads in each paired run and reformatting the outputs to be compatible with DADA2. Briefly, the script uses cutadapt v1.18 to detect and remove primers, while separating forward and reverse

reads in each paired sequence file to produce two pairs of sequence files per sample named R1F/R2R and R2F/R1R. Cutadapt parameters (Table S3) were set to require an overlap over the full length of the primer (default: 3 nt), with 2–4 nt mismatches allowed for ribosomal loci, and 7 nt mismatches allowed for COI (default: 10%). Each identified forward and reverse read is then renamed which the correct extension (/1 and /2 respectively), which is a requirement for DADA2 to recognize the pairs of reads. Each pair of renamed sequence files is then re-paired with BBMAP Repair v38.22 in order to remove singleton reads (nonpaired reads). Optionally, sequence file names can also be renamed if necessary using a CSV correspondence file.

2.3.2 | Read correction, amplicon cluster generation and taxonomic assignment

Pairs of Illumina reads were corrected with DADA2 following the online tutorial for paired-end HiSeq data (https://benjjneb.github.io/dada2/bigdata_paired.html). Reads containing ambiguous bases removed and trimming lengths were adjusted based on sequence quality profiles, so that Q-scores remained above 30 (truncLen at 220 for 18S and 16S, 200 for COI, maxEE at 2, truncQ at 11, maxN at 0). Error model calculation (for R1F/R2R read pairs and then R2F/R1R read pairs), read correction, and read merging was performed at default settings. Amplicons were filtered by size, with size ranges set to 330–390 bp for the 18S SSU rRNA marker gene, 300–326 bp for the COI marker gene, and 350–390 bp for the 16S rRNA marker gene, based on raw size distributions observed. Chimera removal and taxonomic assignment were performed with default methods implemented in DADA2.

A second taxonomic assignment method was optionally implemented in the pipeline, allowing assigning ASVs using basic local alignment search tool BLAST+ (v2.6.0) based on minimum similarity and minimum coverage (-perc_identity 70 and -qcov_hsp 80). An initial test implementing BLAST+ to assign taxonomy only to the COI data set using a 96% percent identity threshold led to the exclusion of the majority of the clusters. Given observed interspecific mitochondrial DNA divergence levels of up to 30% within a same polychaete genus (Zanol et al., 2010) or among some closely related deep-sea shrimp species (Shank et al., 1999), and considering our interest in the identities of multiple, largely unknown taxa in poorly characterized communities, more stringent BLAST thresholds were not implemented at this stage. However, additional filters were performed during downstream processing described below, and only clusters with assignments reliable at phylum-level were retained in the analysis. The Silva132 reference database was used for 16S and 18S SSU rRNA marker genes (Quast et al., 2012), and MIDORI-UNIQUE (Machida et al., 2017) was used for COI. The databases were downloaded from the DADA2 website (<https://benjjneb.github.io/dada2/training.html>) and from the FROGS website (http://genoweb.toulouse.inra.fr/frogs_databanks/assignment). Finally, to evaluate the effect of swarm clustering, ASV tables were clustered with swarm v2 (Mahé et al., 2015) in FROGS (<http://frogs.toulouse.inra.fr/>) at *d*-values (i.e., nucleotide differences) ranging from 1 to 13 ($d = 1, 3, 4, 5, 11$ for 18S/16S, and $d = 1, 5, 6, 7, 13$ for COI), based on settings previously used in the literature (Andújar, Arribas, Gray, et al., 2018; Atienza et al., 2020; Clare et al., 2016; Cordier et al., 2019; Djurhuus et al., 2017; Laroche et al., 2018; Sawaya et al., 2019; Turon et al., 2020; Wood et al., 2019). Resulting OTUs were chimera-filtered and taxonomically assigned via RDP and BLAST+ with the databases stated above, using standard FROGS procedures.

Molecular clusters were refined in R v.3.5.1 (R Core Team, 2018). A blank correction was made using the decontam package v.1.2.1 (Davis et al., 2018), removing all clusters that were prevalent (more frequent) in negative control samples. ASV/OTU tables were refined based on their BLAST or RDP taxonomy. For both assignment methods, clusters unassigned at phylum-level were removed. With BLAST, assigned clusters represented 33% of COI data, 76% of 18S data, and 97% of 16S data. With RDP, assigned clusters represented 95%–99% of data. Nontarget clusters (i.e., either nonmetazoan or nonbacterial) were removed. Additionally, for metazoans, clusters with terrestrial assignments (taxonomic groups known to be terrestrial-only, such as Insecta, Arachnida, Diplopoda, Amphibia, terrestrial mammals, Stylommatophora, Aves, Onychophora, Succineidae, Cyclophoridae, Diplomatiniidae, Megalomastomatidae, Pupinidae, Veronicellidae) were removed. Samples were checked to ensure that a minimum of 10,000 reads were left after refining. Finally, as tag-switching is to be expected in multiplexed metabarcoding analyses (Schnell et al., 2015), an abundance renormalization was performed to remove spurious positive results due to reads assigned to the wrong sample (Wangenstein & Turon, 2016), the original R script being available at https://github.com/metabarpark/R_scripts_metabarpark.

To test LULU curation (Frøslev et al., 2017), refined 18S and COI ASVs/OTUs were curated with LULU v.0.1 following the online tutorial (<https://github.com/tobiasgf/lulu>). The LULU algorithm detects erroneous clusters by comparing their sequence similarity and co-occurrence rate with more abundant (“parent”) clusters. LULU was applied on the full data set (mock and environmental samples) with a minimum relative co-occurrence of 0.95 (default), using a minimum similarity threshold (minimum match) at 84% (default) and slightly higher at 90%, following recommendations of the authors for less variable loci than ITS. The design of the mock samples was not ideal to test LULU, as some mock species were not occurring (or rarely occurring) in environmental samples, but all species were always co-occurring in the mock samples and this at consistent abundance ratios. With the minimum ratio parameter at the default value of 1, this led to the loss of closely related but true mock species for 18S, due to random amplification biases leading to consistent read abundance patterns. In order to remove only errors and avoid losing true mock species, we thus tested minimum ratio at 100 and 1000, which allows removing only clusters that are 100/1,000 times less abundant than a potential parent OTU.

The vast majority of prokaryotes usually show low levels (< 1%) of intragenomic variability for the 16S SSU rRNA gene (Acinas et al., 2004; Pei et al., 2010). These low intragenomic divergence levels can be efficiently removed with swarm clustering at low *d*-values.

The vast majority of prokaryotes usually show low levels (< 1%) of intragenomic variability for the 16S SSU rRNA gene (Acinas et al., 2004; Pei et al., 2010). These low intragenomic divergence levels can be efficiently removed with swarm clustering at low *d*-values.

Although LULU curation may still be useful to merge redundant phylogenies in specific cases such as haplotype network analyses, this was not tested in this study. Indeed, parallelization not being currently available for LULU curation, the richness of prokaryote communities implied unrealistic calculation times, even on a powerful cluster (e.g., LULU curation was at 20%–40% after four days of calculation on our cluster).

In order to have reliable BLAST phylum assignments for pipeline comparison, final data sets were taxonomically filtered by retaining only clusters having a minimum hit identity of 86% for rRNA loci and 80% for COI. These values were chosen as they represent approximate minimum identities for reliable phylum assignment (Stefanni et al., 2018).

2.4 | Statistical analyses

Data was analysed using R with the packages phyloseq v1.22.3 (McMurdie & Holmes, 2013) following guidelines on online tutorials (<http://joey711.github.io/phyloseq/tutorials-index.html>), and vegan v2.5.2 (Oksanen et al., 2018). The data sets were normalized by rarefaction to their common minimum sequencing depth (COI: 15,575; 18S: 33,916; 16S: 70,474), before analysis of mock communities and environmental samples.

To evaluate the functionality of the bioinformatic tools with the mock communities, taxonomically assigned metazoan clusters were considered as derived from one of the 10 species used in the mock communities when the assignment delivered the corresponding species, genus, family, or class. Clusters not fitting the expected taxa were labelled as “Others”. These nontarget clusters may originate from contamination by external DNA or from DNA of associated microfauna, or gut content in the case of whole polyps used for cnidarians.

Alpha diversity detected using each pipeline in the environmental samples was evaluated with the number of observed clusters in the rarefied data sets via analyses of variance (ANOVA) on generalized linear models based on quasipoisson distribution models. Homogeneity of multivariate dispersions were verified with the betadisper function of the betapart package v.1.5.1 (Baselga & Orme, 2012). The effect of site and LULU curation on community composition was tested by PERMANOVA, using the function adonis2 (vegan), with Jaccard incidence dissimilarities for metazoans and Bray-Curtis dissimilarities for prokaryotes, and significance was evaluated by permuting 999 times. Beta-diversity patterns were visualised via non-metric multidimensional scaling (NMDS) using the same dissimilarities stated above.

Finally, BLAST and RDP taxonomic assignments were compared at the most adequate pipeline settings for each locus. BLAST and RDP data sets were compared on ASV-level for prokaryotes, and OTU-level for metazoans (swarm $d = 1$, LULU with minimum match at 84% and minimum ratio at 1 for COI, and 90% and 100, respectively for 18S). As trials on MIDORI-UNIQUE resulted in very poor performance of RDP for COI (assignments belonging mostly to Insecta),

the comparison was performed with MIDORI-UNIQUE subsampled to marine taxa only. For the global data set, full ranges of BLAST hit identities and phylum-level bootstraps were plotted and numbers of clusters left after phylum-level and genus-level quality filtering were calculated, while for evaluation on the mock samples, rarefied data was subsampled to reliable phylum-level assignments (i.e., $\geq 80\%/86\%$ BLAST hit identity, $\geq 80\%$ phylum-level bootstraps).

3 | RESULTS

3.1 | Alpha diversity in mock communities

A total of 1.5 million (COI) and 2 million (18S) raw reads were obtained from the two mock communities (Table S4). After refining (decontamination, renormalisation, removal of nontarget taxa, and clusters unassigned at phylum-level or with unreliable phylum-level assignments), these numbers were decreased to 0.7 million for COI and 1.3 million for 18S.

All 10 mock species were detected in the COI data set (Table 1), even with minimum relative DNA abundance levels as low as 0.7% (Mock 5). With 18S, seven species were recovered and the three bivalve species remained unresolved. Taxonomic assignments were correct at the genus-level for six species with COI and three species for 18S, but all mock species produced ASVs/OTUs correctly assigned up to family or class level. Dominant species generally produced more reads in both the clustered and nonclustered data sets, with the notable exception of the gastropod *Paralepetopsis* sp, which was poorly detected with 18S (Table S5).

When ASVs were clustered with swarm v2, this generally led to a reduction in taxonomic recovery: the two bivalves *P. kilmeri* and *C. regab* were taxonomically misidentified with COI at $d \geq 1$ and *Chorocaris* sp. was not detected with 18S at $d > 1$. Clustering ASVs with swarm v2 reduced the number of molecular clusters produced per species, but some species still produced multiple OTUs even at d values as high as $d = 13$ for COI (*D. dianthus*, *A. muricola*, *Chorocaris* sp., and *Paralepetopsis* sp.) and $d = 11$ for 18S (*A. arbuscula*, *A. muricola*, *Munidopsis* sp., and *E. norvegica*).

Curating ASVs/OTUs with LULU allowed reducing the number of clusters produced per species for both loci, and optimal results were obtained in data sets clustered at $d \geq 1$ for COI and $d = 1$ for 18S. The number of unexpected clusters (“Others”) was hardly affected by LULU curation (Table 1). In the COI data set, curating with LULU at 84% or 90% minimum match resulted in similar OTU numbers, although 84% performed slightly better in Mock 3 (Table 1). Increasing the minimum ratio parameter to 100 or 1,000 resulted in the retention of more error OTUs and thus higher OTU numbers in each mock species (data not shown). For 18S, both LULU minimum match and minimum ratio affected species recovery. LULU curation with minimum ratio = 1 led to the loss of the shrimp *Chorocaris* sp. at both minimum match levels and the gastropod *Paralepetopsis* sp. at 84% minimum match (Table S6). With minimum ratio at 100, *Chorocaris* sp. was retained in the data set

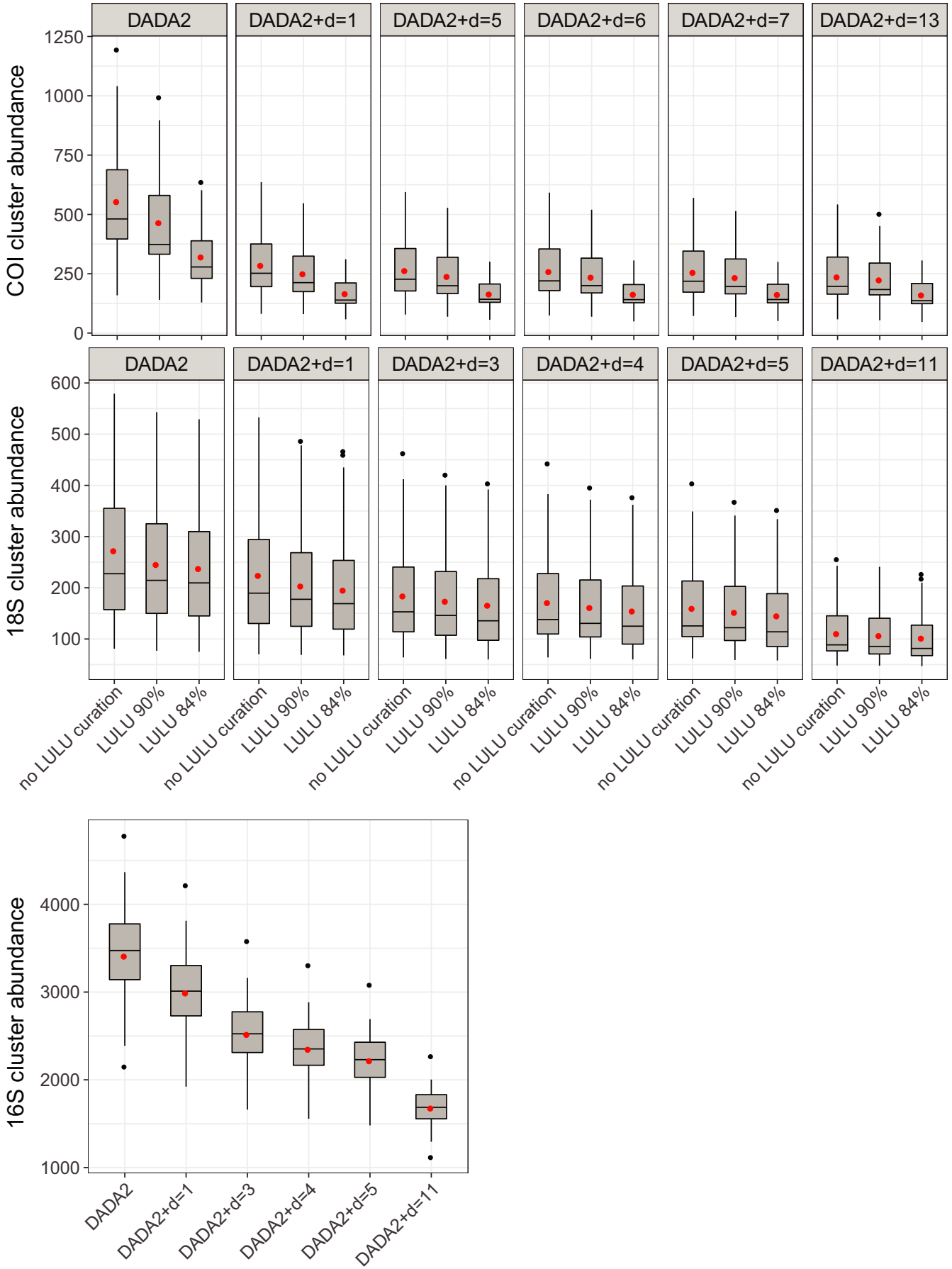


FIGURE 1 Number of metazoan (COI, 18S) and prokaryote (16S) clusters detected in sediment of 14 deep-sea sites with ASV versus OTU-centred data sets. ASVs were obtained with the DADA2 metabarcoding pipeline, and clustered with swarm at different d values. Metazoan ASVs and OTUs were curated with LULU at 84% and 90% minimum match. LULU curation was performed with minimum ratio = 100 for 18S and minimum ratio = 1 for COI. Cluster abundances were obtained from data sets rarefied to same sequencing depth. Boxplots represent medians with first and third quartiles. Red dots indicate means

(Table S4). The final data sets contained ~5 million (COI) to ~8 million (18S) marine metazoan target reads and ~7 million prokaryotic 16S reads (Table S4). COI reads produced 13,397 ASVs, 3,518–5,563 OTUs after swarm clustering ($d = 1$ –13), and 1,758–10,028 OTUs after LULU curation (Table S7). Final 18S reads comprised 8,280 ASVs, 1,869–6,015 OTUs after swarm clustering ($d = 1$ –11), and 1,469–6,909 OTUs after LULU curation. The prokaryote data set produced 53,815 target ASVs and 12,800–38,972 OTUs after swarm clustering ($d = 1$ –11).

3.2.2 | Number of clusters among pipelines

The number of metazoan clusters detected in the deep-sea sediment samples varied significantly with bioinformatic pipeline and site (Table 2). The pipeline effect was consistent across sites (Table 2), although mean cluster numbers detected per sample spanned a wide range in all loci (50–500 for 18S, 100–1,000 for COI, and 1,500–4,000 for 16S, Figure 1).

As expected, clustering significantly reduced the number of detected molecular clusters per sample for all loci. Consistent to results observed in mock communities, clustering at $d = 1$ –13 resulted in comparable OTU numbers for COI, while significantly higher OTU numbers were obtained at $d = 1$ than with $d > 1$ for ribosomal loci (Figure 1, Table 2). DADA2 detected on average 555 (SE = 42) metazoan COI ASVs per sample, and clustering reduced this number to around 250, regardless the d -value. For ribosomal loci, clustering at $d = 3$ –5 reduced OTU numbers of around ~30% compared to without clustering, while at $d = 11$, cluster numbers were more than halved.

LULU curation of ASVs or OTUs decreased the number of COI and 18S clusters detected (Figure 1). This decrease was significant for both ASVs and OTUs with COI, but less marked for 18S as LULU's minimum ratio was set to 100 (Table 2). For COI, where LULU curation was performed with minimum ratio = 1, the minimum match parameter had a strong influence on alpha diversity. Indeed, LULU curation of ASVs or OTUs with minimum match at 90% resulted in significantly more clusters than at 84% (Table 2). In contrast, the magnitude of the minimum match parameter did not significantly affect the number of clusters for 18S, where LULU curation was performed with minimum ratio = 100. LULU curation of ASVs resulted in more OTUs than swarm clustering for both loci, with both minimum match levels tested (Figure 1, Table 2). Similarly, LULU curation of ASVs resulted in significantly more clusters than LULU curation of OTUs produced with any d -value (Figure 1, Table 2).

Looking at mean ASV and OTU numbers detected per phylum with each pipeline showed consistent effects of swarm clustering and LULU curation, but highlighted strong differences in the amount of intragenomic variation between taxonomic groups. For all loci investigated, some taxa displayed high ASV to OTU ratios, while

others were hardly affected by clustering or LULU curation in terms of numbers of clusters detected (Figure S1).

3.3 | Patterns of beta-diversity between pipelines

PERMANOVAs confirmed that sites differed significantly in terms of community structure, accounting from 46% to 89% of variation in data. Evaluating the effect of LULU curation for metazoans showed that LULU-curated data resolved similar community compositions than noncurated data, accounting for < 1% of variation in data (Figure 2).

Although ASV and OTU data sets detected similar amounts of variation due to sites in PERMANOVAs, clustering levels affected the ecological patterns resolved by ordinations in rRNA loci (Figure 2). Metazoan 18S ASVs showed strong segregation by ocean basin, with samples grouped by depth within each basin, and prokaryote ASVs showed both strong segregation by ocean basin and depth (Figure 2). Clustering at d -values >1 decreased differences among deep sites (> 1,000 m) across ocean basins, emphasizing the depth effect over the basin effect. This change in ecological pattern occurred consistently with d -values from 3 to 11 (Figure 2, Figure S2).

3.4 | Taxonomic assignment quality

Assigning with BLAST resulted in mock community assignments comparable to those described above. With COI, eight of the 10 species produced one single OTU, with six correctly assigned at genus-level, and two species were taxonomically correctly assigned only to class-level and produced 2–3 OTUs (Figure S3). With 18S, seven species were recovered (four correctly assigned at genus-level), with two producing more than one OTU, and the three vesicomyid bivalve species were taxonomically unresolved and assigned up to family-level while generating 2 OTUs. Assigning the COI data set with RDP using the MIDORI-UNIQUE database resulted in assignments of the mock samples that did not match the expected taxa and were mostly belonging to arthropods, a problem not observed with BLAST (data not shown). When the database was reduced to marine-only taxa, RDP results were comparable to BLAST, with seven species correctly assigned at genus-level. Assigning the 18S data set with RDP produced results comparable to BLAST, although taxonomic assignments were less accurate for two species.

BLAST and RDP assigned similar amounts of OTUs in the prokaryote data set, but BLAST assigned 20% (18S) and 70% (COI) less OTUs at phylum-level than RDP in the metazoan data sets, even at minimum hit identity of 70% (Table S8). BLAST hit identities of the overall data sets varied strongly depending on phyla and marker gene (Figure 3). For 18S, 90% of metazoan OTUs had assignment identities

TABLE 2 Effect of pipeline and site on the number of metazoan and prokaryote clusters

LOCUS	F-value	p-value	Significant pairwise comparisons
COI			
Pipeline	135.2	<.001	Dada2 > DS***
Site	226.7	<.001	DS(<i>d</i> = 1) > DS(<i>d</i> = 13)**; DS > DSL84%***; D(S)L90% > D(S)L84%***
Pipeline × Site	0.15	>.05	Dada2 > DL***; DL90% > DS(L)***; DL84% > DS(<i>d</i> = 5–13)***; DL > DSL***
18S V1-V2			
Pipeline	67.2	<.001	Dada2 > DS***
Site	263.1	<.001	DS(<i>d</i> = 1) > DS(<i>d</i> = 3–11)***; DS(<i>d</i> = 11) < DS(<i>d</i> = 1–5)***
Pipeline × Site	0.3	>.05	Dada2 > DL84%*; DL > DS(<i>d</i> = 3–11)***; DL > DSL***
16S V4-V5			
Pipeline	188.7	<.001	Dada2 > DS***
Site	18.3	<.001	DS(<i>d</i> = 1) > DS(<i>d</i> = 3–11)***; DS(<i>d</i> = 3) > DS(<i>d</i> = 5)***; DS(<i>d</i> = 11) < DS(<i>d</i> = 1–5)***
Pipeline × Site	0.06	>.05	

Note: Results of the analysis of variance (ANOVA) of the rarefied cluster richness for the three genes studied. Pairwise comparisons were performed with Tukey's HSD tests. DS: Dada2 + swarm; DSL: Dada2 + swarm + LULU; *d*: swarm *d*-value. LULU curation was performed with minimum match at 84% and 90%, and with minimum ratio = 100 for 18S and minimum ratio = 1 for COI. Significance codes: ****p* < .001; ***p* < .01; **p* < .05.

≥ 86%, corresponding roughly to accurate phylum-level (Edgar, 2017; Stefanni et al., 2018). Only 34% had reliable genus-level assignments, for example with > 95% similarity (Table S8). For COI, only 30% of metazoan assignments were reliable at phylum-level (≥ 80%), and only 1% at genus-level (> 93%). BLAST hit identity was much higher for prokaryotes, with 98% of ASVs assigned with ≥ 86% similarity to sequences in databases, and 65% had reliable genus-level assignments (> 95% similarity). With RDP, 77% of metazoan 18S OTUs and 96% of prokaryote 16S ASVs had phylum-level bootstraps ≥ 80%, and 59% and 76% also had genus-level bootstraps ≥ 80%, respectively. For COI, applying a minimum phylum-level bootstrap of 80% resulted in an unviable decrease in the number of target OTUs, as only 242 metazoan OTUs (~1%) remained after filtering, and only 112 (0.3%) with acceptable genus-level bootstraps (Table S8). Indeed, most OTUs, primarily assigned to arthropods, cnidarians, molluscs, vertebrates, and poriferans still had phylum-level bootstraps < 60% (Figure 3).

4 | DISCUSSION

4.1 | ASVs versus OTUs: A choice depending on taxon of interest and research question

ASVs have recently been advocated to replace OTUs “as the standard unit of marker-gene analysis and reporting” (Callahan et al., 2017): an advice for microbiologists that may not apply when studying metazoans. Life histories of organisms, together with intrinsic properties of marker genes, determine the level of intragenomic and intraspecific diversity. Metazoans are well known to exhibit variable

and sometimes very high intraspecific polymorphism. This intraspecific variation is a recognised problem in metabarcoding, known to generate spurious clusters (Brown et al., 2015), especially in the COI barcode marker. Indeed, this gene region has increased intragenomic variation due to its high evolutionary rate (Machida & Knowlton, 2012; Machida et al., 2012), but also due to heteroplasmy and the abundance of pseudogenes, such as NUMTs, playing an important part of the supernumerary OTU richness in COI-metabarcoding (Bensasson et al., 2001; Song et al., 2008). Concerted evolution, a common feature of SSU rRNA markers such as 16S (Hashimoto et al., 2003; Klappenbach et al., 2001) and 18S (Carranza et al., 1996), limits the amount of intragenomic polymorphism. In metazoans, a lower level of diversity is thus expected for 18S than for COI. This is reflected in the lower ASV (DADA2) to OTU (DADA2 + swarm) ratios observed here for 18S (1.4–2.5) compared to COI (2.3–3.2), at clustering *d*-values comprised between one and seven (Table S7), underlining the different influence – and importance – of clustering on these loci, and the need for a versatile, marker by marker choice for clustering parameters.

The results on the mock samples showed that even single individuals produced very different numbers of ASVs, suggesting that ASV-centred data sets do not accurately reflect species composition in metazoans. Intragenomic and intraspecific polymorphism are highly variable across taxa (Plouviez et al., 2009; Teixeira et al., 2013), as confirmed by the very variable decrease in cluster numbers observed with clustering in this study for different phyla (Figure S1). The taxonomic compositions of samples based on ASVs may thus reflect genetic rather than species diversity. This distinction is important to keep in mind, as the species, that is “a lineage or group of

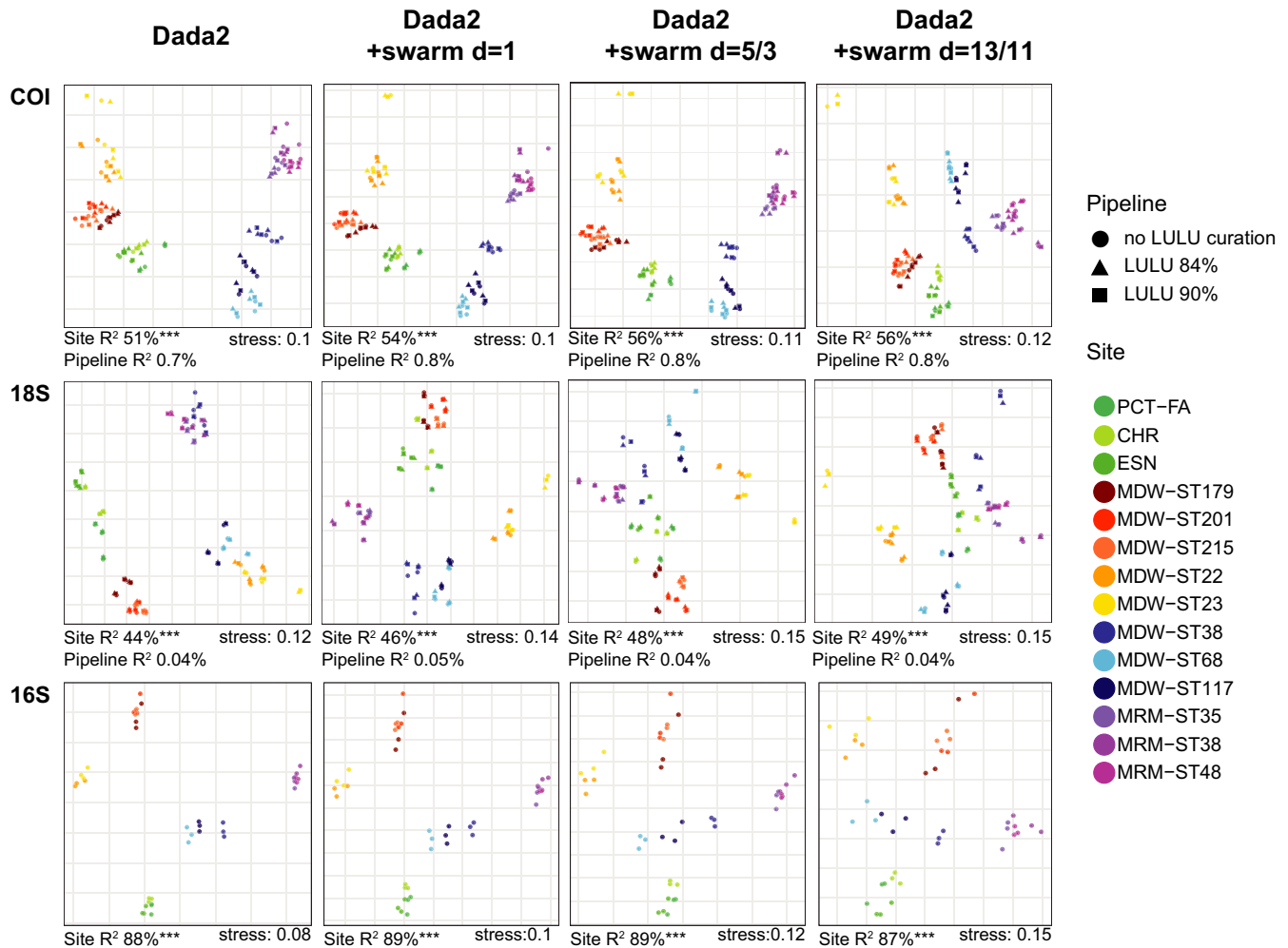


FIGURE 2 Metazoan (COI, 18S) and prokaryote (16S) beta-diversity patterns in ASV and OTU-centred data sets. Nonmetric multidimensional scaling (NMSD) ordinations showing community differentiation observed between sites with different clustering scenarios. ASVs were obtained with the DADA2 metabarcoding pipeline, and clustered with swarm at $d = 1, 5$, and 13 (COI) and $d = 1, 3, 11$ (18S, 16S). Metazoan ASVs and OTUs were curated with LULU at 84% and 90% minimum match. LULU curation was performed with minimum ratio = 100 for 18S and minimum ratio = 1 for COI. R^2 values and associated p -values obtained in PERMANOVAs are shown under the ordination plots. Significance codes: *** $p < .001$; ** $p < .01$; * $p < .05$. Site colour codes: Green: Mediterranean >1,000 m; Red: Mediterranean Gibraltar Strait 300–1,000 m; Yellow: Atlantic Gibraltar Strait 300–1,000 m; Blue: North Atlantic > 1,000 m; Purple: Arctic >1,000 m

connected lineages with a distinct role” (Freudenstein et al., 2017), constitutes the core of biodiversity inventories for biological and ecological studies. The species is a core concept in ecology and evolution that helps organizing agriculture, trade, and industry (e.g., species used for the production of biomaterial), as well as measuring the impact of human activity on Earth’s ecosystems (e.g., biomarker taxa and pathogenic or invasive species). While biotic diversity can be valued and assessed at various levels, including that of the individual organism and the genetic locus, many theoretical and applied developments in ecology are deeply rooted in the species concept, and species richness, while not perfect, remains an essential metric (Freudenstein et al., 2017).

Clustering ASVs into OTUs alleviated the numerical inflation in the mock samples, but some species still produced more than one OTU, even at d -values as high as $d = 11$ – 13 . While clustering improved numerical results in the mock communities, it led to poorer

taxonomic assignments, for example the vesicomyid bivalves only being identified up to class-level in clustered data sets with both loci. With 18S, clustering at d -values > 1 even led to the loss of the shrimp species *Chorocaris* sp., which was merged to the closely related *A. muricola* (Table 1). Similarly, a d -value at 11 led to significantly lower OTU numbers than any other tested d -value for both ribosomal loci (Table 2), explaining the much higher ASV to OTU ratios observed (4.1–4.4, Table S7). When studying natural habitats, very likely to harbour closely related co-occurring species, clustering at d -values higher than 1 is thus likely to lead to the loss of true species diversity, particularly in taxa known to be poorly resolved (e.g., cnidarians with COI, Hebert et al., 2003), and in general with markers having lower taxonomic resolution such as 18S.

The reproductive mode and pace of selection in microbial populations may lead to locally lower levels of intraspecific variation than those expected for metazoans. Prokaryotic alpha diversity was

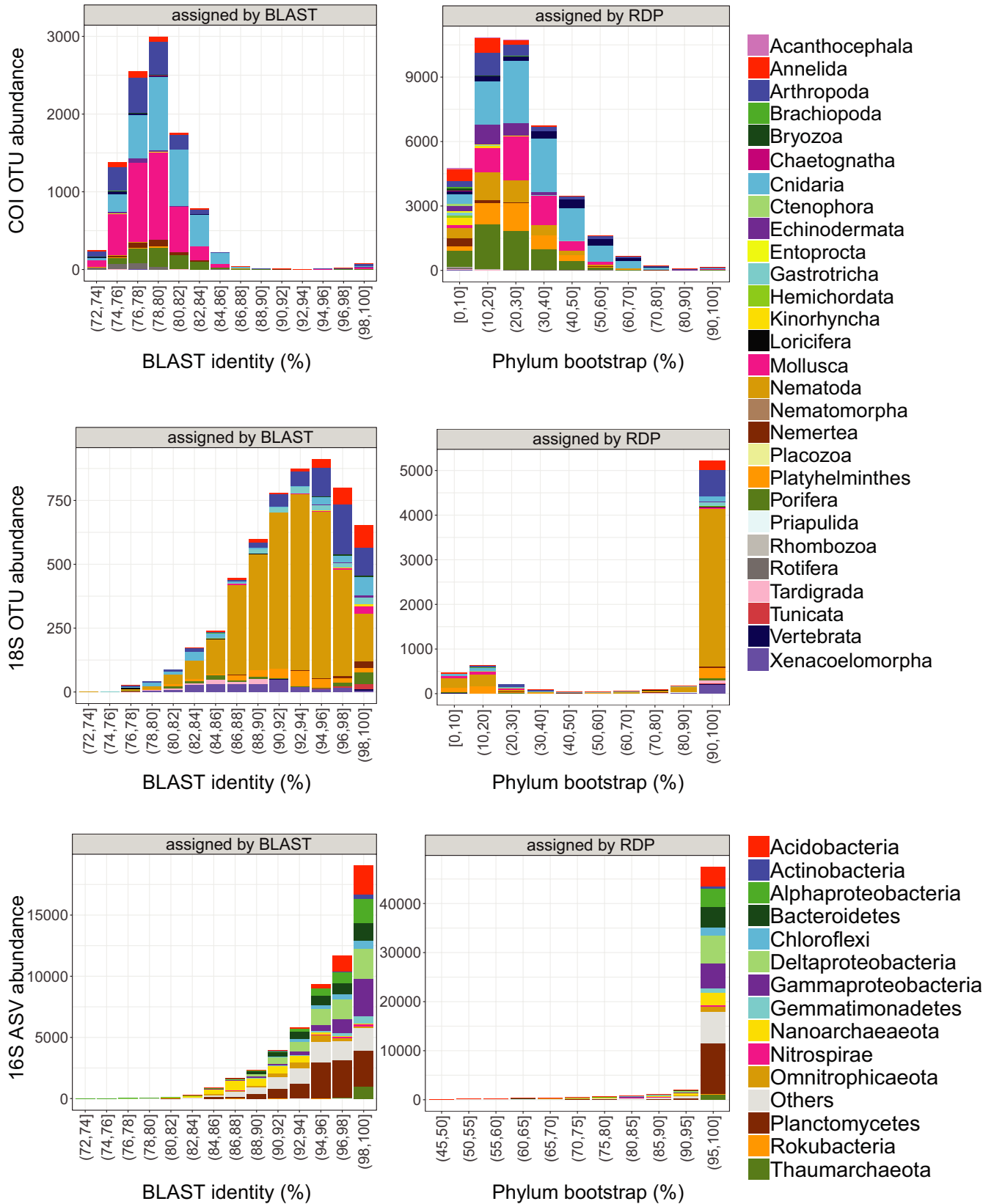


FIGURE 3 Taxonomic assignment quality of BLAST and RDP methods on metazoan (COI, 18S) and prokaryote (16S) metabarcoding data sets of 14 deep-sea sites. Metazoan data was clustered with swarm at $d = 1$ and curated with LULU at 90% (minimum ratio = 100) for 18S and 84% (minimum ratio = 1) for COI. Taxonomic assignments were performed on the Silva132 database for 18S and 16S, and on the MIDORI-UNIQUE database subsampled to marine taxa for COI

however also affected by the clustering of ASVs (Figure 1), supporting the estimation of a 2.5-fold greater number of 16S rRNA variants than the actual number of bacterial "species" (Acinas et al., 2004). The significant decrease in the number of OTUs after clustering at $d = 1$ (Table 2, Figure 1, decrease of ~30%) suggests the occurrence of very closely related 16S rRNA sequences, possibly belonging to the same ecotype/species. Such entities may still be important to define in studies aiming for example at identifying species associations (i.e., symbiotic relationships) across large distances and ecosystems, where drift or selection can lead to slightly different ASVs in space and time, with their function and association remaining stable.

Finally, apart from alpha diversity estimates, clustering also affected the resolution of ecological patterns in ribosomal loci when d -values were higher than 1 (Figure 2). This can be explained by the fact that clustering gives more weight to large distinct OTUs compared to many small (i.e., with low read numbers) ASVs. The deep Atlantic and Mediterranean sites, segregating at the ASV-level (possibly due to vicariance by distance), thus appeared more similar at high d -values, revealing the occurrence of distinct ASVs belonging to many shared OTUs and thus suggesting an ecological signal in finescale sequence variants. This is in accordance with other studies reporting differences in beta diversity patterns in ASV versus OTU data sets for ribosomal loci, when large divergence thresholds were used for clustering (Bokulich et al., 2013; Xiong & Zhan, 2018). This also reveals the interdependence of alpha and beta diversity components, so that clustering ASVs into OTUs and thereby reducing alpha diversity, leaves more space for beta diversity to be expressed, as observed in both population genetics (Beaumont & Nichols, 1996; Jost, 2008) and community analysis (Jost, 2007). Overall, these results confirm the advantage of combining error-correction tools with clustering and post-clustering curation tools, as this allows access to both interspecies and intraspecies information (Turon et al., 2020).

4.2 | Importance of parameter adjustment for LULU curation

LULU curation proved effective in limiting the number of multiple clusters produced by single individuals in the mock samples, confirming its efficiency to correct for intragenomic diversity (Table 1). Moreover, the fact that the number of unexpected clusters ("Others", Table 1) was not affected by LULU curation also shows that LULU specifically removes spurious OTUs and not true species diversity. However, careful adjustment of LULU parameters was needed, particularly for the minimum ratio, as at default level (1) it led to the loss of up to two mock species with 18S. This need for relaxed minimum ratio values can be explained by the nonideal design of the mock samples. Indeed, LULU should be applied on data sets containing as many samples as possible, which should have compositional similarities (i.e., overlapping species lists). If this is not the case, LULU will work as a pure clustering algorithm, at defined minimum match levels. Here, all species were co-occurring in the mock samples at consistent abundance ratios and some mock species were not occurring

(or rarely) in environmental samples. For those, random amplification biases leading to consistently low read numbers in both mock samples resulted in LULU merging them to closely related mock species. Increasing the minimum ratio, i.e., the expected minimal abundance ratio between a true OTU and an associated spurious sequence, allowed detecting all mock species with 18S. With minimum ratio at 100, one mock species (the gastropod *Paralepetopsis* sp.) was still lost when minimum match was at 84%, which could indicate that minimum match at 90% is more appropriate for 18S. However, as all mock species were retained at both minimum match levels with minimum ratio at 1,000, the loss of that species at 84% may also just reflect the nonideal mock design (*Paralepetopsis* sp. being very poorly amplified by 18S, it got merged to a bivalve OTU as their similarity was greater than 84%). Given the fact that 18S is evolving much slower than COI, this marker is taxonomically much less resolutive and phylum-level similarity is at ~86% (Stefanni et al., 2018). As error OTUs are produced within each individual, it is reasonable to think that their similarity to their parent OTUs will be greater than phylum-level similarity, justifying the use of 90% minimum match. This increased minimum match also has the added benefit to decrease calculation time on large data sets. For COI, although results in the mock samples showed the best performance at minimum ratio of 1 and little effect of the minimum match parameter (90% vs. 84%), both minimum match levels resulted in significantly different OTU numbers in the environmental samples (Table 2, Figure 1). This was not the case for 18S, where both 84% and 90% minimum match resulted in similar numbers of OTUs in the environmental samples (minimum ratio at 100). Thus, increasing the minimum ratio parameter is essential for not losing species in sample-poor data sets, and will be more correct than adjusting the minimum match.

The mock communities used in this study, apart from being taxonomically limited to just 10 species, did unfortunately not contain several haplotypes of the same species (intraspecific variation). This could explain the comparable results obtained with LULU curation of ASVs and LULU curation of OTUs in the mock samples, and lead to the hasty conclusion of a limited effect of clustering. Communities detected in environmental samples are much more complex, probably comprising many different haplotypes of the same species. However, LULU curation of ASVs cannot substitute clustering algorithms to account for natural haplotype diversity. Indeed, not all haplotypes co-occur and when they do so, they may vary in proportion and dominance relationships, making clustering the best tool to account for natural haplotypic diversity. This is in line with LULU developers (Frøslev et al., 2017), who recommend clustering ASVs for addressing the average intraspecific variation of the target group, and subsequent curation with LULU. In the environmental samples, LULU curation of the ASV data sets led to significantly more OTUs than swarm clustering at intermediate to high d -values and than LULU curation of swarm-clustered OTUs, with both metazoan loci (Table 2). This indicates that LULU curation merges less ASVs than the amount grouped through clustering, and highlights the different purposes of both tools, LULU effectively removing spurious OTUs, while clustering allows removing haplotype diversity.

4.3 | Taxonomic resolution and assignment quality

The COI locus allowed the detection of all 10 species present in the mock samples, compared to seven in the 18S data set (Table 1). It also provided much more accurate assignments, most of them correct at the genus (and species) level, confirming that COI uncovers more metazoan species and offers a better taxonomic resolution than 18S (Andújar et al., 2018; Clarke et al., 2017; Tang et al., 2012). Our results also support approaches combining nuclear and mitochondrial markers to achieve more comprehensive biodiversity inventories (Coward et al., 2015; Drummond et al., 2015; Zhan et al., 2014). Indeed, strong differences exist in amplification success among taxa (Bhadury et al., 2006; Carugati et al., 2015), exemplified by nematodes, which are well detected with 18S but not with COI (Bucklin et al., 2011). The 18S barcode marker performed better in the detection of nematodes, annelids, platyhelminths, and xenacoelomorphs while COI mostly detected cnidarians, molluscs, and poriferans (Figure 3, Figure S1), highlighting the complementarity of these two loci. This high complementarity of COI and 18S in terms of targeted taxa also supports the approach taken by Stefanni et al. (2018), indeed subsampling each gene data set for its “best targeted phyla” and subsequently combining both, seems to be a very efficient way to produce comprehensive and nonredundant biodiversity inventories.

Finally, similar taxonomic assignments were observed using BLAST or the RDP Bayesian Classifier in the mock samples for 18S and for COI when using the MIDORI-UNIQUE marine-only database (Figure S3), in line with the comparable performances of taxonomy prediction algorithms reported by Edgar (2018). Poor performance of RDP using the full COI MIDORI database is likely due to the size of the database, and to its low coverage of deep-sea species. Indeed, smaller databases with reduced sampling bias, taxonomically similar to the targeted communities, and with sequences of the same length as the DNA fragment of interest, are known to maximise accurate identification (Edgar 2018; Macheriotou et al., 2019; Ritari et al., 2015). The problem of underrepresentation of deep-sea taxa is especially apparent with the BLAST assignments, which generally displayed low identities to sequences in databases, especially for COI (Figure 3). Minimum similarities of 80% for COI and 86% for 18S as cutoff values for metazoans have been used to improve the taxonomic quality of metazoan metabarcoding data sets (Stefanni et al., 2018). However, phylogenies of marine invertebrates are characterised by high levels of species divergence (up to ~30%), even within genera (Zanol et al., 2010). Studies on deep-sea taxa have found that some invertebrate species had COI sequences diverging more than 20% from any other species present in molecular databases (Herrera et al., 2015; Shank et al., 1999). At present, it thus seems difficult to work at low taxonomic levels with deep-sea metazoan metabarcoding data when using large public databases (as revealed by the strong decreases in OTU numbers observed after genus-level quality filtering, Table S8). With the reduced marine-only COI database, RDP provided slightly more accurate assignments in the mock samples (Figure S3). However, filtering to accurate phylum-level bootstraps

(≥ 80) drastically reduced the number of OTUs in the overall data set (1% of OTUs left, Table S8). The development of custom-built RDP training sets, without overrepresentation of terrestrial species, is therefore needed for this Bayesian assignment method to be effective on deep-sea metazoan data sets. At present, if more accurate taxonomic assignments are sought while using universal primers, we advocate assigning taxonomy in two steps: first, using BLAST and a large database including all phyla amplifiable by the primer set, as BLAST performs better than RDP in terms of speed. The clusters belonging to the groups of interest can then be extracted and re-assigned using RDP and a smaller, taxon-specific database.

5 | CONCLUSIONS AND PERSPECTIVES

Using mock communities and environmental samples, we evaluate several recent algorithms and assess their capacity to improve the quality of molecular biodiversity inventories of metazoans and prokaryotes. Our results support the fact that ASV data should be produced and communicated for reusability and reproducibility following the recommendations of Callahan et al. (2017). This is especially useful in large projects spanning wide geographic zones and time scales, as different ASV data sets can easily be merged a posteriori, and clustered if necessary afterwards. However, our results confirm that both ASVs and OTUs describe relevant, yet different levels of biotic diversity. ASVs comprehensively describe genetic diversity (including intraspecies) while OTUs more accurately reflect interspecies diversity. Considering 16S polymorphism observed in prokaryotic species (Acinas et al., 2004) and the possible geographic segregation of their populations, using OTUs may also be suitable in prokaryotic data sets, for example in studies screening for species associations, as symbionts may be prone to differential fixation through enhanced drift (Shapiro et al., 2016).

This study emphasized that swarm clustering needs to be adapted to each genetic marker and taxonomic compartment, in order to identify an optimal balance between the correction for spurious clusters and the loss of species. Specifically, d -values > 1 appeared to be less appropriate with 18S for metazoans. Our results also demonstrated that LULU effectively curates metazoan biodiversity inventories obtained through metabarcoding. They underline the need to adapt parameters for LULU curation, in particular the minimum ratio level in the case of sample-poor data sets, where co-occurrence and abundance patterns may be distorted.

Finally, this study also showed that accurate taxonomic assignments of deep-sea species can be obtained with the RDP Bayesian Classifier, but only with reduced databases containing ecosystem-specific sequences.

ACKNOWLEDGEMENTS

This work is part of the “Pourquoi Pas les Abysses?” project, funded by Ifremer and the eDNAbyss (AP2016-228) project, funded by France Génomique (ANR-10-INBS-09) and Genoscope-CEA. We wish to thank Caroline Dussart for her contribution to the early

developments of bioinformatic scripts, Patrick Durand for bioinformatic support, and Cathy Liautard-Haag for her help with lab management. We also wish to express our gratitude to the participants and mission chiefs of the EssNaut16 (Marie-Anne Cambon Bonavita), MarMine (Eva Ramirez Llodra, project 247626/O30, NRC-BI), PEACETIME (Cécile Guieu and Christian Tamburini), CanHROV (Marie Claire Fabri) and MEDWAVES (Covadonga Orejas) cruises. The MEDWAVES cruise was organised in the framework of the ATLAS Project, supported by the European Union 2020 Research and Innovation Programme under grant agreement number: 678760. We thank Stefaniya Kamenova, Tiago Pereira, and four anonymous reviewers for their comments on previous versions of this manuscript. An earlier version of this manuscript has been peer-reviewed and recommended by Peer Community In Ecology (<https://dx.doi.org/10.24072/pci.ecology.100043>).

AUTHOR CONTRIBUTIONS

Miriam I. Brandt and Sophie Arnaud-Haond designed the study, Miriam I. Brandt and Julie Poulain carried out the laboratory work; Miriam I. Brandt and Blandine Trouche performed the bioinformatic and statistical analyses. Laure Quintric assisted in the bioinformatic development and participated in the study design. Miriam I. Brandt, Babett Günther, and Sophie Arnaud-Haond wrote the manuscript. All authors contributed to the final manuscript.

DATA AVAILABILITY STATEMENT

The raw data of this work can be accessed in the European Nucleotide Archive (ENA) database (project: PRJEB33873), please refer to the Data S1 for ENA file names. The data set, including raw sequences, reference databases, and ASV/OTU tables, is accessible via <https://doi.org/10.12770/0b5d250b-8418-4dda-b39c-960c4481df93>. Bioinformatic scripts, config files, and R scripts are available on Gitlab (<https://gitlab.ifremer.fr/abyss-project/>).

ORCID

Miriam I. Brandt  <https://orcid.org/0000-0002-5734-0087>

Blandine Trouche  <https://orcid.org/0000-0003-4307-4782>

Babett Günther  <https://orcid.org/0000-0003-1225-4262>

Sophie Arnaud-Haond  <https://orcid.org/0000-0001-5814-8452>

REFERENCES

- Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rnr operons. *Journal of Bacteriology*, 186(9), 2629–2635. <https://doi.org/10.1128/JB.186.9.2629-2635.2004>.
- Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2017). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, 9(1), 134–147. <https://doi.org/10.1111/2041-210X.12849>.
- Andújar, C., Arribas, P., Gray, C., Bruce, C., Woodward, G., Yu, D. W., & Vogler, A. P. (2018). Metabarcoding of freshwater invertebrates to detect the effects of a pesticide spill. *Molecular Ecology*, 27(1), 146–166. <https://doi.org/10.1111/mec.14410>.
- Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the Metazoa. *Molecular Ecology*, 27(20), 3968–3975. <https://doi.org/10.1111/mec.14844>.
- Atienza, S., Guardiola, M., Præbel, K., Antich, A., Turon, X., & Wangensteen, O. S. (2020). DNA metabarcoding of Deep-Sea sediment communities using COI: Community Assessment, Spatio-Temporal Patterns and Comparison with 18S rDNA. *Diversity*, 12(4), 123. <https://doi.org/10.3390/D12040123>.
- Baselga, A., & Orme, C. D. L. (2012). Betapart: An R package for the study of beta diversity. *Methods in Ecology and Evolution*, 3(5), 808–812. <https://doi.org/10.1111/j.2041-210X.2012.00224.x>.
- Bazin, E., Glémin, S., & Galtier, N. (2006). Population size does not influence mitochondrial genetic diversity in animals. *Science*, 312(5773), 570–572. <https://doi.org/10.1126/science.1122033>.
- Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263(1377), 1619–1626. <https://doi.org/10.1098/rspb.1996.0237>.
- Bensasson, D., Zhang, D. X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial Pseudogenes: Evolution's Misplaced Witnesses. *Trends in Ecology & Evolution*, 16(6), 314–321. [https://doi.org/10.1016/S0169-5347\(01\)02151-6](https://doi.org/10.1016/S0169-5347(01)02151-6).
- Bhadury, P., Austen, M. C., Bilton, D. T., Lamshead, P. J. D., Rogers, A. D., & Smerdon, G. R. (2006). Molecular detection of marine nematodes from environmental samples: Overcoming eukaryotic interference. *Aquatic Microbial Ecology*, 44(1), 97–103. <https://doi.org/10.3354/Ame044097>.
- Bik, H. M., Sung, W., De Ley, P., Baldwin, J. G., Sharma, J., Rocha-Olivares, A., & Thomas, W. K. (2012). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in Deep-Sea and Shallow Water Sediments. *Molecular Ecology*, 21(5), 1048–1059. <https://doi.org/10.1111/j.1365-294X.2011.05297.x>.
- Bista, I., Carvalho, G., Walsh, K., Christmas, M., Hajibabaei, M., Kille, P., Lallias, D., & Creer, S. (2015). Monitoring lake ecosystem health using metabarcoding of environmental DNA: Temporal persistence and ecological relevance. *Genome*, 58(5), 197.
- Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nature Methods*, 10(1), 57–59. <https://doi.org/10.1038/nmeth.2276>.
- Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J. B., Kiszka, J. J., Kulbicki, M., Manel, S., Robbins, W. D., Vigliola, L., & Mouillot, D. (2018). Environmental DNA illuminates the dark diversity of sharks. *Science Advances*, 4(5), eaap9661. <https://doi.org/10.1126/sciadv.aap9661>.
- Brannock, P. M., & Halanych, K. M. (2015). Meiofaunal community analysis by high-throughput sequencing: comparison of extraction, quality filtering, and clustering methods. *Marine Genomics*, 23, 67–75. <https://doi.org/10.1016/j.margen.2015.05.007>.
- Brown, E. A., Chain, F. J. J., Crease, T. J., Macisaac, H. J., & Cristescu, M. E. (2015). Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe Zooplankton Communities? *Ecology and Evolution*, 5(11), 2234–2251. <https://doi.org/10.1002/ece3.1485>.
- Bucklin, A., Steinke, D., & Blanco-Bercial, L. (2011). DNA barcoding of marine Metazoa. *Annual Review of Marine Science*, 3(1), 471–508. <https://doi.org/10.1146/annurev-marine-120308-080950>.
- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, 11(12), 2639–2643. <https://doi.org/10.1038/ismej.2017.119>.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-Resolution Sample Inference from Illumina Amplicon Data. *Nature Methods*, 13(7), 581–583. <https://doi.org/10.1038/nmeth.3869>.

- Carranza, S., Giribet, G., Ribera, C., Baguñà, J., & Riutort, M. (1996). Evidence that two types of 18S rDNA coexist in the genome of *Dugesia* (Schmidtea) Mediterranea (Platyhelminthes, Turbellaria, Tricladida). *Molecular Biology and Evolution*, 13(6), 824–832. <https://doi.org/10.1093/oxfordjournals.molbev.a025643>
- Carugati, L., Corinaldesi, C., Dell'Anno, A., & Danovaro, R. (2015). Metagenetic tools for the census of marine meiofaunal biodiversity: An overview. *Marine Genomics*, 24, 11–20. <https://doi.org/10.1016/j.margen.2015.04.010>.
- Clare, E. L., Chain, F. J. J., Littlefair, J. E., & Cristescu, M. E. (2016). The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. *Genome*, 59(11), 981–990. <https://doi.org/10.1139/gen-2015-0184>
- Clarke, L. J., Beard, J. M., Swadling, K. M., & Deagle, B. E. (2017). Effect of Marker choice and thermal cycling protocol on Zooplankton DNA Metabarcoding studies. *Ecology and Evolution*, 7(3), 873–883. <https://doi.org/10.1002/ece3.2667>.
- Cohan, F. M. (2001). Bacterial species and speciation. *Systematic Biology*, 50(4), 513–524. <https://doi.org/10.1080/10635150118398>.
- Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, 21(8), 1834–1847. <https://doi.org/10.1111/j.1365-294X.2012.05550.x>
- Cordier, T., Frontalini, F., Cermakova, K., Apothéloz-Perret-Gentil, L., Treglia, M., Scantamburlo, E., Bonamin, V., & Pawlowski, J. W. (2019). Multi-marker EDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Marine Environmental Research*, 146, 24–34. <https://doi.org/10.1016/j.marenvres.2018.12.009>.
- Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., & Arnaud-Haond, S. (2015). Metabarcoding is powerful yet still blind: A comparative analysis of morphological and molecular surveys of Seagrass communities. *PLoS One*, 10(2), e0117562. <https://doi.org/10.1371/journal.pone.0117562>.
- Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The Ecologist's Field Guide to Sequence-Based Identification of Biodiversity. *Methods in Ecology and Evolution*, 7(9), 1008–1018. <https://doi.org/10.1111/2041-210X.12574>.
- Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple Statistical Identification and Removal of Contaminant Sequences in Marker-Gene and Metagenomics Data. *Microbiome*, 6(1), 226. <https://doi.org/10.1186/s40168-018-0605-2>.
- de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*, 102(Suppl 1), 6600–6607. <https://doi.org/10.1073/pnas.0502030102>.
- Deiner, K., Fronhofer, E. A., Mächler, E., Walser, J. C., & Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyor belts of biodiversity information. *Nature Communications*, 7(1), 12544. <https://doi.org/10.1038/ncomms12544>.
- Deiner, K., Walser, J.-C.-C., Mächler, E., Altermatt, F., Mächler, E., & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, 183, 53–63. <https://doi.org/10.1016/j.biocon.2014.11.018>.
- Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., & Miaud, C. (2011). Persistence of Environmental DNA in freshwater ecosystems. *PLoS One*, 6(8), e23398. <https://doi.org/10.1371/journal.pone.0023398>.
- Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., Holdaway, R. J., Lear, G., Makiola, A., Morales, S. E., Powell, J. R., & Weaver, L. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources*, 18(5), 940–952. <https://doi.org/10.1111/1755-0998.12907>
- Djurhuus, A., Port, J., Closek, C. J., Yamahara, K. M., Romero-Maraccini, O. C., Walz, K. R., Goldsmith, D. B., Michisaki, R., Breitbart, M., Boehm, A. B., & Chavez, F. P. (2017). Evaluation of filtration and DNA extraction methods for environmental DNA biodiversity assessments across multiple trophic levels. *Frontiers in Marine Science*, 4(October), 314. <https://doi.org/10.3389/fmars.2017.00314>.
- Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C. M., Heled, J., Ross, H. A., Tooman, L., Grosser, S., Park, D., Demetras, N. J., Stevens, M. I., Russell, J. C., Anderson, S. H., Carter, A., & Nelson, N. (2015). Evaluating a Multigene Environmental DNA Approach for Biodiversity Assessment. *Gigascience*, 4(1), 46. <https://doi.org/10.1186/s13742-015-0086-1>.
- Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J. W., & Cordier, T. (2019). SLIM: A flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinformatics*, 20(1), 88. <https://doi.org/10.1186/s12859-019-2663-2>.
- Edgar, R. C. (2017). SINAPS: Prediction of Microbial Traits from Marker Gene Sequences. *BioRxiv*. <https://doi.org/10.1101/124156>.
- Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, 6(4), e4652. <https://doi.org/10.7717/peerj.4652>.
- Eren, A. M., Vineis, J. H., Morrison, H. G., & Sogin, M. L. (2013). A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One*, 8(6), e66643. <https://doi.org/10.1371/journal.pone.0066643>.
- Escudé, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S., & Pascal, G. (2018). FROGS: Find, rapidly, OTUs with galaxy solution. *Bioinformatics*, 34(8), 1287–1294. <https://doi.org/10.1093/bioinformatics/btx791>.
- Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y. Y., Jerde, C. L., Mahon, A. R., Pfrender, M. E., Lamberti, G. A., & Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, 16(1), 29–41. <https://doi.org/10.1111/1755-0998.12433>.
- Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguët-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from EDNA metabarcoding data. *Molecular Ecology Resources*, 15(3), 543–556. <https://doi.org/10.1111/1755-0998.12338>.
- Fonseca, V. G. (2018). Pitfalls in relative abundance estimation using edna metabarcoding. *Molecular Ecology Resources*, 18(5), 923–926. <https://doi.org/10.1111/1755-0998.12902>.
- Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power, D. M., Neill, S. P., & Packer, M. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, 1(1), 98. <https://doi.org/10.1038/ncomms1095>.
- Freudenstein, J. V., Broe, M. B., Folk, R. A., & Sinn, B. T. (2017). Biodiversity and the species concept - lineages are not enough. *Systematic Biology*, 66(4), 644–656. <https://doi.org/10.1093/sysbio/syw098>.
- Frøslev, T. G., Kjølner, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, 8(1), 1188. <https://doi.org/10.1038/s41467-017-01312-x>.
- Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., de Peer, Y. V., Vandamme, P., Thompson, F. L., & Swings, J. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, 3(9), 733–739. <https://doi.org/10.1038/nrmicr01236>.

- Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., & Spear, S. F. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, 7(11), 1299–1307. <https://doi.org/10.1111/2041-210X.12595>.
- Hashimoto, J. G., Stevenson, B. S., & Schmidt, T. M. (2003). Rates and consequences of recombination between rRNA Operons. *Journal of Bacteriology*, 185(3), 966–972. <https://doi.org/10.1128/JB.185.3.966-972.2003>.
- Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 270(suppl_1), S96–S99. <https://doi.org/10.1098/rsbl.2003.0025>.
- Herrera, S., Watanabe, H., & Shank, T. M. (2015). Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology*, 24(3), 673–689. <https://doi.org/10.1111/mec.13054>.
- Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P. M., Woodcock, P., Edwards, F. A., Larsen, T. H., Hsu, W. W., Benedick, S., Hamer, K. C., Wilcove, D. S., Bruce, C., Wang, X., Levi, T., Lott, M., ... Yu, D. W. (2013). Reliable, Verifiable and Efficient Monitoring of Biodiversity via Metabarcoding. *Ecology Letters*, 16(10), 1245–1257. <https://doi.org/10.1111/ele.12162>.
- Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10), 2427–2439. <https://doi.org/10.1890/06-1736.1>.
- Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology*, 17(18), 4015–4026. <https://doi.org/10.1111/j.1365-294X.2008.03887.x>.
- Klappenbach, J. A., Saxman, P. R., Cole, J. R., & Schmidt, T. M. (2001). Rndb: The ribosomal RNA operon copy number database. *Nucleic Acids Research*, 29(1), 181–184. <https://doi.org/10.1093/nar/29.1.181>.
- Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lear, G., & Pochon, X. (2018). A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Marine Pollution Bulletin*, 127, 97–107. <https://doi.org/10.1016/j.marpolbul.2017.11.042>.
- Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10, 34. <https://doi.org/10.1186/1742-9994-10-34>.
- Macheriotou, L., Guilini, K., Bezerra, T. N., Tytgat, B., Nguyen, D. T., Phuong Nguyen, T. X., Noppe, F., Armenteros, M., Boufahja, F., Rigaux, A., Vanreusel, A., & Derycke, S. (2019). Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecology and Evolution*, 9(1), 1–16. <https://doi.org/10.1002/ece3.4814>.
- Machida, R. J., & Knowlton, N. (2012). PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences. *PLoS One*, 7(9), e46180. <https://doi.org/10.1371/journal.pone.0046180>.
- Machida, R. J., Kweskin, M., & Knowlton, N. (2012). PCR primers for metazoan mitochondrial 12S ribosomal DNA sequences. *PLoS One*, 7(4), e35887. <https://doi.org/10.1371/journal.pone.0035887>.
- Machida, R. J., Leray, M., Ho, S. L., & Knowlton, N. (2017). Data descriptor: Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, 4, 170027. <https://doi.org/10.1038/sdata.2017.27>.
- Mahé, F., Rognes, T., de Quince, C., Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3(December), e1420. <https://doi.org/10.7717/peerj.1420>.
- Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W. H. C. F., Logares, R., ... de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, 17(10), 4035–4049. <https://doi.org/10.1111/1462-2920.12955>.
- Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a Zoologist*. Columbia University Press. <http://www.hup.harvard.edu/catalog.php?isbn=9780674862500>
- McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>.
- Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biology*, 12(11), R112. <https://doi.org/10.1186/gb-2011-12-11-r112>
- Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, 6, e5364. <https://doi.org/10.7717/peerj.5364>.
- Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., Green, R. E., & Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, 18(5), 927–939. <https://doi.org/10.1111/1755-0998.12895>.
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlenn, D., Minchin, P. R., & O'Hara, R. B. (2018). *Vegan: Community Ecology Package*. <https://cran.r-project.org/package=vegan>
- Pansu, J., Giguet-Covex, C., Ficetola, F., Gielly, L., Boyer, F., Coissac, E., Domaizon, I., Zinger, L., Poulenard, J., & Arnaud, F. (2015). Environmental DNA metabarcoding to investigate historic changes in biodiversity. *Genome*, 58(5), 264.
- Parada, A. E., Needham, D. M., & 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), 1403–1414. <https://doi.org/10.1111/1462-2920.13023>.
- Pawlowski, J. W., Christen, R., Lecroq, B., Bachar, D., Shahbazkia, H. R., Amaral-Zettler, L., & Guillou, L. (2011). Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing. *PLoS One*, 6(4), e18169. <https://doi.org/10.1371/journal.pone.0018169>.
- Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., Jin, Z., Lee, P., Yang, L., Poles, M., Brown, S. M., Sotero, S., DeSantis, T., Brodie, E., Nelson, K., & Pei, Z. (2010). Diversity of 16S rRNA genes within individual prokaryotic genomes. *Applied and Environmental Microbiology*, 76(12), 3886–3897. <https://doi.org/10.1128/AEM.02953-09>.
- Phillips, J. D., Gillis, D. J., & Hanner, R. H. (2019). Incomplete estimates of genetic diversity within species: Implications for DNA barcoding. *Ecology and Evolution*, 9(5), 2996–3010. <https://doi.org/10.1002/ece3.4757>.
- Plouviez, S., Shank, T. M., Faure, B., Daguin-Thiebaut, C., Viard, F., Lallier, F. H., & Jollivet, D. (2009). Comparative phylogeography among hydrothermal vent species along the east Pacific rise reveals vicariant processes and population expansion in the South. *Molecular Ecology*, 18(18), 3903–3917. <https://doi.org/10.1111/j.1365-294X.2009.04325.x>.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1), D590–596. <https://doi.org/10.1093/nar/gks1219>.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>

- Ritari, J., Salojärvi, J., Lahti, L., & de Vos, W. M. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics*, *16*(1), 1056. <https://doi.org/10.1186/s12864-015-2265-y>.
- Salazar, G., Cornejo-Castillo, F. M., Benitez-Barrrios, V., Fraile-Nuez, E., Alvarez-Salgado, X. A., Duarte, C. M., Gasol, J. M., & Acinas, S. G. (2016). Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME Journal*, *10*(3), 596–608. <https://doi.org/10.1038/ismej.2015.137>.
- Sawaya, N. A., Djurhuus, A., Closek, C. J., Hepner, M., Olesin, E., Visser, L., Kelble, C., Hubbard, K., & Breitbart, M. (2019). Assessing eukaryotic biodiversity in the Florida keys national marine sanctuary through environmental DNA metabarcoding. *Ecology and Evolution*, *9*(3), 1029–1040. <https://doi.org/10.1002/ece3.4742>.
- Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated - Reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. <https://doi.org/10.1111/1755-0998.12402>.
- Shank, T. M., Black, M. B., Halanych, K. M., Lutz, R. A., & Vrijenhoek, R. C. (1999). Miocene radiation of deep-sea hydrothermal vent shrimp (Caridea: Bresiliidae): Evidence from mitochondrial cytochrome Oxidase Subunit I. *Molecular Phylogenetics and Evolution*, *13*(2), 244–254. <https://doi.org/10.1006/mpev.1999.0642>.
- Shapiro, B. J., Leducq, J. B., & Mallet, J. (2016). What Is Speciation? *PLoS Genetics*, *12*(3), e1005860. <https://doi.org/10.1371/journal.pgen.1005860>.
- Sinniger, F., Pawlowski, J. W., Harii, S., Gooday, A. J., Yamamoto, H., Chevaldonné, P., Cedhagen, T., Carvalho, G., & Creer, S. (2016). Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic knowledge of deep-sea benthos. *Frontiers in Marine Science*, *3*(June), 92. <https://doi.org/10.3389/fmars.2016.00092>.
- Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., De La Rasilla, M., Lalueza-Fox, C., Rosas, A., Soressi, M., Knul, M. V., Miller, R., Stewart, J. R., Dereviānko, A. P., Jacobs, Z., Li, B., Roberts, R. G., Shunkov, M. V., de Lumley, H., Perrenoud, C., Gusic, I., ... Meyer, M. (2017). Neandertal and denisovan DNA from pleistocene sediments. *Science*, *356*(6338), 605–608. <https://doi.org/10.1126/science.aam9695>.
- Sokal, R. R., & Crovello, T. J. (1970). The Biological species concept: A critical evaluation. *The American Naturalist*, *104*(936), 127–153.
- Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(36), 13486–13491. <https://doi.org/10.1073/pnas.0803076105>.
- Stat, M., Huggett, M. J., Bernasconi, R., Dibattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with EDNA: Metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, *7*(1), 12240. <https://doi.org/10.1038/s41598-017-12501-5>.
- Stefanni, S., Stanković, D., Borme, D., de Olazabal, A., Juretić, T., Pallavicini, A., & Tirelli, V. (2018). Multi-Marker Metabarcoding Approach to Study Mesozooplankton at Basin Scale. *Scientific Reports*, *8*(1), 12085. <https://doi.org/10.1038/s41598-018-30157-7>.
- Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–1793. <https://doi.org/10.1111/j.1365-294X.2012.05542.x>.
- Tang, C. Q., Leasi, F., Obertegger, U., Kieneker, A., Barraclough, T. G., & Fontaneto, D. (2012). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the Meiofauna. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(40), 16208–16212. <https://doi.org/10.1073/pnas.1209160109>.
- Teixeira, S., Olu, K., Decker, C., Cunha, R. L., Fuchs, S., Hourdez, S., Serrão, E. A., & Arnaud-Haond, S. (2013). High connectivity across the fragmented chemosynthetic ecosystems of the deep atlantic equatorial belt: Efficient dispersal mechanisms or questionable endemism? *Molecular Ecology*, *22*(18), 4663–4680. <https://doi.org/10.1111/mec.12419>.
- Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2020). From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecological Applications*, *30*(2), e02036. <https://doi.org/10.1002/eap.2036>.
- Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution*, *24*(2), 110–117. <https://doi.org/10.1016/j.tree.2008.09.011>.
- Valentini, A., Taberlet, P., Miaud, C., Civade, R. R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J. M., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942. <https://doi.org/10.1111/mec.13428>.
- Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J. M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., ... Karsenti, E. (2015). Eukaryotic plankton diversity in the Sunlit Ocean. *Science*, *348*(6237), <https://doi.org/10.1126/science.1261605>.
- Wangensteen, O. S., & Turon, X. (2016). Metabarcoding Techniques for Assessing Biodiversity of Marine Animal Forests. In S. Rossi, L. Bramanti, A. Gori, & C. Orejas Saco del Valle (Eds.), *Marine animal forests* (pp. 1–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-17001-5_53-1.
- Wood, S. A., Pochon, X., Laroche, O., von Ammon, U., Adamson, J., & Zaiko, A. (2019). A comparison of droplet digital polymerase chain reaction (PCR), Quantitative PCR and metabarcoding for species-specific detection in environmental DNA. *Molecular Ecology Resources*, *19*(6), 1407–1419. <https://doi.org/10.1111/1755-0998.13055>.
- Xiong, W., & Zhan, A. (2018). Testing clustering strategies for metabarcoding-based investigation of community-environment interactions. *Molecular Ecology Resources*, *18*(6), 1326–1338. <https://doi.org/10.1111/1755-0998.12922>.
- Yoccoz, N. G., Brathen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., von Stedingk, H., Brysting, A. K., Coissac, E., Pompanon, F., Sønstebo, J. H., Miquel, C., Valentini, A., De Bello, F., Chave, J., Thuiller, W., Wincker, P., Cruaud, C., Gavory, F., ... Taberlet, P. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, *21*(15), 3647–3655. <https://doi.org/10.1111/j.1365-294X.2012.05545.x>.
- Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623. <https://doi.org/10.1111/j.2041-210X.2012.00198.x>.
- Zanol, J., Halanych, K. M., Struck, T. H., & Fauchald, K. (2010). Phylogeny of the Bristle Worm Family Eunicidae (Eunicida, Annelida) and the phylogenetic utility of noncongruent 16S, COI and 18S in combined analyses. *Molecular Phylogenetics and Evolution*, *55*(2), 660–676. <https://doi.org/10.1016/j.ympev.2009.12.024>.
- Zhan, A., Bailey, S. A., Heath, D. D., & Macisaac, H. J. (2014). Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources*, *14*(5), 1049–1059. <https://doi.org/10.1111/1755-0998.12254>.

Zinger, L., Chave, J., Coissac, E., Iribar, A., Louisanna, E., Manzi, S., Schilling, V., Schimann, H., Sommeria-Klein, G., & Taberlet, P. (2016). Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biology and Biochemistry*, *96*, 16–19. <https://doi.org/10.1016/j.soilb.2016.01.008>.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Brandt MI, Trouche B, Quintric L, et al. Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Mol Ecol Resour.* 2021;00: 1–19. <https://doi.org/10.1111/1755-0998.13398>

4. Comparison of two 16S rRNA amplicon primers and metagenomic data for deep-sea benthic archaeal diversity studies

Blandine Trouche¹, Ferial Bouderkha¹, Clemens Schauburger², Jean-Christophe Auguet³, Caroline Belser⁴, Miriam Brandt³, Julie Poulain⁴, Bo Thamdrup², Patrick Wincker⁴, Ronnie N. Glud², Sophie Arnaud-Haond³ and Loïs Maignien^{1,5}

¹ Univ Brest, CNRS, IFREMER, Microbiology of Extreme Environments Laboratory (LM2E), F-29280 Plouzané, France

² Hadal & Nordcee, Department of Biology, University of Southern Denmark, Odense, Denmark

³ MARBEC, Univ Montpellier, Ifremer, IRD, CNRS, Sète, France

⁴ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Évry, Université Paris-Saclay, 91057 Evry, France

⁵ Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Woods Hole, MA, United States

Prepared for submission in PeerJ

Résumé de l'article en français

De récentes études ont souligné la grande diversité et le rôle écologique essentiel des Archées dans les sédiments marins profonds. A l'interface entre les écosystèmes pélagiques et les communautés de subsurface, les Archées des couches benthiques sont particulièrement intéressantes, mais encore peu explorées dans les zones abyssales et hadales. Dans cette étude nous avons utilisé 46 échantillons de sédiments de surface (0-30 cm) issus de deux fosses hadales pour comparer quatre méthodes à haut débit de caractérisation de la diversité archées. Deux de ces approches sont basées sur le séquençage d'amplicons de la région V4V5 du gène ARNr 16S, et utilisent dans un cas un couple d'amorces universelles (ciblant les Bactéries et les Archées) et dans l'autre un mélange d'amorces spécifiques des Archées. Les deux autres approches sont basées sur des données métagénomiques, évitant ainsi les biais de la PCR ciblée : d'une part l'extraction de fragments de la petite sous-unité de l'ARNr (appelés miTAGs) à partir de données non assemblées, et d'autre part la détection après assemblage de gènes centraux à copie unique (SCG). Les résultats montrent que les amorces universelles et les miTAGs produisent des profils de diversité archéale similaires. Ils soulignent également de façon frappante le manque de détection des Nanoarchaeota par les amorces spécifiques des Archées. Les deux jeux de données de métabarcoding montrent des signes de biais de PCR envers certains groupes plus rares tels que les Halobacterota et les Thermoprotei. Les profils de SCG ne produisent pas d'estimations fiables de la diversité des Archées benthiques, ce qui est potentiellement dû aux complexités d'assemblage des gènes de protéines ribosomales, gènes centraux considérés ici, pour des populations à l'abondance faible. Enfin, la diversité des estimations de la proportion d'Archées présentes dans les échantillons considérés a mis en valeur l'impossibilité de tirer des conclusions à partir des résultats obtenus pour chaque méthode. Cependant, la similarité des variations entre échantillons d'un même jeu de données pour les amplicons universels et les miTAGs semble suggérer une certaine fiabilité et reproductibilité

CHAPTER 1

de cette information au sein d'un même jeu de données. Nous concluons de cette étude que les amplicons générés à l'aide des amorces universelles et les miTAGs sont de bons *proxies* pour l'étude de la diversité globale des Archées dans les sédiments de surface des grands fonds, une observation intéressante à répéter pour d'autres environnements et lignées d'Archées afin de tester la validité plus générale de ces résultats.

Abstract

Recent studies have highlighted the wide diversity and important ecological role of Archaea in deep sea sediments. At the interface between pelagic ecosystems and seafloor communities, Archaea of the benthic layers are of particular interest, though as yet underexplored at abyssal and hadal depths. Here we used 46 surface sediment samples (0 - 30 cm) from two hadal trenches to compare four high throughput characterization methods. Two of these approaches relied on 16S rRNA V4V5 amplicon sequencing, comparing universal (targeting both Bacteria and Archaea) *versus* archaea-specific primers. The other two approaches were based on metagenomic data, and thus not subject to targeted-PCR bias: extraction of small subunit rRNA fragments (miTAGs) from unassembled data on the one hand, and single copy core genes (SCG) detection after assembly on the other hand. We found that universal primers and miTAGs provided similar profiles of archaeal diversity. More importantly, those results highlighted the lack of detection of Nanoarchaeota by the archaea-specific primers. Both metabarcoding datasets showed signs of PCR bias against specific groups among the rarer ones such as Halobacterota and Thermoprotei. SCG profiles did not give reliable estimates of the benthic archaeal diversity, probably due to complexities in the assembly of ribosomal protein genes, that were the core genes considered here, from low abundance populations. Finally, comparison of the estimated relative abundance of Archaea in these samples highlighted the impossibility to draw absolute conclusions from any of the four sets of results obtained. However, the similarity of the trends observed for the proportion of Archaea in the universal and miTAG datasets seemed to indicate a reliability and reproducibility of the information between samples of the same dataset. Overall, we found that universal primers and miTAGs were good *proxies* to investigate the general diversity of Archaea in deep sea surface sediment, an interesting observation to be repeated with other environments and lineages to test for the generic nature of this pattern.

Introduction

In the last decade, research efforts have widely expanded our knowledge of Archaea: first characterized as mostly extremophilic organisms shared between two phyla, Crenarchaeota and Euryarchaeota, they were later shown to be ubiquitous, fulfilling important ecosystemic services, and with a much wider phylogenetic diversity (reviewed in Spang et al., 2017; Baker et al., 2020). Archaea have also played a central role in evolutionary studies since the 1970s, with Carl Woese's work leading to the definition of the three domains of life (Woese and Fox, 1977; Woese et al., 1990). More recently, assembly of Asgard Archaea genomic sequences has led to suggest changes to the topology of the tree of life, and raised new hypotheses as to the last common ancestor between the different domains (Hug et al., 2016a; Zaremba-Niedzwiedzka et al., 2017; Castelle and Banfield, 2018; Spang, 2019).

Archaeal populations have been reported in pelagic ecosystems for decades (DeLong, 1992; Fuhrman et al., 1992), and deep-sea environments in particular have been instrumental in the continued description of the diversity of archaeal lineages (Fuhrman and Davis, 1997; Takai and Horikoshi, 1999; Vetriani et al., 1999; Huber et al., 2002b; Wang et al., 2005; Reysenbach et al., 2006; Jørgensen et al., 2013). Deep-sea benthic boundary layer archaeal populations are of particular interest, because they are located at the interface between the pelagic realm and seafloor communities, where Archaea play a crucial role in biogeochemical cycles (Biddle et al., 2006; Sørensen and Teske, 2006; Vuillemin et al., 2019). Indeed, according to recent studies (Petro et al., 2017; Starnawski et al., 2017; Kirkpatrick et al., 2019; Petro et al., 2019), surface sediment populations are the first step in the assembly of subsurface communities, and are important contributors to early diagenesis (Durbin and Teske, 2011; Teske et al., 2011; Molari et al., 2013). Yet, the archaeal diversity found in abyssal and hadal benthic sediments (> 4000 meters below sea level (mbsl)) is still sparsely described.

Cultivation efforts are bringing necessary insights into the populations harbored by these environments (Imachi et al., 2020). However, deep-sea archaeal isolates are more often recovered from hot environments (Huber et al., 2002a; Zeng et al., 2009; François et al., 2021),

CHAPTER 1

and overall, due to the extreme set of conditions these environments exhibit, cultivation remains challenging (Sun et al., 2019; Hu et al., 2021). In this context, high throughput methods based on new generation sequencing, such as metabarcoding and metagenomics, became invaluable tools in the study of microbial diversity during the last decades, and technical advances in sequencing and computational approaches were reflected in the exploration of archaeal diversity (for example Dombrowski et al., 2020; Farag et al., 2020; Kerou et al., 2021).

Metabarcoding of the 16S rRNA gene has been applied to large-scale ecosystemic studies in the marine realm such as the TARA Oceans expedition (Bork et al., 2015) and Deep Carbon Observatory/Census of Deep Life (Schiffries et al., 2019), based mostly on two primer sets targeting the V4V5 region of the 16S rRNA gene. The first one (Parada et al., 2015) is designed to capture both bacterial and archaeal diversity in marine environments, while the second primer set, first proposed by Topçuoğlu et al. (2016), selectively amplifies archaea, and has been employed in several marine subsurface studies (Trembath-Reichert et al., 2019; Hoshino et al., 2020; Teske et al., 2021). A number of studies have highlighted the critical importance of primer design and the impact that even a single base pair mismatch has on the reliability of metabarcoding results (Bru et al., 2008; Parada et al., 2015). In particular, the consequent gaps in representative sequences used for primer design for Archaea, whose diversity is still being uncovered, may lead to significant blindspots in amplification (Eloe-Fadrosh et al., 2016; Bahram et al., 2019). Besides introducing possible bias in the detection of some lineages depending on primer adequacy, the PCR required for metabarcoding may distort the relative abundance of lineages in the molecular results compared to the environmental communities (Suzuki and Giovannoni, 1996; Polz and Cavanaugh, 1998; Kobschull and Zador, 2015; O'Donnell et al., 2016).

Thus, approaches based on whole genome sequencing (metagenomics) have been put forward as an alternative to metabarcoding because of their reduced PCR bias (Logares et al., 2014), at the cost of more complex bioinformatic processes to reconstruct marker genes or genomes. The inventory of the diversity of archaeal lineages in a given environment can be

CHAPTER 1

achieved through the extraction of small subunit (SSU) rRNA sequences, also called miTAGs, from the unassembled metagenomes (Kopylova et al., 2012; Gruber-Vodicka et al., 2020), or through the detection of single copy core genes (SCG) after assembly of the raw reads (Darling et al., 2014; Parks et al., 2015).

To contribute to the direction of future efforts to describe deep sea benthic archaeal communities, we compared the results obtained with i) the two primer sets targeting 16S rRNA V4V5 region, ii) miTAGs and iii) SCG profiles on 46 surface sediment samples (top 30 cm) from 6 stations in two south Pacific hadal trenches and adjacent abyssal plains. We hypothesized that both metagenomic methods would perform best, by being less subjected to PCR bias, and that the archaeal primers would be more appropriate for the high resolution characterization of deep sea archaeal sedimentary communities. We present here the results of our assessment of the relative advantages and weaknesses of each method in terms of inventory of the overall benthic archaeal diversity, detection of specific lineages, and of potential application to beta-diversity studies.

Material & Methods

1. Environmental samples collection

The samples in this study were collected during two cruises in November 2017 and March 2018 as part of the HADES-ERC project. During the first cruise targeting the Kermadec trench, north of New Zealand, two sites were sampled: one in the trench (9555 meters below sea-level (mbsl)) using a multicorer, and the other on the adjacent subducting plate (6080 mbsl) using a boxcorer. The second set of samples originated from the Atacama trench, off the coast of Chile, with two trench sites (7770 and 7915 mbsl) and two sites located on the adjacent abyssal plain (5500 mbsl) or continental margin (4050 mbsl). All samples from the Atacama trench were collected using a multicorer. Triplicate cores were recovered for one Atacama trench site and the adjacent abyssal plain site to assess fine-scale variations in diversity.

The recovered sediment cores were sliced immediately after getting onboard in a 3°C cold room using equipment previously bleached and rinsed with ethanol and nanopure water. Each core was sliced into depth layers following a standard scheme (0-1 cm, 1-3 cm, 3-5 cm, 5-10 cm, 10-15 cm, and 15-30 cm), with always at least 1 cm left on the extruder to avoid contamination. Slicing was performed using spatulas also bleached and rinsed with nanopure water before each use. Samples were then transferred into zip-lock bags, homogenized, and frozen at -80°C on board before being shipped on dry ice to the laboratory where they were also kept at -80°C. Empty bags were also conditioned on board to be later used as sampling controls.

CHAPTER 1

2. DNA extraction

DNA extractions were performed in a sterile shore lab, using approximately 10 g of sediment with the PowerMax Soil DNA Isolation Kit according to the manufacturer's instructions (MO BIO Laboratories, Inc.; Qiagen, Hilden, Germany) with modifications: the elution buffer was left on the spin filter membrane for 10 min at room temperature before centrifugation in order to increase DNA yield. Extraction controls were added by using an empty tube from the kit for each series of extraction or extraction kit batch. In total 7 extraction negative controls were produced. In addition, field controls were prepared onboard with the first solution of the kit poured into the control ziplock bag before following the usual extraction steps. Each of the resulting 5-mL DNA solutions were stored at -80°C.

3. Libraries construction and sequencing

3.1. Metabarcoding

Two primer sets targeting the V4V5 region of the 16S rRNA gene were used to generate amplicon datasets: a universal primer pair (515F-926R) (Parada et al., 2015) and an archaea-specific primer mix (517F-958R) (Topçuoğlu et al., 2016) (Table 1). PCR amplifications were carried out at Génoscope (Evry, France) as part of the eDNAbyss project (see Chapter 1 Supplementary Material for amplification details). Three amplicon libraries were prepared for each sample by non-directional ligation of Illumina adapters on 100 ng of amplicons following the Kapa Hifi HotStart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA, USA). After quantification and quality control, replicate libraries of each sample were pooled, library concentrations were normalized to 10 nM, and 8–9 pM of each library containing a 20% PhiX spike-in were sequenced on a HiSeq2500 (System User Guide Part # 15035786) instruments in a 250 bp paired-end mode.

CHAPTER 1

Table 1: Primer sequences

Universal primer set (Parada et al., 2015)	(515F-Y) 5'-GTGYCAGCMGCCGCGGTAA	(926R) 5'-CCGYCAATTYMTTTRAGTTT
Archaea-specific primer set (Topçuoğlu et al., 2016)	(517F) 5'-GCCTAAAGCATCCGTAGC	(958R) 5'-CCGGCGTTGANTCCAATT
	(517F) 5'-GCCTAAARCGTYCGTAGC	
	(517F) 5'-GTCTAAAGGGTCYGTAGC	
	(517F) 5'-GCTTAAAGNGTYCGTAGC	
	(517F) 5'-GTCTAAARCGYYCGTAGC	

3.2. Metagenomics

Metagenomic libraries were prepared from 10 ng or less of the same DNA extracts as before with the NEBNext Ultra II DNA Library prep kit (New England Biolabs, MA, USA). After quantification and quality control, library concentrations were normalized to 10 nM and applied to cluster generation according to the Illumina Cbot User Guide (Part #15006165). Sequencing of libraries was performed according to the Novaseq 6000 System User Guide Part #20023471 (Illumina, San Diego, CA, USA) in paired-end mode (2x150 bp). See Supplementary Material for the details of library preparation and sequencing carried out at Génoscope (Evry, France).

4. *In silico* primer specificity evaluation with TestPrime

To evaluate the *a priori* performance of the primer sets considered, we used TestPrime 1.0 (Klindworth et al., 2013) via the arb server (<https://www.arb-silva.de/search/testprime>) and ran *in silico* PCRs on the SILVA non redundant SSU database version 138, allowing no mismatches. We ran the *in silico* PCR six times, once for each possible primer pair, since the universal primer set is a simple pair while the archaea-specific primer set is composed of a mix of five forward primer variants and one reverse primer. TestPrime provided estimates of

CHAPTER 1

the coverage of primers for all available taxonomic groups present in the database. We focused on the results at domain level and archaeal phylum level.

5. Amplicon datasets bioinformatic analysis

Bioinformatic analyses of the metabarcoding datasets were performed using a standardized pipeline (Brandt et al., 2021), available on Gitlab (<https://gitlab.ifremer.fr/abyss-project/abyss-pipeline>), on a home-based cluster (DATARMOR, Ifremer).

Due to non-directional adapter ligation, inserts were sequenced in different orientations. After demultiplexing and renaming steps, we used Cutadapt v1.9 (Martin, 2011) to identify the primer sequence in each read and sort them according to two criteria: forward or reverse primer and forward or reverse sequencing. Data for each sample was thus split into 4 sequence files (R1F, R1R, R2F, R2R). Cutadapt then removed the identified primer sequences and BBMAP repair (Bushnell, 2014) was used to ensure that reads were still paired by sorting reads using the information present in their description line and removing unmatched reads.

For each sequencing run, we determined Amplicon Sequence Variants, merged read pairs and removed chimeras using the DADA2 package v.1.10 (Callahan et al., 2016), following guidelines from the online tutorial for paired-end HiSeq data (<https://benjjneb.github.io/dada2/bigdata.html>). The script implementing DADA2 was applied separately to the two pairs of sequence files R1F/R2R and R2F/R1R. The parameters used for filtering and trimming reads were as follows: truncation length of 220 base pairs, maxN= 0, maxEE= 2 and truncQ= 11. The error learning step was based on nbases= 1e8. Merged sequences were size-filtered by keeping sequences with a length between 350 and 390 bps. The Amplicon Sequence Variants (ASVs) tables produced by DADA2 for each run were then merged, collapsing ASVs based on DNA sequence identity. Taxonomic assignment was performed with the implementation of the RDP naive Bayesian classifier (Wang et al., 2007)

CHAPTER 1

available in DADA2 v.1.10, using the SILVA v138 reference database (Quast et al., 2013) and a bootstrap threshold of 80.

The ASV and taxonomy tables produced by this pipeline were then combined in a phyloseq object (phyloseq v1.28.0, McMurdie and Holmes, 2013) in an R v3.6.1 environment. Reads from the same amplicon library, but originating from different Illumina runs, were merged under the same sample name before removing sequences from unwanted taxa (Eukaryota, Chloroplast and Mitochondria affiliated sequences). The dataset was decontaminated using extraction, PCR and field controls with the decontam package (v1.4.0, Davis et al., 2018), or handpicking in the case of the ASV dominating bacterial control libraries reads (see reproducible workflow on github). Samples totaling less than 80,000 reads after decontamination were removed, the appropriate metadata added, and the final object saved as a phyloseq object for further analysis in R.

6. Metagenomic reads analysis

6.1. Ribosomal SSU sequences (miTags)

The quality filtration of the demultiplexed metagenomic raw reads was carried out with Illumina-Utills python scripts (Eren et al., 2013b) following recommendations by Minoche et al. (2011). We then used Phyloflash v3.3b3 (Gruber-Vodicka et al., 2020) to identify SSU (short subunit) rRNA fragments using SILVA v138 database as reference, with default parameters and a clustering id of 100%. Shortly, Phyloflash maps reads to a reference database (here SILVA v138) using BMap (Bushnell, 2014) and extracts sequences based on a minimum identity threshold, 70% by default. It can also implement different assemblers to try and reconstruct 16S sequences, but in our case such reconstruction was unsuccessful for the archaeal part of the community.

CHAPTER 1

In addition to the taxonomic assignment inherited from Phyloflash's algorithm, we also assigned taxonomy to the SSU rRNA fragments (or miTAGs) using the NBC (Naive Bayesian Classifier) implemented in DADA2 (Callahan et al., 2016). To do this, we filtered out miTAG sequences shorter than 50 bp, dereplicated them and then used the NBC with SILVA v138.1. We then removed the sequences affiliated with Eukaryota, Mitochondria or Chloroplasts from both resulting miTAG tables, before saving them as phyloseq objects for further analysis.

6.2. Single copy core genes

Finally, we obtained a single copy core gene observation table from the metagenomes using the following method. We assembled each quality-filtered metagenome using Megahit (Li et al., 2015) with preset meta-sensitive and minimum contig length of 1000 bp. We then used the Anvi'o metagenomics snakemake workflow (Köster and Rahmann, 2012; Shaiber et al., 2020; Eren et al., 2021) to construct a contigs database, map back the metagenomic reads on the corresponding assembly (Langmead et al., 2009; Danecek et al., 2021), identify genes (Prodigal, Hyatt et al., 2010) and finally examine taxonomic profiles of our metagenomes by comparing the identified single copy core genes to the GTDB database (Buchfink et al., 2015; Parks et al., 2018, 2020). The universal single-copy core genes considered here were a set of 22 ribosomal protein genes as implemented in Anvi'o v7.

7. Comparison of the datasets

Generation of the phylum-level taxonomic plots and evolution of the ratio of Archaea in samples was done in R v3.6.1, using phyloseq (v1.28.0, McMurdie and Holmes, 2013) and ggplot2 (v3.3.0, Wickham, 2016) packages. We represented the relative composition of the datasets at higher taxonomic ranks around a dendrogram using Graphlan (Asnicar et al., 2015).

CHAPTER 1

To account for phylogenetic distance in dataset comparison, we placed the sequences from the three datasets in the SILVA v138 reference archaeal tree extracted from the full reference tree using ARB (Ludwig et al., 2004). To this end, we first extracted all ASV sequences affiliated to Archaea from both metabarcoding datasets. For miTAGs, we used SINA's "search and classify" option (v1.6.1, Pruesse et al., 2012) to filter out bacterial sequences. We then aligned the full set of query sequences and tree sequences using MAFFT with default parameters (v7.273, Katoh, 2002), and masked the positions with more than 95% gaps using the *easel* functions in HMMER (Johnson et al., 2010). Finally, we generated the phylogenetic placement file using EPA-ng (Barbera et al., 2018).

The resulting phylogenetic placement file was analyzed using *gappa* (Czech et al., 2020). It was first split into 56 independent files by sample according to the ASV table using *gappa edit split* function. We then computed the edge PCA with function *gappa analyze edgepca*. Finally, we visualized the resulting projection of samples into principal coordinate space in R v3.6.1 with *ggplot2*.

Results

1. *In silico* primer specificity evaluation using SILVA v138 16S rRNA database

We first assessed the potential coverage of the two primer sets on the SILVA v138 database, allowing no mismatches (Table 2). As expected, the archaea-specific primers presented no coverage of the diversity of domains Bacteria and Eukaryota. They also matched at most 61.1% of the overall database of archaeal sequences, even combining all 5 primer pairs. Conversely, the universal primers showed comparably good coverage on the three domains, between 80.8 and 84.5.

The archaeal primer mix exhibited a generally lower coverage in all archaeal phyla except for Aenigmarchaeota, Euryarchaeota and Korarchaeota, far outperforming the universal primers in the latter (77 to 14.8). On the contrary, there was a strong discrepancy in coverage for phyla Hadarchaeota, Iainarchaeota and Nanoarchaeota in favor of the universal primers, with the archaeal set matching respectively 13.2 to 18.4%, 13.6 to 15.9% and 0.2% of database sequences when universal primers matched 77.3%, 46.7% and 59.7%. In addition to these differences in specificity, both primer sets exhibited good coverage of Asgardarchaeota, Euryarchaeota and Hydrothermarchaeota, and both performed poorly on sequences affiliated with phyla Aenigmarchaeota, Altiarchaeota and Nanohaloarchaeota (no sequence coverage).

CHAPTER 1

Table 2: Potential coverage of primer pairs determined using TestPrime 1.0 against the SILVA RefNR database v138, at domain level, and phylum level for Archaea.

Primer pairs \ Taxon	Universal primers	Archaeal primer 1	Archaeal primer 2	Archaeal primer 3	Archaeal primer 4	Archaeal primer 5	Overall archaeal primers
Bacteria (domain)	84.5	0	0	0	0	0	0
Eukaryota (domain)	80.8	0	0	0	0	0	0
Archaea (domain)	81.0	10.3	34.2	5.6	9.2	1.8	34.2 - 61.1
Aenigmarchaeota	3.5	2.3	5.3	0.3	4.6	16.5	16.5 - 29
Altiarchaeota	7.5	0	0	0	2.4	0	2.4
Asgardarchaeota	86.2	0.6	1.5	0	65.4	5.4	65.4 - 72.9
Crenarchaeota	83.4	29	9.4	0	11.9	0.4	29 - 50.7
Euryarchaeota	78.4	0	81.4	0	0.9	1.3	81.4 - 83.6
Hadarchaeota	77.3	0	13.2	0	0.3	4.9	13.2 - 18.4
Halobacterota	86.8	0	46	20.2	7.9	2.4	46 - 76.5
Hydrothermarchaeota	86.3	0	79	0	1.0	0	79 - 80
Iainarchaeota	46.7	0	2.3	0	13.6	0	13.6 - 15.9
Korarchaeota	14.8	0	77	0	0	0	77
Micrarchaeota	28.2	0	2.9	0	1.5	5.9	5.9 - 10.3
Nanoarchaeota	59.7	0	0	0	0	0.2	0.2
Nanohaloarchaeota	0	0	0	0	0	0	0
Thermoplasmatota	82.1	0.1	54.7	0	2.4	2.3	54.7 - 59.5

2. Datasets description

The environmental dataset used to compare the methods was composed of 60 sediment samples, 30 samples collected from three hadal sites and 30 samples recovered from three abyssal sites. As highlighted in the methods section, these samples came from two different trench regions, triplicate cores were recovered for one hadal and one abyssal site, and the cores were sliced up to 30 cm below the seafloor (cmbsf). Construction of metagenomic

CHAPTER 1

libraries failed for 4 of these 60 samples (Table S3), which were subsequently discarded from the metabarcoding datasets.

2.1. Universal primers dataset

A total of 56 sequence files were thus obtained after amplification with the universal primers and sequencing, producing 47,627,181 raw sequences of the 16S rRNA gene V4-V5 region, with a mean of 850,485 reads by library. A total of 17,360,183 reads were recovered from the additional 23 control libraries that were constructed and sequenced simultaneously, including sampling (empty bags of storage conditioned on-board research vessels at the end of some of the sampling sessions), extraction (empty kit processed through all extraction steps together with the samples) and PCR (nanopure water) controls.

After processing with DADA2 the dataset included 42,469,064 sequences distributed among 81,066 ASVs. Of those, 2019 (about 2.5%) were found in control libraries, with 1460 ASVs specific to these libraries. A specific ASV accounted for 99% of the negative control libraries. This ASV, affiliated with partial 16S sequences of *Sphingobium* strains, has been recognized by the manufacturer as a pervasive contaminant of Taq-Phusion reagents (Salter et al., 2014), a contamination that occurred in all commercial kits up to 2019.

After bioinformatic processing, decontamination and reduction to the 46 samples that successfully went through bioinformatic processing for the three methods, the dataset comprised a total of 46 libraries with 22,672,216 sequences representing 69,157 ASVs. 16.4% of these ASVs (11,320 ASVs adding up to 16.7% of sequences) were assigned to domain Archaea.

CHAPTER 1

2.2. Archaeal primers dataset

Only 48 out of the 56 sample libraries were obtained by amplification with the archaeal primer mix and sequenced, producing 29,150,201 raw sequences of the 16S rRNA gene V4-V5 region, with a mean of 607,296 reads by library. A total of 602,022 reads were recovered from the additional 21 control libraries. After processing with DADA2 the dataset was composed of 15,718,397 sequences and 3,239 ASVs. Of those, 171 (about 5.3%) were found in control samples, representing less than 1% of the total number of reads. After bioinformatic processing, decontamination and reduction to samples successfully processed for the three methods, the dataset comprised 46 libraries including 15,458,643 sequences representing 2,928 archaeal ASVs.

For both amplicon datasets, rarefaction curves were plotted (Fig. S1) and confirmed that sampling and sequencing efforts captured most of the sample diversity.

2.3. miTAGs dataset and SCG profiles

The metagenomic dataset was composed of 56 metagenomic libraries with a mean of 159,868,288 raw reads. On average, 0.037% of the quality filtered reads were identified as miTAGs by Phyloflash.

When taxonomy was assigned using Phyloflash, the dataset reduced to the 46 samples successfully amplified and processed for the three methods comprised 8501 phylotypes, out of which 386 were archaeal and added up to 277,544 sequences. By contrast, assigning taxonomy to the sequences using the Naive Bayesian Classifier implemented in DADA2 led to a dataset composed of 1483 microbial phylotypes, among which 121 were archaeal and represented by a total of 225,439 sequences.

After metagenome assembly and annotation, we observed that ribosomal protein S11 was the most abundant single-copy core gene in our dataset. We produced SCG-based taxonomic profiles with several core genes with similar outcomes and thus present here only the results

CHAPTER 1

based on S11 sequences. On average, we found a coverage of 3620.42x of ribosomal protein S11 in our metagenomes, varying between 1993.24x and 7506.57x. For Archaea, the mean coverage in these samples was 810.97x.

3. Comparative taxonomic profiles on deep-sea benthic samples

We extracted the percentage of Archaea-assigned sequences from all datasets whenever possible and examined its variation across samples (Fig. 8A). There was a higher proportion of archaea in abyssal samples than in hadal ones (22.1%, 15.4%, 7.8%, 33.1% vs 12.8%, 7.7%, 3.7% and 12.3% respectively with the universal primer pair, miTAG (phyloflash), miTAG (NBC) and SCG dataset, $p\text{-value} < 2.10^{-6}$). The datasets yielded different estimates of archaeal relative abundance for each sample, however the values based on the 16S rRNA gene, be it using universal primer amplicons or miTAGs, followed a remarkably similar pattern across samples (Fig. 8A). On the contrary, the relative abundance of archaea estimated by SCG profiles showed sharp and singular variations, ranging from 1.4% to 49.6% of the sequences identified, and delivered lower estimates in the deep horizons of the hadal sites and of abyssal site A9 in particular. This drop in archaeal ratio seemed to be due to a taxonomic bias of the recovered SCGs, as it coincided with more diverse samples (i.e. less dominated by Nitrososphaeria) for the universal primers and miTAG dataset, with diversity that was mostly missed by the SCGs (Fig. 8F).

In terms of taxonomy, all datasets indicated the predominance of class Nitrososphaeria (previously part of phylum Thaumarchaeota), especially in the abyssal zone (Fig. 8B-F). Nanoarchaeota represented an important part of the universal and miTAG datasets (11.7 to 22%) (Fig. 8C, D and E). Contrastingly, they were not detected in the archaeal primers dataset (Fig. 8B) and barely present in the SCG data (Fig. 8F). Overall, the taxonomic profiles obtained with the universal primer amplicon dataset were remarkably similar to the ones obtained based on unassembled metagenomic data (miTAGs). They differed mostly in the relative percentage

CHAPTER 1

of sequences assigned to the relatively rare Hydrothermarchaeota, these sequences making up 3 to 3.8% of the miTAG dataset, depending on the taxonomic assignment method employed, and only 1% of the universal primer dataset (Fig. S1). Asgardarchaeota sequences were also less represented in the universal primer dataset than in the profiles obtained from the miTAGs assigned using Phyloflash, though this might be an artefact of the taxonomic affiliation method since this discrepancy was also observable when assigning miTAG sequences using DADA2's implementation of the NBC. This last method of taxonomic assignment was included for comparison purposes, but yields a non-negligible amount of unassigned sequences, both at domain level (17%) and at deeper taxonomic ranks.

Apart from their lack of Nanoarchaeota detection, the archaeal primers produced few sequences affiliated with Asgardarchaeota (Fig. 8B).

Finally, the taxonomic profiles obtained from assembled metagenomic data through the most abundant single copy core gene (ribosomal S11 gene here) captured little of the archaeal diversity, with only 4 phyla represented (Fig. 8F), and was thus not included in further comparisons. Sequences affiliated with Asgard-, Nano-, Had-archaeota and Bathyarchaeia were only detected in samples from the deeper horizons of hadal sites and abyssal site A9.

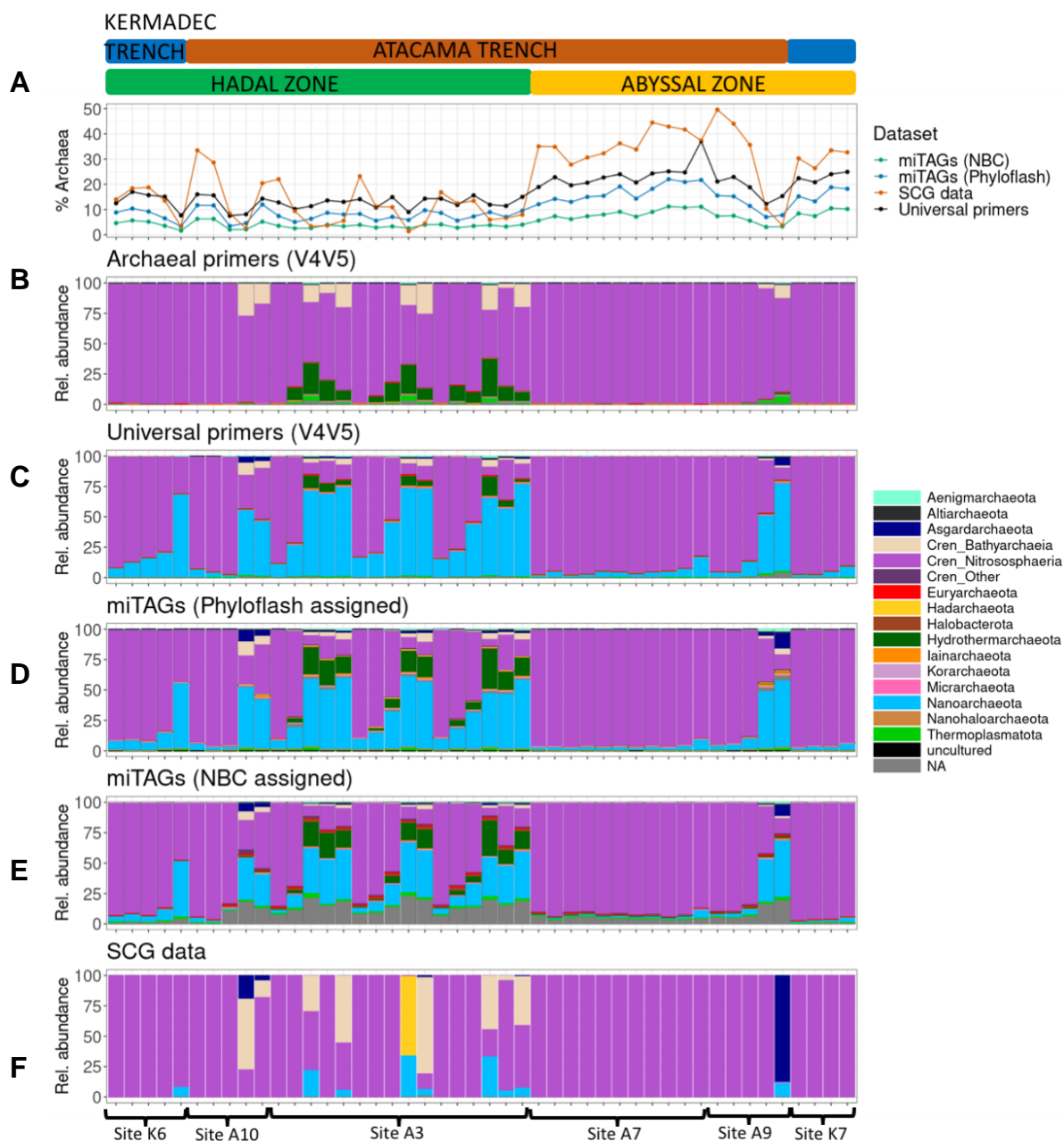


Figure 8: Composition of datasets at domain and phylum level. **A)** Percentage of sequences identified as Archaea in the three datasets detecting both Bacteria and Archaea (miTAGs are represented after taxonomic assignment with two methods). Archaeal phyla diversity profiles in the four datasets considered: **B)** archaea-specific amplicon dataset, **C)** universal amplicon dataset, **D)** miTAG dataset assigned with Phyloflash, **E)** miTAG dataset assigned with DADA2’s implementation of the Naive Bayesian Classifier and **F)** SCG profiles based on Ribosomal gene S11. The samples are ordered by increasing depth in the sediments for each site, and phylum Crenarchaeota has been split between Bathyarchaeia, Nitrososphaeria and Others for clarity.

CHAPTER 1

In order to better understand the specific differences in the results obtained with the metabarcode and miTAG methods, we represented the partition of each dataset at deeper taxonomic levels on a dendrogram (Fig. 9).

Once again, the lack of detection of Nanoarchaeota (order Woesearchaeales) by the archaeal primers was evident (Fig. 9B), as well as the predominance in all datasets of Nitrososphaeria, and more precisely *Nitrosopumilaceae*, most of them assigned to *Candidatus Nitrosopumilus* (Fig. 9). Indeed, in all datasets the sum of the relative abundance of Nanoarchaeota and Crenarchaeota (defined based on SILVA 138) added up to more than 93% of the archaeal community (Fig. S2) except for the NBC-assigned miTAGs, for which it amounted to 87.17%, probably due to the higher percentage of unassigned sequences. In this last dataset, the proportion of unassigned sequences was highest at all taxonomic levels, although it was overall relatively high also with other methods, with at best 68,9% of sequences assigned at genus level in the archaea-specific amplicon dataset (Fig. S6). Interestingly, Phyloflash was the method that assigned sequences to the largest diversity of genera, most of these being uncultured representatives.

Differences in detection between the metabarcoding and metagenomic datasets were observable in the less abundant taxa: Halobacterota and Euryarchaeota (Methanofastidiosales) were assigned less than 0.1% of the sequences (at least 1 order of magnitude less than in the miTAG datasets), and Thermoprotei were not detected in the ASV data. Conversely, Aenigmarchaeota members (Deep Sea Euryarchaeotic Group and Aenigmarchaeia) had very similar relative abundance profiles across all datasets.

CHAPTER 1

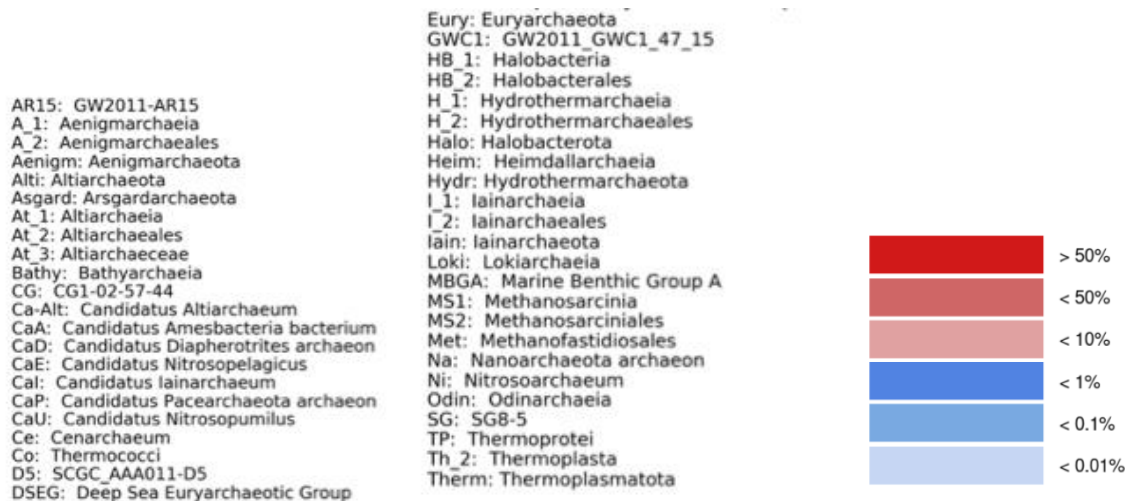


Figure 9: Dendrogram of the taxa representing more than 0.025% of the datasets, with nodes colored in green if found in the specific dataset presented: **A)** universal amplicon data, **B)** archaea-specific amplicon data, **C)** miTAG dataset assigned using Phyloflash and **D)** miTAG dataset assigned using DADA2's implementation of the Naive Bayesian Classifier. Pixel-shaped nodes represent uncultured archaeons. The heatmap layers around the dendrogram represent the percentage of the dataset made up by the considered taxon at a given taxonomic rank. Grey color indicates no detection.

4. Multi-dimensional analysis based on phylogenetic distance

Considering that taxonomic assignment methods may blur the interpretation of the differences among diversity characterization methods, we also compared the datasets based on phylogenetic distance between sequences (Fig. 10).

Abyssal samples for the three datasets clustered together. Similarly to the results presented in Fig. 8 and 9, differences were observed for the hadal sites. Surface horizon samples for the metabarcoding datasets seemed to cluster closer together than with miTAGs, at the top of the ordination space. However, samples from the deep horizons of the hadal sites and abyssal site A9 formed a distinct group on the left side of the ordination space for the universal primer dataset and the miTAG dataset (Fig. 10D).

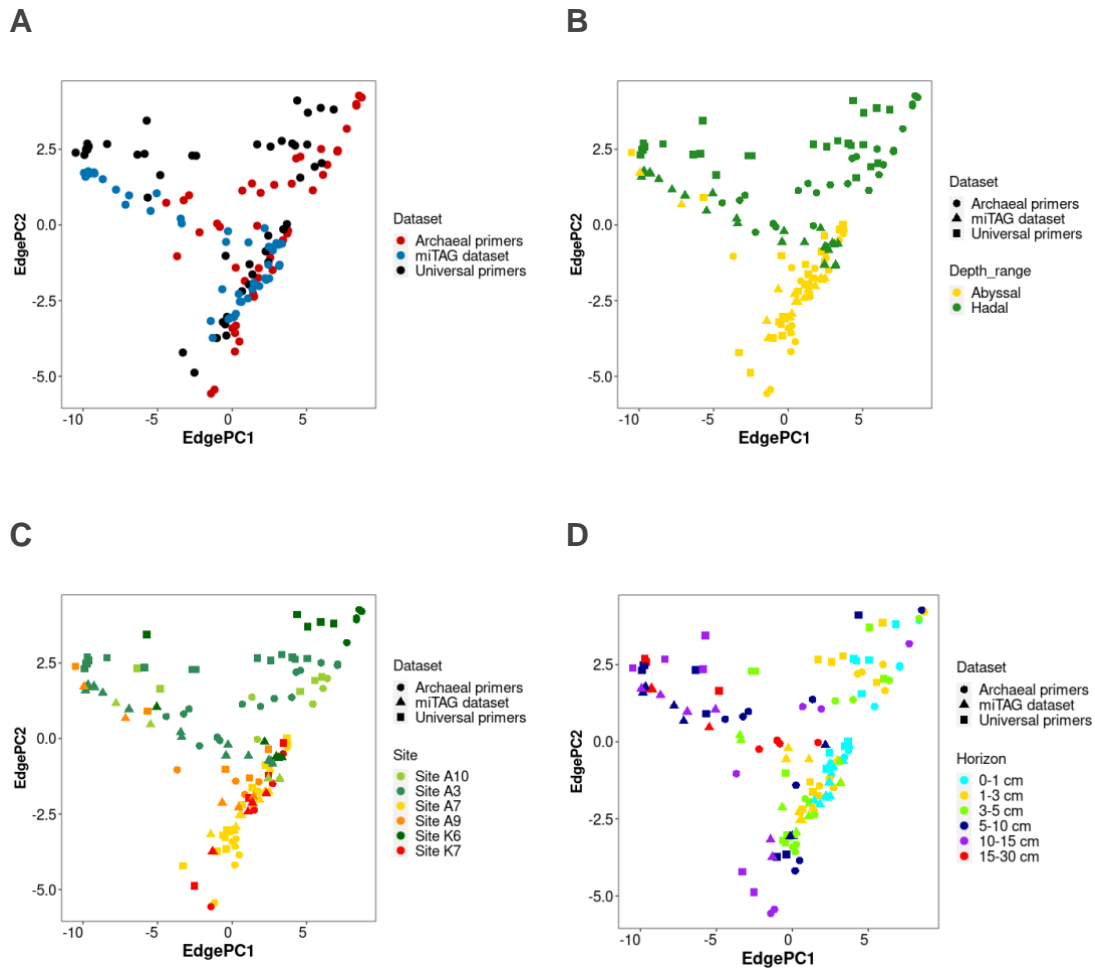


Figure 10: Edge Principal Components Analysis of the three datasets (universal primers data, archaeal primers and miTAGs) based on placement of the sequences in the SILVA reference tree. The dataset is illustrated by shape, and the points are colored according to **A)** dataset, **B)** the depth range of the sampling sites (hadal or abyssal), **C)** the sampling sites and **D)** the sediment horizon.

Discussion

Screening for the optimal way to inventory the biodiversity of deep sea benthic archaeal communities, this study based on the comparison of two 16S rRNA V4V5 metabarcoding datasets with metagenome-extracted SSU rRNA sequences (miTAGs), and single copy core gene profiles, supported the accuracy of universal metabarcoding primers (Parada et al., 2015) and miTAGs. Both delivered highly similar results in terms of community composition, while single copy core gene profiles failed in uncovering the diversity of archaeal lineages. More surprisingly, the set of archaea-specific primers chosen for screening here because of its wide use in subsurface studies showed important diversity coverage gaps both *in silico* and on the environmental samples used for this comparison.

miTAGs have been proposed as an alternative to 16S amplicons to avoid targeted PCR biases (Logares et al., 2014). Here we used them as a starting point to assess the performance of the two metabarcoding primer sets, universal and archaea-specific. The taxonomic profiles (Fig. 8B, C, D and E) showed that the results obtained with the universal primers were remarkably comparable to the miTAGs results, a result reinforced by the Edge PCA ordinations (Fig. 10), particularly for deeper hadal sites. They mostly differed from the archaeal primer dataset in the detection of phylum Nanoarchaeota. We also showed here that this poor detection by the archaea-specific primer set could be expected from the results of the *in silico* PCR using SILVA TestPrime, which predicted their coverage of Nanoarchaeota at 0.2% (Table 2). On the contrary, the *in silico* results at phylum level did not reflect the difference in detection of Hydrothermarchaeota in favor of the archaeal primers that was observed on real data. This highlights the fact that these *a priori* results need to be confirmed through experimental data since a single base pair mismatch between primer and sequence can lead to a significant skewing of the results (Bru et al., 2008; Parada et al., 2015; Eloë-Fadrosh et al., 2016). It also means that these results need to be examined with the aim of the study in mind, since a different conclusion might be reached depending on the taxa of interest.

CHAPTER 1

Nanoarchaeota -more precisely Woese archaeales- and *Nitrosopumilaceae* (part of the Thaumarchaeota phylum in former SILVA 132 release) dominated the archaeal communities found in our samples, a trend also observed in previous hadal studies (Cui et al., 2019; Peoples et al., 2019). On the less abundant phyla, the archaeal primers showed again important coverage gaps compared to their counterparts (Fig. 9B), with a lower percentage of sequences assigned to Iain-, Alti-, Eury-, and Asgard-archaeota, expected *in silico* for the first two lineages (Table 2). In addition, a bias in both ASV datasets was visible in their lower detection of Halobacterota, and the lack of Thermoprotei (Fig. 9A, B). It was difficult however to determine whether some differences in abundance profiles resulted from PCR or assignment biases, since the profiles were similar between universal ASV data and NBC-assigned miTAGs, but differed from Phyloflash-assigned miTAGs (e.g. *Candidatus Altiarchaeum*, Heimdall- and Odin-archaeia, and Bathyarchaeia).

Other blindspots that might exist in the metabarcoding ASV data, according to the TestPrime results (Table 2), are a lower coverage of the diversity in phyla Aenigm-, Micr- Kor-, and Nanohalo-archaeota. The last two were indeed undetected, yet we acknowledge that they represented a very small fraction of the sequences extracted from the metagenomes (0.00044% to 0.00108% and 0.0036% respectively) (Fig. S1). Though it is probable that both sets of primers tested here would present coverage gaps for these phyla, it is difficult to conclude. Regarding Aenigmarchaeota, surprisingly, similar abundance profiles were retrieved in all datasets (Fig. 9), which seemed to indicate an adequate performance of both primer sets.

The taxonomic profiles produced from assembled metagenomic data by detecting single copy core genes (results presented here for ribosomal gene S11) diverged from the results based on the 16S rRNA gene (amplicons and miTAGs) (Fig. 8F). Although they showed the same predominance of Nitrososphaeria, they did not accurately capture the diversity of archaeal lineages, or only in deeper horizon samples where these populations, as seen with the other methods, became relatively more abundant (Fig. 8). This is most likely due to the specific challenges of metagenomic assembly (Wang et al., 2019; Lapidus and

CHAPTER 1

Korobeynikov, 2021) and the fact that targeted populations represented a very small fraction of the microbial communities. There is also a discrepancy between the abundance and the genetic diversity of some clades: in the universal dataset, 22% of archaeal sequences were affiliated with phylum Nanoarchaeota but were split in 9628 ASVs, while Crenarchaeota accounted for 75.5% of the sequences but only 758 ASVs. This could partly explain the increased difficulty in reconstructing Nanoarchaeota-affiliated single copy core genes. It also highlights the much wider sequence diversity among the 16S rRNA gene within the Nanoarchaeota phylum compared to those affiliated to Crenarchaeota.

Overall, miTAGs seemed to be an interesting method to establish taxonomic profiles of benthic archaeal diversity. However, since they are extracted from random sequencing of the whole 16S rRNA gene, it is not possible to constitute OTUs (Operational taxonomic units) or ASVs based on sequence identity. This method thus entirely relies on taxonomic assignment followed by aggregation of sequences belonging to the same lineage into phylotypes, or alignment and phylogenetic placement to obtain phylogenetic distances. This can become a problem for beta-diversity or network analysis, where high resolution is needed. In the ecosystem studied here, due to the relative uniformity of taxonomic diversity in all samples (Fig. 8), and the low level of assignment at higher taxonomic levels (31 to 67% unassigned at genus level) (Fig. S6), ordinations based on miTAG phylotypes did not illustrate archaeal community structure as well as ASVs obtained with the universal primers (data not shown).

A rigorous estimate of the proportion of Archaea compared to Bacteria in deep sea benthic sediments cannot be obtained without the help of experimental data such as qPCR or CARD-Fish microscopy. A few studies have described the relative proportion of Bacteria and Archaea in surface sediments of various oceanic regions using these techniques (Molari and Manini, 2012; Giovannelli et al., 2013; Jørgensen et al., 2013; Danovaro et al., 2016). However, they highlighted limitations due to the probes used in the case of CARD-Fish data,

CHAPTER 1

and observed widely varying estimates depending on the region, bathymetric depth and sediment depth, making it difficult to extrapolate these results to our study. Interestingly however, all the datasets exhibited similar trends in the relative proportion of archaea, with the exception of the one based on single copy core genes. miTAG data assigned using the Naive Bayesian Classifier most probably underestimated this proportion since 17% of the sequences were not assigned at domain-level. Nevertheless, the similarity in pattern underlined a reproducibility that makes it possible to compare this ratio among samples characterized using the same method.

Conclusion

The results presented here show that the universal primers are capable of producing diverse taxonomic profiles comparable to the miTAG dataset for archaea found in deep sea surface sediments, with the exception of some of the rarest lineages. They also clearly exhibit better reproducibility than the archaea-specific primer mix tested here, as they seemed less affected by the low amount of DNA extracted from deep-sea sediment samples. Additionally, these primers have been widely used, making our results comparable to other large-scale studies of the ocean microbiome (Peoples et al., 2019; Tara Oceans Coordinators et al., 2020). Thus we argue that, for deep-sea benthic archaeal diversity studies where resolution up to the ASV level is necessary, such as beta-diversity or network analyses, metabarcoding based on these universal primers is a cost-effective and appropriate choice.

This conclusion should be reevaluated depending on the aim of the study, be it a specific target lineage or a different biome where the relative proportion of Archaea might be lower and thus more difficult to access. It supposes a high sequencing depth that might not always be achievable. Thus, depending on the results, it might be appropriate to use a combination of domain-specific primers, or to design new primers, a task that can be informed by the miTAG method considered here, as also proposed by McNichol et al. (2021).

Supplementary Figures

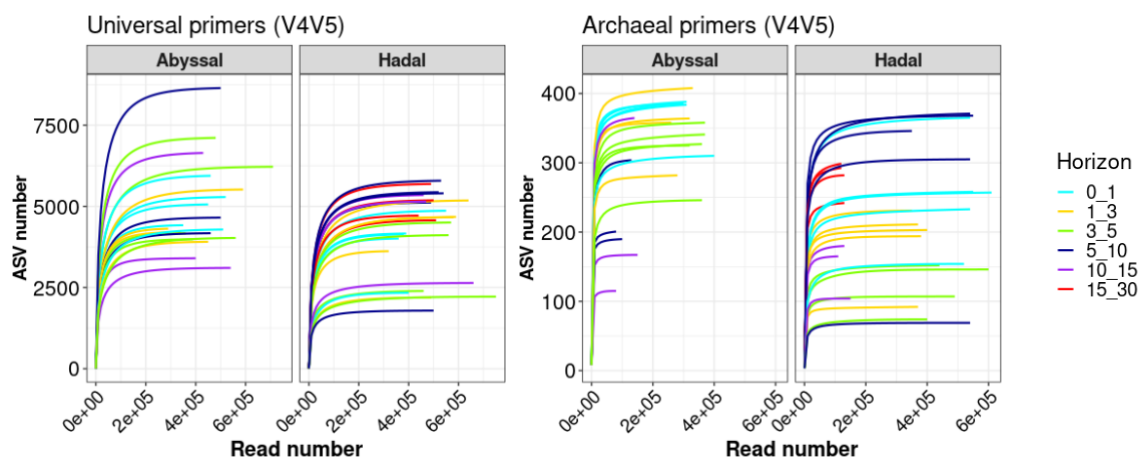


Figure S1: Rarefaction curves for the 2 metabarcoding datasets, colored by sediment horizon.

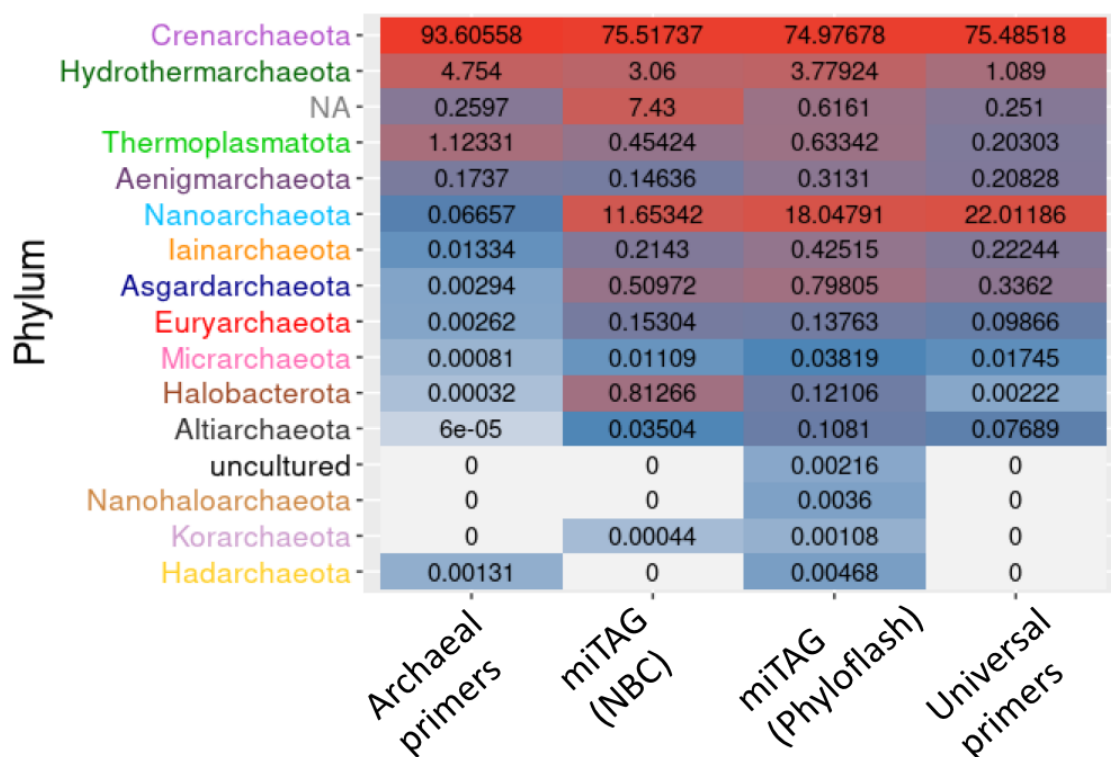


Figure S2: Heatmap of the relative proportion of each phylum in the four datasets: archaeal metabarcoding dataset, miTAGs assigned using the NBC in DADA2, miTAGs assigned in Phyloflash and universal metabarcoding dataset.

CHAPTER 1

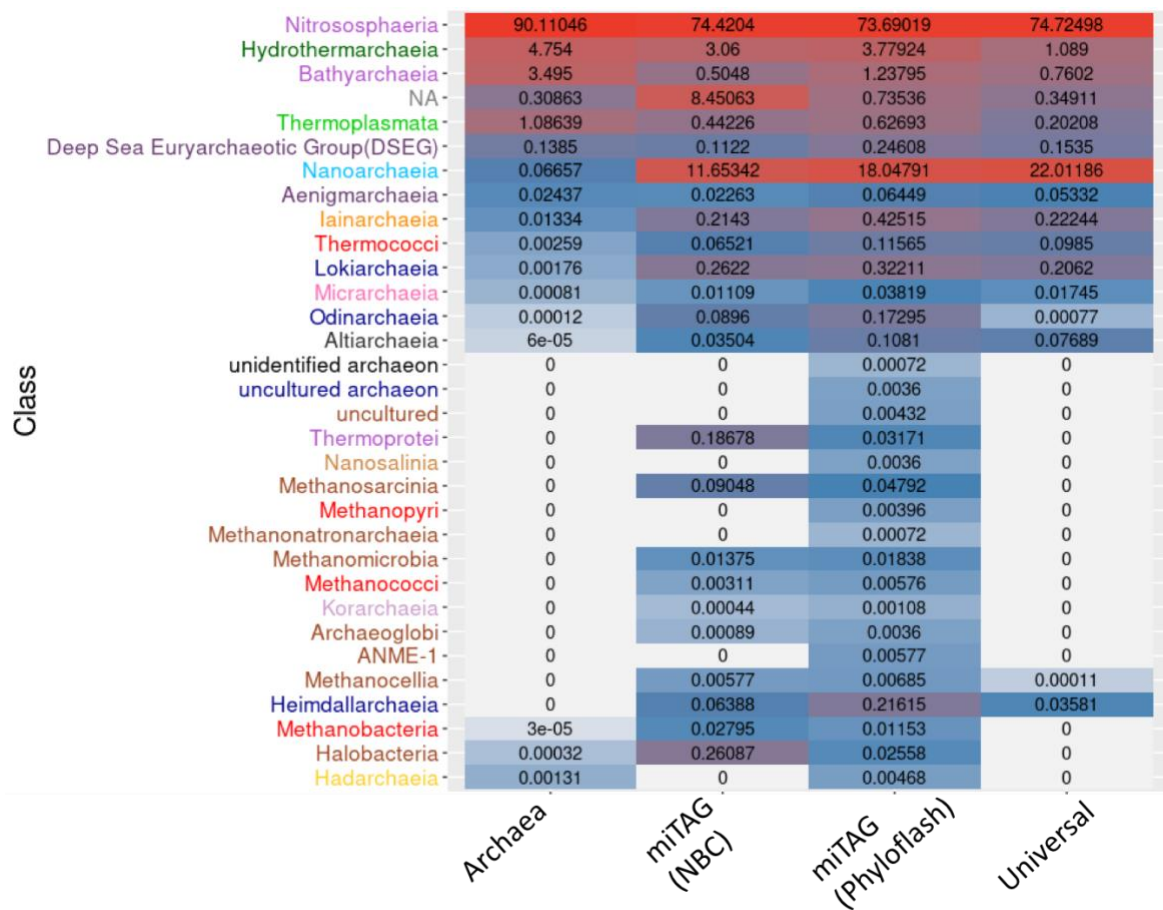


Figure S3: Heatmap of the relative proportion of each taxonomic class in the four datasets. The names of the classes are colored by phylum.

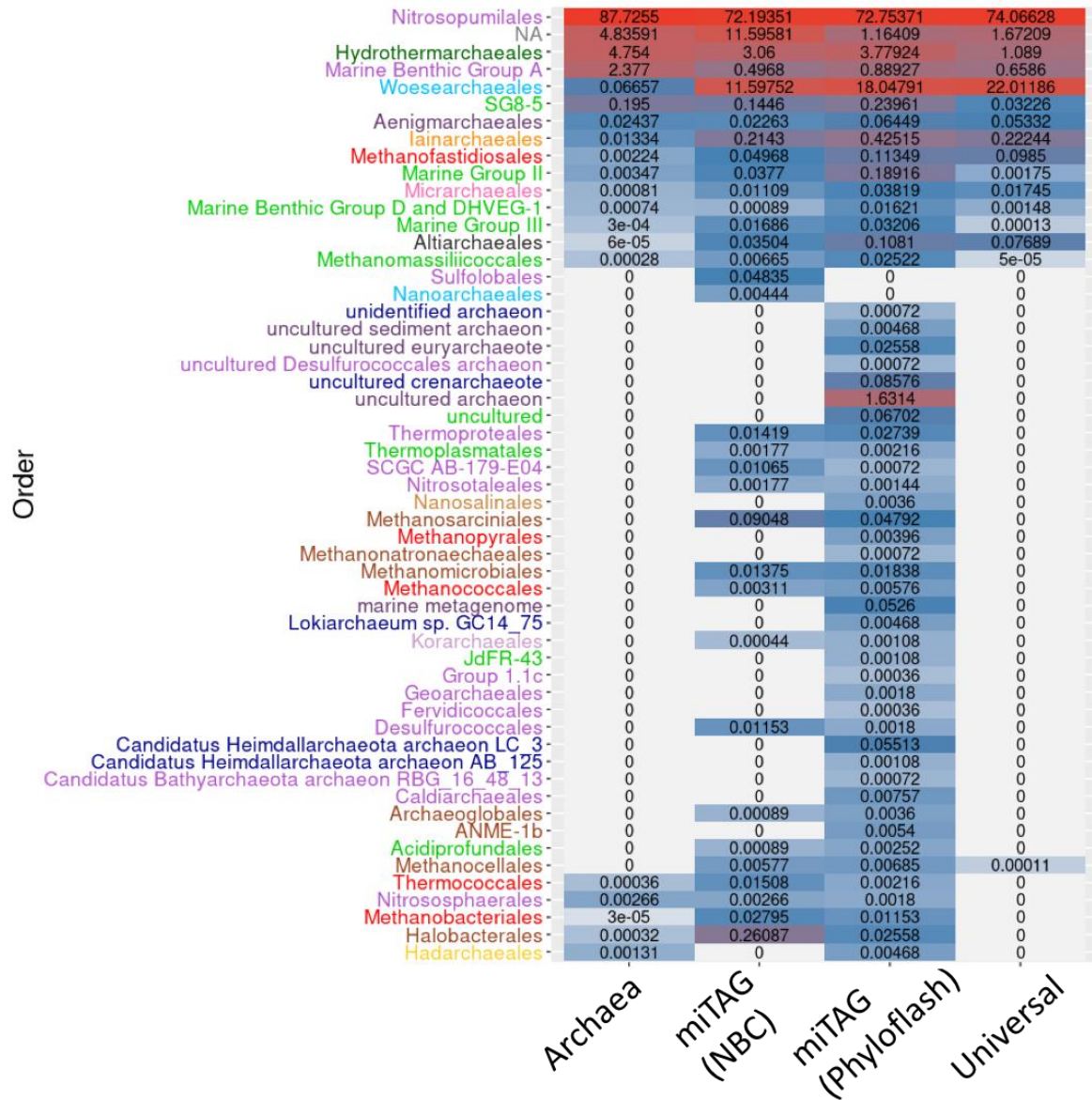


Figure S4: Heatmap of the relative proportion of each taxonomic order in the four datasets. The names of the orders are colored by phylum, see Fig. S3 for legend.

CHAPTER 1

Family	87.7255	72.19351	72.75371	74.06628
Nitrosopumilaceae	12.26352	24.04155	12.52803	16.16939
NA	0.00705	0.1726	0.32896	0.1584
Candidatus Iainarchaeum	0.00351	0.04701	0.23349	0.02871
GW2011	0.00047	0.02528	0.11961	0.01208
CG1-02-57-44	6e-05	0.03504	0.1081	0.07689
Altiarchaeaceae	1e-05	0.00577	0.01441	0.0094
CG1-02-32-21	0	0.04835	0	0
Sulfobacterales	0	0.00044	0	0
Pyrodictiaceae	0	0.00222	0	0
Nanopusillaceae	0	0.00044	0	0
Nanoarchaeaceae	0	0.00133	0	0
Methanomethylphilaceae	0	0.00399	0	0
Acidilobaceae	0	0	0.00036	0
unidentified euryarchaeote	0	0	0.0508	0
unidentified archaeon	0	0	0.00108	0
uncultured Thermoplasmatales archaeon	0	0	0.00108	0
uncultured Thermoplasmata archaeon	0	0	0.00468	0
uncultured sediment archaeon	0	0	0.00036	0
uncultured Nitrosopumilales archaeon	0	0	0.00144	0
uncultured Methanobolus sp.	0	0	0.00432	0
uncultured marine group II/III euryarchaeote KM3_133_A04	0	0	0.00144	0
uncultured marine group II euryarchaeote HF130_40G09	0	0	0.00216	0
uncultured marine group II euryarchaeote DeepAnt-JyKC7	0	0	0.01585	0
uncultured marine group II euryarchaeote	0	0	0.00108	0
uncultured marine euryarchaeote DH148-W1	0	0	0.00036	0
uncultured marine eukaryote	0	0	0.00432	0
uncultured marine crenarchaeote AD1000-23-H12	0	0	0.00937	0
uncultured marine benthic group E euryarchaeote	0	0	0.00252	0
uncultured marine archaeon	0	0	0.00036	0
uncultured haloarchaeon	0	0	1.43573	0
uncultured euryarchaeote	0	0	0.00036	0
uncultured crenarchaeote	0	0	1.21496	0
uncultured bacterium	0	0	0.1095	0
uncultured archaeon W5-61a	0	0	0.05332	0
uncultured archaeon CRA13-11cm	0	0	0.00072	0
uncultured archaeon APA7-17cm	0	0	0.1639	0
uncultured archaeon	0	0	5.50909	0
uncultured	0	0	0.13799	0
Thermoproteaceae	0	0.01198	0.0245	0
Thermoplasmataceae	0	0	0.00108	0
Thermofilaceae	0	0	0.00288	0
Terrestrial Hot Spring Gp (THSCG)	0	0	0.00036	0
Syntrophoarchaeaceae	0	0	0.00108	0
Parcubacteria group bacterium GW2011_GWA2_37_10	0	0	0.00901	0
Nitrosotaleaceae	0	0.00177	0.00144	0
Nanosaliniaceae	0	0	0.0036	0
Nanoarchaeota archaeon JGI OTU-1	0	0	0.02738	0
Methanothermobacteriaceae	0	0	0.00072	0
Methanospirillaceae	0	0.00044	0.00252	0
Methanosarcinaceae	0	0.05544	0.01009	0
Methanosetaeaceae	0	0	0.01585	0
Methanoregulaceae	0	0.00044	0.00252	0
Methanopyraceae	0	0	0.00396	0
Methanoperedenaceae	0	0	0.00216	0
Methanonatronarchaeaceae	0	0	0.00072	0
Methanomicrobiaceae	0	0.00355	0.01261	0
Methanomassiliococcaceae	0	0.00133	0.00036	0
Methanofastidiosaceae	0	0	0.0036	0
Methanococcaceae	0	0	0.00252	0
Methanocellaceae	0	0	0.00036	0
Methanocaldococcaceae	0	0.00222	0.00324	0
metagenome	0	0	0.34188	0
marine metagenome	0	0	0.8114	0
Marine Group III euryarchaeote SCGC AAA288-E19	0	0	0.0018	0
Marine group II euryarchaeote REDSEA-S19_B7N8	0	0	0.00108	0
Korarchaeaceae	0	0.00044	0.00108	0
Ignicoccaceae	0	0.00044	0.00036	0
hydrothermal vent metagenome	0	0	0.00036	0
Halobacteriaceae	0	0.00532	0.00072	0
Haloadaptaceae	0	0	0.00072	0
Geothermarchaeaceae	0	0	0.00036	0
Feravidicoccaceae	0	0	0.00036	0
Ferroplasmaceae	0	0.00089	0.00072	0
Euryarchaeota archaeon SGB-5	0	0	0.0036	0
Desulfurococcaceae	0	0.00222	0.00144	0
Desulfobacteraceae bacterium 4572_19	0	0	0.00396	0
crenarchaeote SRI-298	0	0	0.00072	0
Candidatus Yanofskybacteria bacterium RIFCSPHIGHO2_01_FULL_41_26	0	0	0.01549	0
Candidatus Pacearchaeota archaeon CG1_02_31_27	0	0	0.03279	0
Candidatus Nomurabacteria bacterium RIFCSPLOWO2_02_FULL_42_17	0	0	0.00468	0
Candidatus Micrarchaeum	0	0	0.00036	0
Candidatus Hydrothermarchaeota archaeon JdFR-18	0	0	0.00324	0
Candidatus Amesbacteria bacterium GW2011_GWC1_47_15	0	0	0.00144	0
Candidatus Aenigmarchaeota archaeon JGI 0000T06-F11	0	0	0.00108	0
Caldiarchaeaceae	0	0	0.00685	0
archaeon GW2011_AR16	0	0	0.01261	0
Archaeoglobaceae	0	0.00089	0.0036	0
ANME-2a-2b	0	0	0.01081	0
Aciduliprofundaceae	0	0.00089	0.00252	0
SCGC AAA286-E23	0	0.02883	0.03747	0.00207
SCGC AAA011-D5	0	2.796	3.02453	8.04
GW2011_GWC1_47_15	0	0.3824	0.70224	1.439
Thermococcaceae	0.00036	0.01508	0.00216	0
Nitrososphaeraceae	0.00266	0.00266	0.0018	0
Methanobacteriaceae	3e-05	0.01597	0.01081	0
Halomicrobiaceae	3e-05	0.0102	0.00757	0
Haloferacaceae	5e-05	0.0794	0.00613	0
Halococcaceae	0.00024	0.00044	0.00036	0
Candidatus Aenigmarchaeum	0.00078	0.00089	0.00108	0

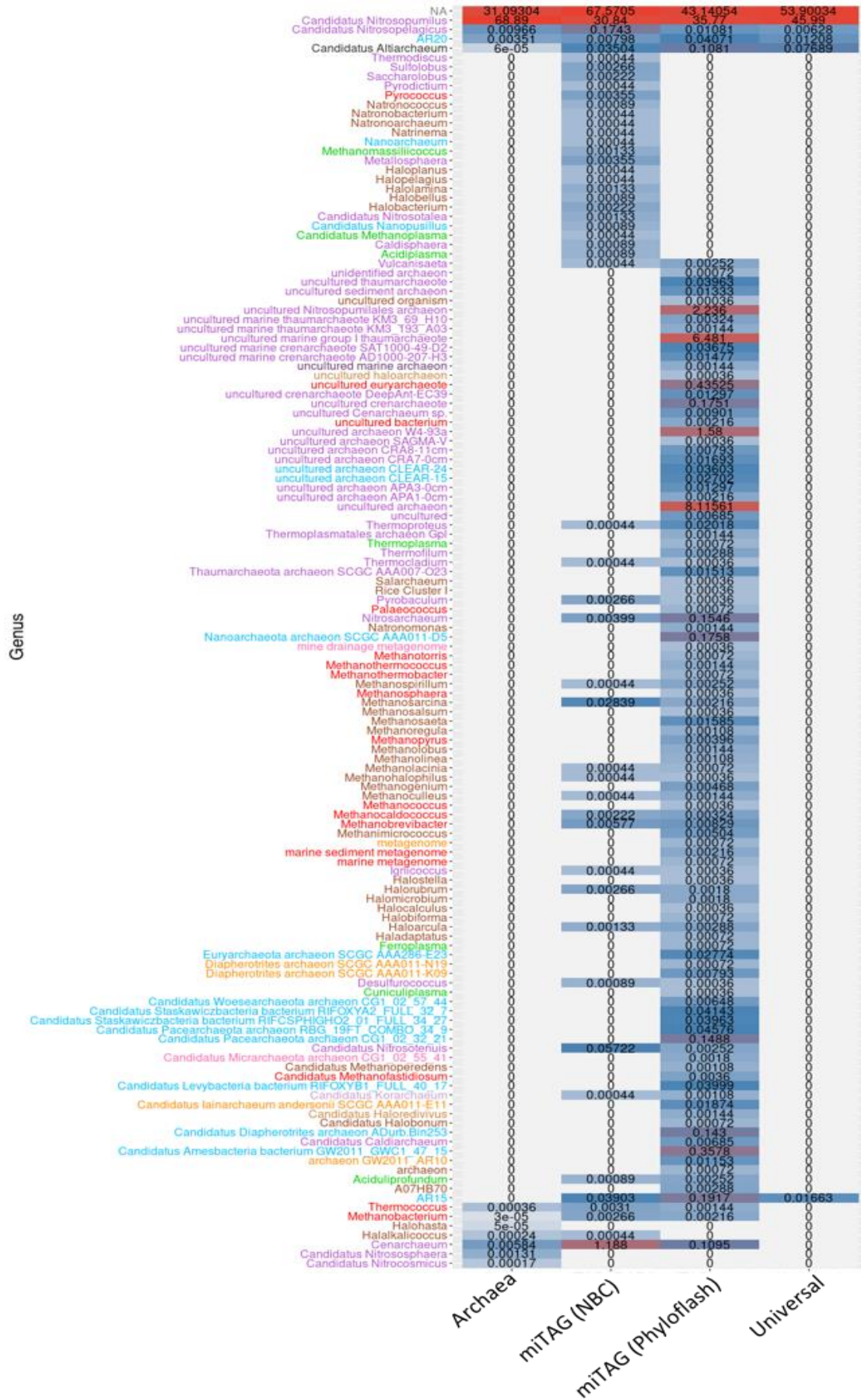
Archaea
miTAG (NBC)
miTAG (Phyloflash)
Universal

CHAPTER 1

Figure S5: Heatmap of the relative proportion of each taxonomic family in the four datasets. The names of the families are colored by phylum, see Fig. S3 for legend.

Figure S6: Heatmap of the relative proportion of each taxonomic genus in the four datasets. The names of the genera are colored by phylum, see Fig. S3 for legend.

CHAPTER 1



CHAPTER 2

Diversity and biogeography of
bathyal and abyssal seafloor
Bacteria and Archaea along a
Mediterranean - Atlantic gradient

Diversity and biogeography of bathyal and abyssal seafloor

Bacteria and Archaea along a Mediterranean – Atlantic gradient

Blandine Trouche^{*1,a}, Miriam Brandt^{2,b}, Caroline Belser^{3,c}, Covadonga Orejas^{4,d}, Stéphane Pesant^{5,e}, Julie Poulain^{3,f}, Patrick Wincker^{3,g}, Jean-Christophe Auguet^{6,h}, Sophie Arnaud-Haond^{†2,i} and Loïs Maignien^{†1,7,j}

† These authors have contributed equally to this work and share last authorship

¹ Univ Brest, CNRS, IFREMER, Microbiology of Extreme Environments Laboratory (LM2E), Plouzané, France

² MARBEC, Univ Montpellier, Ifremer, IRD, CNRS, Sète, France

³ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Évry, Université Paris-Saclay, 91057 Evry, France

⁴ Centro Oceanográfico de Baleares, Instituto Español de Oceanografía, Muelle de Poniente, S/N, 07015, Palma de Mallorca, Spain

⁵ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

⁶ MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France

⁷ Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Woods Hole, MA, United States

Article accepted in Frontiers in Microbiology.

Résumé de l'article en français

Les sédiments marins recouvrent la majorité de la surface de la planète Terre, et les micro-organismes qui les habitent jouent un rôle central dans les cycles biogéochimiques marins. Pourtant, la description de la distribution et de la biogéographie de la diversité microbienne sédimentaire est encore trop limitée pour permettre d'évaluer la contribution des processus donnant lieu à cette distribution, comme par exemple l'influence respective de la dérive, de la connectivité et de la spécialisation. Pour apporter des éléments de réponse à cette question nous avons analysé 210 librairies de métabarcoding visant les Bactéries et les Archées, générées à partir d'une collection d'échantillons standardisés et découpés en horizons provenant de 18 stations organisées selon un gradient longitudinal entre l'Est de la Méditerranée et l'Ouest de l'Atlantique. Dans l'ensemble, nous avons observé une différence dans les schémas biogéographiques suivant l'échelle spatiale considérée : à l'échelle locale, l'influence sélective de l'environnement contemporain semble la plus forte, tandis que l'héritage des processus historiques via la limitation de dispersion et la dérive devient plus visible à l'échelle régionale, jusqu'à l'emporter sur les influences contemporaines à l'échelle inter-régionale. En ce qui concerne les facteurs environnementaux, la structure des communautés est principalement liée à la profondeur de l'eau, avec une transition claire entre 800 et 1200 mètres sous la surface. Le bassin océanique, la température de l'eau et la profondeur dans le sédiment sont d'autres facteurs explicatifs importants de la structure des communautés microbiennes. Enfin, nous suggérons que l'augmentation de la limitation de dispersion et de la dérive écologique avec la profondeur dans le sédiment pourrait être un des facteurs résultant dans la divergence accrue observée pour les communautés des horizons profonds.

Abstract

Seafloor sediments cover the majority of planet Earth and microorganisms inhabiting these environments play a central role in marine biogeochemical cycles. Yet, description of the biogeography and distribution of sedimentary microbial life is still too sparse to evaluate the relative contribution of processes driving this distribution, such as the levels of drift, connectivity, and specialization. To address this question, we analyzed 210 archaeal and bacterial metabarcoding libraries from a standardized and horizon-resolved collection of sediment samples from 18 stations along a longitudinal gradient from the eastern Mediterranean to the western Atlantic. Overall, we found that biogeographic patterns depended on the scale considered: while at local scale the selective influence of contemporary environmental conditions appeared strongest, the heritage of historic processes through dispersal limitation and drift became more apparent at regional scale, and ended up superseding contemporary influences at inter-regional scale. When looking at environmental factors, the structure of microbial communities was correlated primarily with water depth, with a clear transition between 800 and 1200 meters below sea level. Oceanic basin, water temperature, and sediment depth were other important explanatory parameters of community structure. Finally, we propose increasing dispersal limitation and ecological drift with sediment depth as a probable factor for the enhanced divergence of deeper horizons communities.

Introduction

Marine sediments cover around 65% of the Earth's surface and accumulate particulate organic matter settling from the water column, thereby representing the largest sink of oceanic organic matter (Jørgensen and Boetius, 2007; Seiter et al., 2005). Bacteria and archaea in these sediments represent the largest pool of biomass in the deep sea, with their abundance estimated to be on the order of 4.9×10^{28} cells in the benthic layer (top 50 cm) and 2.9×10^{29} globally (Kallmeyer et al., 2012; Danovaro et al., 2015). Contrary to meio-, macro-, and mega-fauna, their abundance and biomass does not decrease with water depth, though cell counts decrease logarithmically with depth in the sediments. Benthic bacteria and archaea are essential for the early diagenesis of sinking organic matter and as a consequence, they are crucial contributors to biogeochemical cycles, determining the partitioning between buried organic matter and nutrients released in the water column (Orcutt et al., 2011; Teske et al., 2011). This underlines the importance of the benthic boundary layer microbial communities as a transition between water-column and seafloor communities (Zinger et al., 2011; Walsh et al., 2016).

Thanks to recent technological advances, particularly in sequencing techniques (e.g. Huber et al., 2006, reviewed in Salazar and Sunagawa, 2017), it is now possible to perform near-exhaustive inventories of benthic microbial community diversity across large spatial scales, and to investigate patterns of microbial distribution. Despite their essential role in the marine ecosystem (Nealson, 1997; del Giorgio and Duarte, 2002; Jørgensen and Boetius, 2007; Aristegui et al., 2009; Molari et al., 2013), processes shaping benthic prokaryotic community structure are still poorly understood, and the existence of biogeographic patterns has been questioned owing to their possible unlimited dispersal ability (Green and Bohannan, 2006; Astorga et al., 2012). Nonetheless, recent studies focusing on deep sea benthic microorganisms at local and regional scale (Jacob et al., 2013; Buttigieg and Ramette, 2015; Liu et al., 2020; Li et al., 2021) and meta-analyses (Bienhold et al., 2016; Petro et al., 2017;

CHAPTER 2

Hoshino et al., 2020) have clearly shown geographic structuration in these communities, even at reduced spatial scales.

Biogeographic patterns are usually considered to result from four main evolutionary forces: selection, diversification, dispersal, and drift (Vellend, 2010; Hanson et al., 2012; Nemergut et al., 2013). These processes are often split between deterministic and stochastic, selection being considered wholly deterministic, drift being stochastic, and dispersal and diversification largely accepted as stochastic processes, although they may encompass both deterministic and stochastic components (Zhou and Ning, 2017). One of the most studied biogeographic patterns resulting from these processes is the evolution of community composition with geographic distance. When community similarity decreases with increasing geographic distance, a distance-decay relationship or “isolation by distance” pattern will be observed (Nekola and White, 1999; Horner-Devine et al., 2004; Soininen et al., 2007; Hanson et al., 2012). Coupled with investigation of the link between community and environmental similarity, this approach provides insights into the relative contribution of historical and contemporary processes shaping microbial provinces and habitats, as proposed by Martiny et al. (2006).

Besides, microbial communities display a strong stratification with sediment depth that has traditionally been explained by the redox gradient with depth of electron acceptors that are sequentially consumed by organic matter respiring microorganisms (Emerson et al., 1980; Durbin and Teske, 2011; Orcutt et al., 2011). In addition to the deterministic influence of environmental conditions, recent studies focusing on processes involved in vertical distribution of sedimentary microorganisms have suggested a strong influence of surface community structure on the seafloor community assembly through selective survival, beginning in the very first layers of sediment (Jochum et al., 2017; Petro et al., 2017, 2019; Starnawski et al., 2017; Kirkpatrick et al., 2019; Marshall et al., 2019).

In this study, we aimed at examining benthic microbial community diversity and biogeographic patterns across the Mediterranean - Atlantic basins to determine to what extent the microbial community structure resulted from past historical processes *versus*

CHAPTER 2

contemporary environmental drivers at different spatial scales. Building on previous work suggesting that assembly of seafloor microbial communities initiates in the very first layers of sediment, we also examined the evolution of microbial community structure with increasing depth in the surface sediments of the seafloor.

Material & Methods

1. Sample collection and processing

1.1. Cruises and locations

Samples from 18 stations from the eastern Mediterranean Sea to the northern Atlantic Ocean were collected between April 2016 and May 2017 (Fig. 11A). In the spring of 2016, samples were taken from the upper and lower bathyal zones of the Gulf of Lion during cruises ESSNAUT16 (DOI: 10.17600/16000500) and CanHROV (DOI: 10.17600/16012300). In September 2016, the MEDWAVES cruise (Atlas project H2020) targeted one Mediterranean feature (Seco de los Olivos gullot), and three Atlantic features (Gazul mud volcano in the Gulf of Cádiz, Ormonde seamount off Portugal and Formigas seamount off Azores) (Orejas et al., 2017). In March 2017, samples were collected from the abyssal plains of the North Atlantic Ocean during transect cruise AMIGO1. Finally, in May 2017, the sampling for this study was completed during the PEACETIME cruise (DOI 10.17600/17000300), targeting the lower bathyal zone of the western Mediterranean Sea. Details of the stations are given in Table 3.

CHAPTER 2

Table 3: List of sampling stations and their characteristics. Mbsl = meters below sea-level.

Geographic zone	Location	Latitude	Longitude	Depth (mbsl)	Station name	Sampling cruise
Mediterranean Sea	Mediterranean (abyssal plain) - Ionian Sea	35.4891	10.776	3100	ION	PEACETIME
	Mediterranean (abyssal plain) - Tyrrhenian Sea	39.3402	12.5927	3400	TYRR	
	Mediterranean (undersea canyon)	42.7167	6.1333	2490	Canhrov_ST 1	CanHROV
	Mediterranean (passive margin) – Gulf of Lion	42.9422	6.7422	2417	ESN1	ESSNAUT16
		43.0867	6.4512	334	ESN2	
Mediterranean (abyssal plain)	37.9467	2.9167	2800	FAST	PEACETIME	
Transition	Alborán Sea (Seco de los Olivos gullot)	36.4808	-2.8945	729	ST179	MEDWAVES
		36.5460	-2.8135	381	ST201	
		36.5157	-2.7942	554	ST215	
	Gulf of Cádiz (Gazul mud volcano)	36.5598	-6.9492	470	ST22	
		36.5605	-6.9498	470	ST23	
	Southwest Portugal (Ormonde seamount)	36.8442	-11.3025	1920	ST38	
Atlantic Ocean	Azores (Formigas seamount)	37.34	-24.7552	1325	ST117	
		37.2837	-24.7873	1245	ST68	
	North Atlantic (abyssal plain)	26.89	-22.4442	4931	Amigo1_ST0	AMIGO1
		23.0042	-41.2092	4770	Amigo1_ST1	
		20.3392	-49.4359	4630	Amigo1_ST2	
		18.8175	-54.0836	4630	Amigo1_ST3	

1.2. Sampling protocol

For each station, three cores were collected with a multicorer (MUC) or push-cores deployed from the Nautilie submarine (ESSNAUT16) or a remotely operated vehicle (ROV, CanHROV). The sediment cores were sliced onboard in a lab environment previously cleaned using ~10% bleach solution, rinsed with ethanol and ultrapure water. Each core was sliced into depth layers following a standard scheme: 0-1 cm, 1-3 cm, 3-5 cm, 5-10 cm, 10-15 cm, and when

CHAPTER 2

the cores were long enough 15-30 cm or to 1 cm before the maximum length, to avoid contamination from the core extruder. Slicing was performed using spatulas also bleached and rinsed with ultrapure water before each use. Horizons (slices of sediment) were transferred into zip-lock bags, homogenized, and frozen at -80°C on board before being shipped on dry ice to the laboratory where they were also kept at -80°C .

2. DNA extraction

DNA extractions were performed in a sterile lab, using approximately 10 g of sediment with the PowerMax Soil DNA Isolation Kit according to the manufacturer's instructions (Qiagen, Hilden, Germany) with modifications: the elution buffer was left on the spin filter membrane for 10 min at room temperature before centrifugation in order to increase DNA yield. Extraction controls were performed by using an empty tube from the kit for each series of extraction or extraction kit batch. In total, 8 extraction blanks were produced. When field controls were prepared onboard (empty zip-lock bags), the first solution of the kit was poured into the control ziplock bag, before following the usual extraction steps. Each of the resulting 5 mL DNA solutions were stored at -80°C .

3. Libraries construction and sequencing

A primer pair targeting both Bacteria and Archaea (Parada et al., 2015) was used to amplify the V4V5 region of the 16S rRNA gene (515F: 5'-GTGYCAGCMGCCGCGGTAA, 926R: 5'-CCGYCAATTYMTTTRAGTTT). PCR amplifications were carried out at Genoscope (Evry, France) as part of the eDNAbyss project (see Supporting Information for amplification, purification, and quantification details). Amplicon libraries were prepared for each sample by non-directional ligation of Illumina adapters on 100 ng of amplicons following the Kapa Hifi HotStart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA, USA). After

CHAPTER 2

quantification and quality control, libraries normalized to 8–9 pM concentrations and containing a 20% PhiX spike-in were sequenced on HiSeq2500 instruments in a 250 bp paired-end mode (System User Guide Part # 15035786).

4. Bioinformatic analysis

All bioinformatic analyses were performed using a standardized pipeline (Brandt et al., 2021), available on Gitlab (<https://gitlab.ifremer.fr/abyss-project/abyss-pipeline>), on a home-based cluster (DATARMOR, Ifremer).

First, sequence files were renamed from their Genoscope identifiers to more explicit names. Due to non-directional adapter ligation, inserts were sequenced in different orientations. We thus used Cutadapt v1.9 (Martin, 2011) to identify the primer sequence in each read and sort them according to two criteria: forward or reverse primer and forward or reverse sequencing. Data for each sample was thus split into 4 sequence files (R1F, R1R, R2F, R2R). Cutadapt then removed the identified primer sequences and BBMAP repair (Bushnell, 2014) was used to ensure that reads were still paired by sorting reads using the information present in their description line and removing unmatched reads.

For each sequencing run, we determined Amplicon Sequence Variants, merged read pairs and removed chimeras using the DADA2 package v.1.10 (Callahan et al., 2016), following guidelines from the online tutorial for paired-end HiSeq data (<https://benjjneb.github.io/dada2/bigdata.html>). The script implementing DADA2 was applied separately to the two pairs of sequence files R1F/R2R and R2F/R1R. The parameters used for filtering and trimming reads were as follows: truncation length of 220 base pairs, maxN= 0, maxEE= 2 and truncQ= 11. The error learning step was based on nbases= 1e8. Merged sequences were size-filtered by keeping sequences with a length between 350 and 390 bps. The Amplicon Sequence Variants (ASVs) tables produced by DADA2 for each run were then merged, collapsing ASVs based on DNA sequence identity. Taxonomic assignment was

CHAPTER 2

performed with the implementation of the RDP naive Bayesian classifier (Wang et al., 2007) available in DADA2 v.1.10, using the Silva v138 reference database (Quast et al., 2013) and a bootstrap threshold of 80.

The ASV and taxonomy tables produced by this pipeline were then combined in a phyloseq object (phyloseq v1.28.0, McMurdie and Holmes, 2013) in an R v3.6.1 environment. Reads from the same amplicon library, but originating from different Illumina runs, were merged under the same sample name before removing sequences from unwanted taxa (Eukaryota, Chloroplast and Mitochondria affiliated sequences). Data was decontaminated using extraction, PCR and field controls using the decontam package (v1.4.0, Davis et al., 2018), or handpicking in the case of the ASV dominating control libraries reads (see reproducible workflow on github). Samples totaling less than 40,000 reads after decontamination were removed, the appropriate metadata added, and the final object saved as a phyloseq object for further analysis in R.

Scripts for the reproducible bioinformatic workflow are available at https://github.com/loimai/ABYSS_16S/tree/master/docs.

5. Sediments characterization

Characterization of the sediment samples was carried out by Filab S.A.S (Dijon, France). Granulometry values were obtained using wet Malvern laser scattering together with humidity level and loss on ignition at 550°C (see Supporting Information for more details on the methods used).

Temperature for each sampling station was extrapolated when possible (MEDWAVES expedition) from CTD data from the same sampling stations (Orejas et al., 2017). When this data did not exist, it was set to average temperature recorded in the ocean basin at the depth considered (Mantyla and Reid, 1983; Pierre, 1999; Martín-Cuadrado et al., 2007).

CHAPTER 2

6. Statistical analysis

All subsequent statistical analyses were done in R v3.6.1, using phyloseq (v1.28.0, McMurdie and Holmes, 2013), vegan (v2.5.7, Oksanen et al., 2015) and ggplot2 (v3.3.0, Wickham, 2016) packages to compute alpha diversity, beta diversity, and produce taxonomy barplots.

A map of the sampling stations was generated using ggmap (v3.0.0, Kahle and Wickham, 2013). Description of the sampling sites was based on available metadata using the principal components analysis (PCA) from package FactomineR (Lê et al., 2008). Environmental parameters considered for each sample were ocean depth at sampling station, distance from shore, sediment horizon (depth in the sediment core), temperature above seafloor, and sediment characteristics, namely mean organic matter content by station and horizon, mean humidity level by station and horizon, mean granulometry (μm), and heterogeneity of particle size. Visualization of the distance-decay relationships relied on community similarity computed using a Bray-Curtis index after normalization of the dataset using cumulative sum scaling with the metagenomeSeq package (Paulson et al., 2013). Geographic distance was measured using a 'Vincenty' (ellipsoid) great circle distance to take into account Earth curvature, relying on packages enmSdm, and geosphere (v1.5.10, Hijmans, 2019). Environmental similarity between samples was estimated with euclidean distances on the centered and scaled environmental data table, using the R base *scale* function. The statistical difference between the slopes of the models fitted to the distance-decay relationships was tested using function *diffslope2* from package simba (Jurasinsk and Retzer, 2012).

Beta-diversity variation partitioning analysis were computed using function *varpart* in the vegan package. Ordinations of the microbial data were visualized as nMDS using phyloseq, and environmental data were fitted to the ordinations using the *envfit* function in package vegan. Permutational multivariate analyses of variance were performed when appropriate with the *adonis* function of package vegan, after checking the homogeneity of group dispersions using function *betadisper*. Finally, biomarker detection was done with the DESeq2 package (1.24.0, Love et al., 2014).

CHAPTER 2

The fully reproducible workflow for statistical analysis is available at:

https://github.com/loimai/ABYSS_16S/tree/master/docs.

Results

1. 16s rRNA gene amplicon processing

A total of 230 sample libraries were built and sequenced, producing 195,470,177 raw sequences of the 16S rRNA gene's V4-V5 region, with a mean of 849,870 reads by library. A total of 17,366,268 reads were recovered from the additional 24 control libraries that were constructed and sequenced simultaneously, originating from sampling (empty storage bags conditioned on-board research vessels at the end of some of the sampling sessions), extraction (empty kit processed through all extraction steps together with the samples) and PCR (ultrapure water) blanks.

After processing with DADA2 the dataset included 123,454,421 sequences for a total of 265,198 ASVs. Of those, 1,223 (about 0.5%) were found in control samples, and 728 ASVs were specific to the control libraries. Most of the contamination was dominated by a specific ASV that accounted for 99% of reads in negative control libraries. This ASV, affiliated with partial 16S sequences of *Sphingobium* strains, is a recognized contaminant of Taq-Phusion reagents (Salter et al., 2014).

After bioinformatic processing, taxonomic refining and decontamination, the dataset comprised a total of 210 libraries including 66,826,975 sequences representing 260,567 ASVs (min 40,076 sequences, max 847,227 sequences). Rarefaction curves (Fig. S7) confirmed that sequencing and sampling efforts captured most of the sample diversity.

2. Description of sampling sites

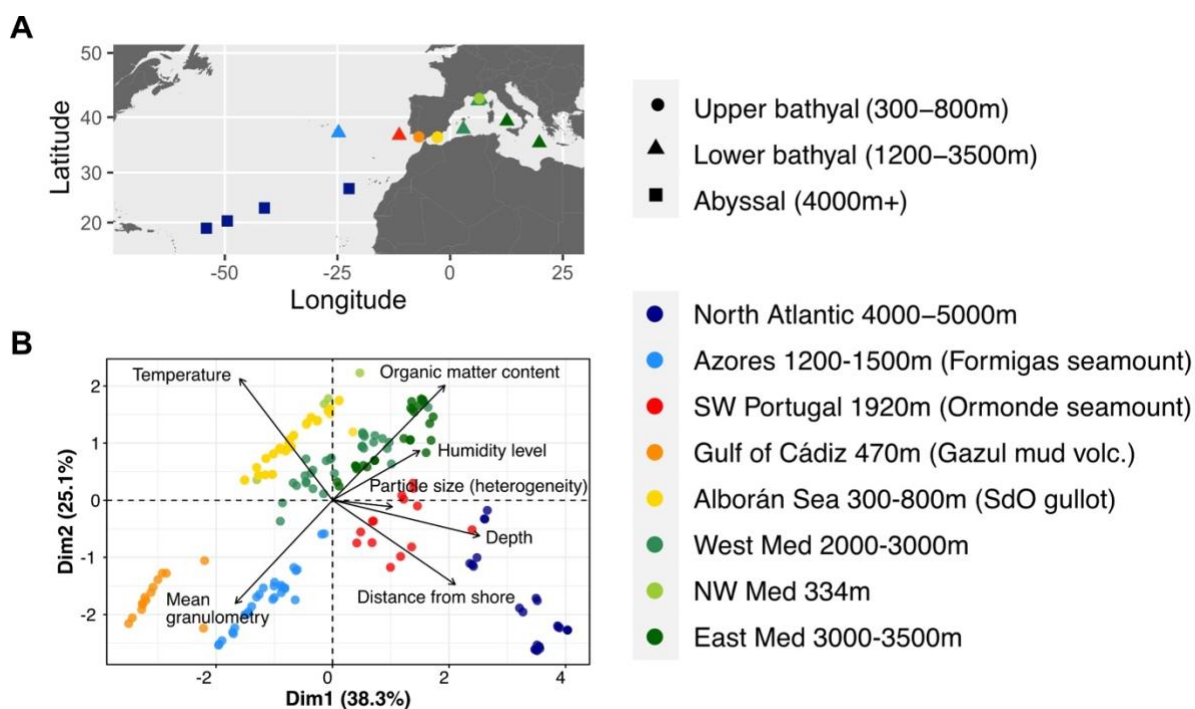


Figure 11: Description of sampling sites: **(A)** Map of the sampling stations across the Mediterranean and Atlantic transition. **(B)** Characterization of the samples based solely on available metadata using a principal components analysis biplot. Arrows represent the decomposition of the variables

Three geographic zones were defined based on the coordinates of the sampling stations (Table 3 and Fig. 11A). From East to West, the Mediterranean zone grouped the stations from the Ionian and Tyrrhenian Sea, the Gulf of Lion and the abyssal plain near the Balearic Islands. The Transition zone around the Gibraltar Strait consisted of the stations from the Alborán Sea, Gulf of Cádiz and southwest Portugal. Finally, the Atlantic zone was composed of the Azores and north Atlantic abyssal plain stations.

We first characterized the samples based on the available environmental parameters (depth, distance from shore, temperature, organic matter (OM) content, humidity level, mean granulometry (μm) and heterogeneity of particle sizes) (Fig. 11B). The first two dimensions of

CHAPTER 2

the Principal Components Analysis summed up 63.4% of the total inertia. Five variables contributed most to these dimensions, namely depth and distance from shore, temperature, OM content and granulometry, leading to a segregation of samples by site rather than oceanic basin. Depth and distance from shore were anti-correlated with temperature, thus creating a gradient of sampling sites from the shallow warm sediments of the Mediterranean Sea to the deep abyssal samples of the Atlantic Ocean. Two groups of sites, from the Azores and Gulf of Cádiz, differed most from the others based on the sediment composition data.

3. Distance-decay relationship between deep sea sediment communities

To explore biogeographic patterns along the longitudinal gradient, we plotted the community similarity between pairs of samples as a function of their geographic distance and their environmental similarity (Fig. 12). Regarding geographic distances, we only compared samples originating from the same sediment layer (horizon) and partitioned the pairwise comparisons according to sampling region (Mediterranean, Atlantic and Transition region) to investigate biogeographic patterns at regional (Fig. 12A) and inter-regional scales (Fig. 12B). Community similarity between sample pairs generally decreased with geographic distance, hence exhibiting a clear distance-decay relationship (DDR) at all scales, both within (Fig. 12A) and between geographic zones (Fig. 12B). Linear regressions of the DDRs showed that the highest rate of decrease in community similarity with distance occurred in the transition zone (slope: -0.00342), followed by the Mediterranean basin (slope: -0.00134), and the Atlantic basin (slope: -0.000715). At inter-regional scale, the slope of the DDR was steepest between the Atlantic and Transition regions and the Mediterranean and Transition regions (Fig. 12B). It was 0.5 times less steep in the Atlantic-Mediterranean comparison, likely due to the most distant samples originating from the deepest stations in the dataset (water depth).

CHAPTER 2

In addition, we observed a generally positive correlation between community similarity and environmental similarity, with a modelled regression slope at least four times higher at intra-regional scale (Fig. 12A, slope: 0.0703 to 0.0835) than at inter-regional scale (Fig. 12B, slope: 0.00601 to 0.0213).

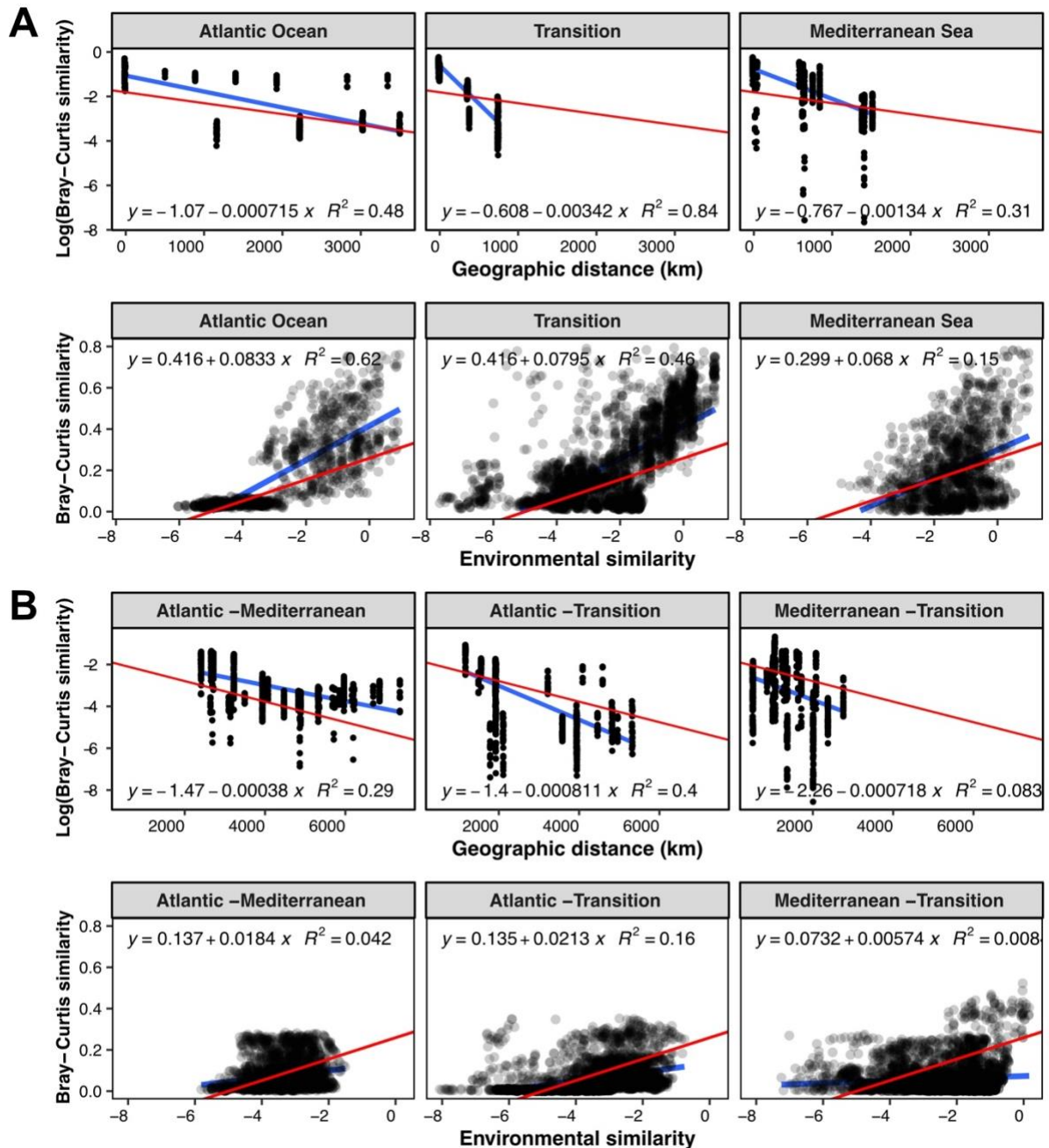


Figure 12: Pairwise Bray-Curtis community similarity between samples with respect to geographic distance (km) and environmental similarity: **(A)** samples from the same geographic zone, **(B)** samples from two different zones. For the evolution with distance, pairwise community similarity was evaluated exclusively between samples of the same horizon.

CHAPTER 2

Blue lines illustrate linear models computed for the subset of samples considered, and red lines represent the overall linear regression when including all the samples. All linear models have a p-value at least inferior to $3.306e^{-10}$.

Overall linear regression Log(Bray-Curtis) vs geographic distance: $y = -1.81 - 0.000491x$, $R^2 = 0.27$, p-value $< 2.2e^{-16}$

Overall linear regression Bray-Curtis vs environmental similarity: $y = 0.261 + 0.0524x$, $R^2 = 0.27$, p-value $< 2.2e^{-16}$

4. Distance-decay relationship depending on sediment horizons

When decomposed for each sediment layer, we observed a clear increase in DDR with sediment horizon, with linear regression slopes approximately 5 times steeper in horizon 15-30 cm compared to the top three horizons (Fig. 13, Table S1). Indeed, slopes ranged from 0.000445 in the first horizon (0-1 cm) to 0.00255 in horizon 15-30 cm (Table S1), and were significantly different between adjacent horizons, except for horizons 1-3 cm and 3-5 cm. The fit of the regression did not clearly improve in deeper horizons, and ranged from 0.17 to 0.4, except for the lowest horizon where it reached 0.97 but was calculated on the lowest number of samples ($n = 12$).

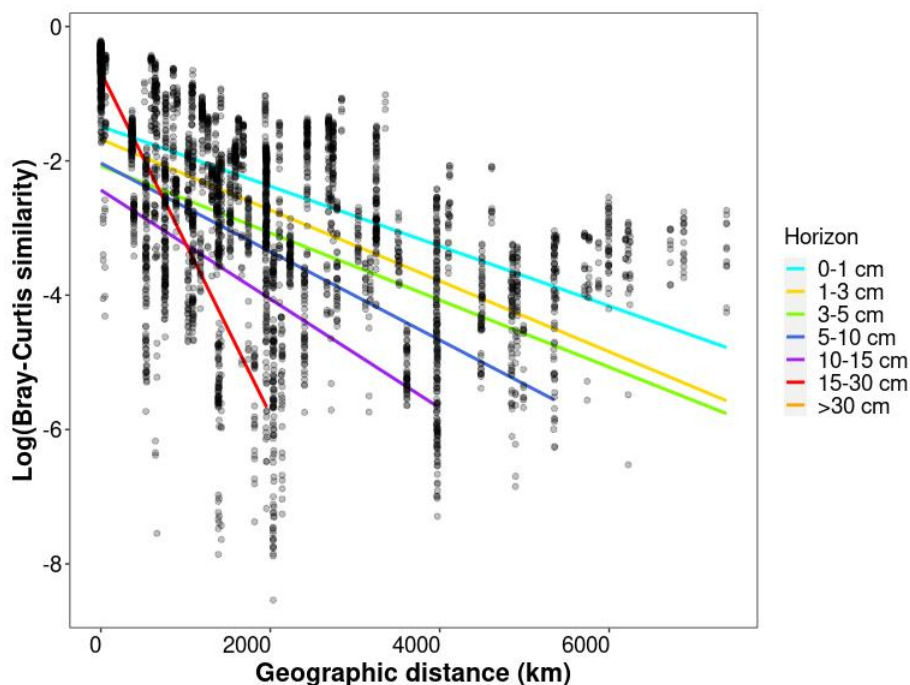


Figure 13: Pairwise Bray-Curtis community similarity with respect to geographic distance (km) between samples. Each point corresponds to a pairwise comparison between samples of the same sediment horizon, and a linear regression has been computed separately for each horizon (equations in Table S1).

5. Environmental parameters structuring microbial communities

In terms of alpha-diversity, Shannon indices were comparable across locations, except in the north Atlantic and eastern Mediterranean abyssal plains (Fig. S8B), where Shannon diversity was significantly lower (Kruskal-Wallis rank sum test, $p < 2.2e^{-16}$). These seemingly low diversity samples were the ones most affected by the Taq-Phusion contamination, most probably due to the low DNA content in the sediment and in the extract, resulting in poor sequencing depths. Alpha-diversity estimates decreased with increasing horizon depth for all locations, except for the sites located above 800 m ocean depth (Alborán Sea, Gulf of Cádiz, and NW Med) (Fig. S8B).

Overall, benthic microbial communities considered in this study exhibited similar dominant phyla, regardless of the origin of the samples in terms of geography, water depth, or horizon depth. Members of the Acidobacteria, Crenarchaeota (mostly Nitrososphaeria, previously a

CHAPTER 2

member of phylum Thaumarchaeota), Planctomycetota, and Proteobacteria (shared between Alpha- and Gammaproteobacteria) were predominant in all samples (Fig. S8A).

When investigating the correlation between microbial community composition and environmental parameters, we found that samples readily clustered by site on the ordination (Fig. 14A, PERMANOVA test $F= 10.772$, $p= 0.001$), as expected from the PCA results (Fig. 11B). We also noticed a clear split between samples originating from shallow (< 800 meters below sea-level (mbsl)) and deep (> 1200 mbsl) stations even when originating from the same oceanic basin (i.e. the Mediterranean). This depth effect was visible on ordination plots (Fig. 14A and B), with a significant fit of the environmental data ($p= 0.005$), and on alpha diversity profiles (Fig. S8B). This difference was backed up by a variation partitioning analysis (Fig. S10). Focusing on samples deeper than 1200 mbsl, the gradual change in community structure with water depth was maintained, with an increased turnover with changes in temperature and oceanic basin (Fig. 14A, B and C).

Few differences in taxonomic composition were observed at phylum level between samples above and below 1000 mbsl (Fig. S8A), except for an increasing importance of Acidobacteria with water depth, while Bathyarchaeia and Desulfobacterota (previously Deltaproteobacteria) were present almost exclusively above 1000 mbsl.

Stratification of community composition with sediment horizon was observed in each depth range (Fig. 14D) and was clearest in terms of taxonomy for the upper bathyal zone, with the following trends: increase in the presence of Acidobacteria, Chloroflexi, and Nanoarchaeota with increasing depth in the sediment cores, while the relative abundance of Alpha- and Gammaproteobacteria decreased (Fig. S8A). This horizon effect was mostly apparent inside each site (PERMANOVA $F= 5.11$, $p= 0.001$).

Contrary to the results from the environmental PCA (Fig. 11B), sediment composition was not strongly linked with community composition. Indeed, humidity and heterogeneity of particle size were not significantly correlated with the ordination axes. Granulometry, and the anti-

CHAPTER 2

correlated OM content, were significantly ($p = 0.01$) correlated with axis 2 of the ordination though more weakly than temperature or depth. These results were reflected in the variation partitioning analyses (Fig. S10C).

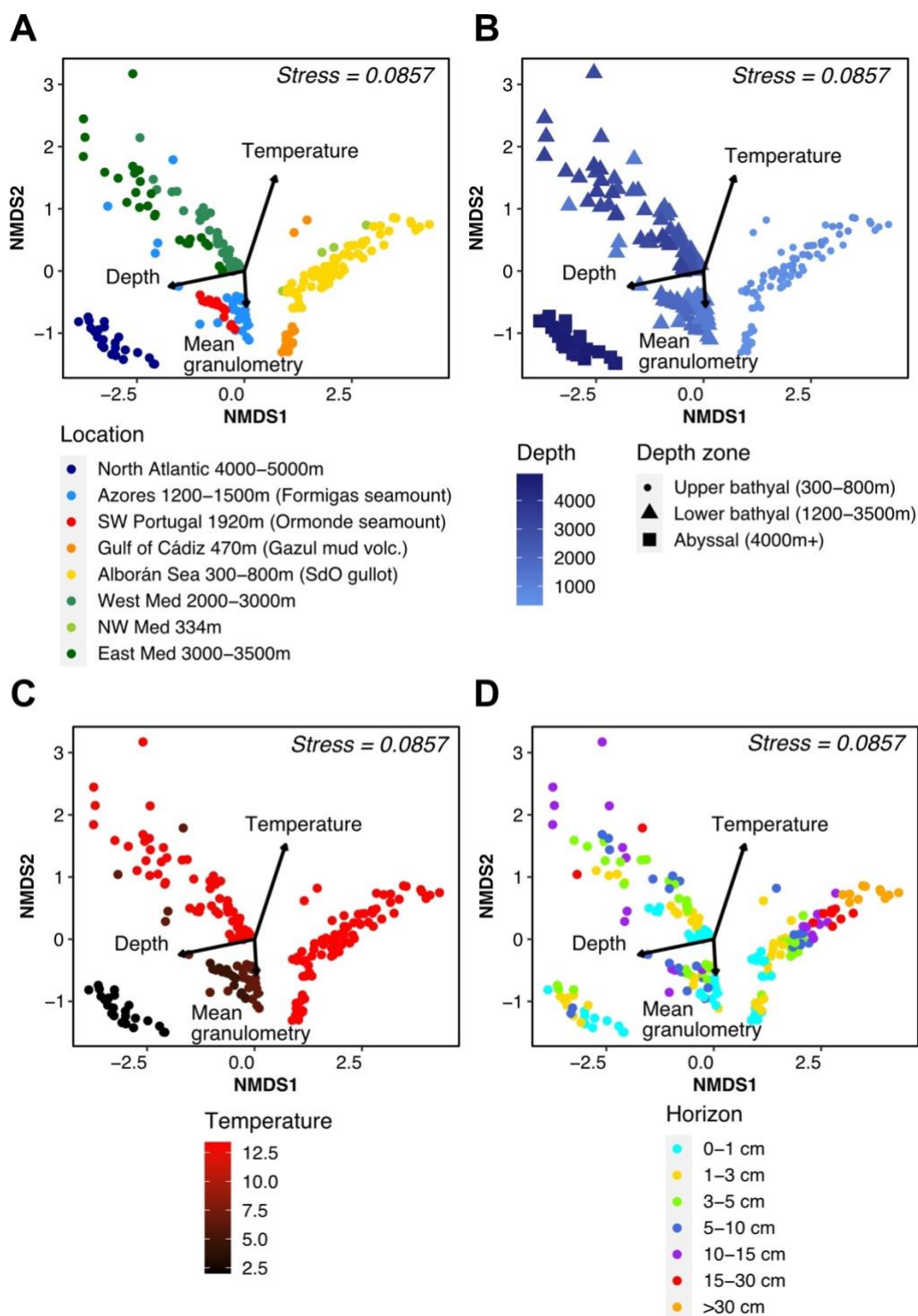


Figure 14: Non-Metric Multidimensional Scaling ordination plot of Bray-Curtis distance between samples, colored according to **(A)** geographic region, **(B)** water depth at sampling station, **(C)** temperature of aboveground water and **(D)** horizon depth in the sediment core (cm). Environmental variables were fitted to the unconstrained ordination and the significant ones were added to the plot with their length illustrating the strength of the correlation.

6. Exploring the link between surface and subsurface communities at local scale

We chose a subset of three sites located in the Alborán Sea (Seco de los Olivos gullot), each sampled with triplicate cores down to 45 cm, to examine community changes at local scale. The sites were sampled at three water depths (381 m, 554 m, 729 m, see Table 3), with a maximum distance of 11 km between two sites.

We observed a distance-decay relationship for Alborán Sea samples (Fig. 15A). However, in contrast with (inter)regional scales, neither the slope of the DDR nor its fit increased with depth in the sediment at local scale (Table S2). A significant correlation between Bray-Curtis and environmental similarities was apparent as well, with a fit ($R^2 = 0.53$) (Fig. 15B) in the same range as what was observed at the regional scale ($R^2 = 0.16-0.62$) (Fig. 12A).

Variation partitioning analysis showed that a larger fraction of variation in data was attributed to the horizon effect (29.6 – 32.7%) than to site effect, conflating spatial distance and variation in depth and temperature (9%), or sediment composition (1.8%) (Fig. 15C).

Biomarker analysis strengthened this observation (Fig. 15D). We split our samples between surface horizons (0-10 centimeters below the seafloor (cmbsf)) and subsurface horizons (10-40 cmbsf), based on the findings of Petro et al. (2019) highlighting a shift in community composition at the bottom of the bioturbation zone. We then determined three sets of site-specific surface biomarker ASVs, together with one set of general surface biomarkers, and one set of general subsurface biomarkers. Represented in Fig. 15D is the contribution of each set of biomarkers to the communities at each site, in the surface (left) and subsurface (right) horizons. Site-specific biomarkers made up a small fraction of the community (< 1.8%) while surface and subsurface ones accounted for 36.9-45.6%. As seen in the taxonomic profiles, the surface biomarkers were assigned to a variety of taxonomic groups, including classes Nitrososphaeria, Alpha- and Gamma-proteobacteria, while the subsurface biomarker ASVs mostly belonged to phyla Desulfobacterota, Acidobacteria, Chloroflexi, and class Bathyarchaeia (Fig. S9A).

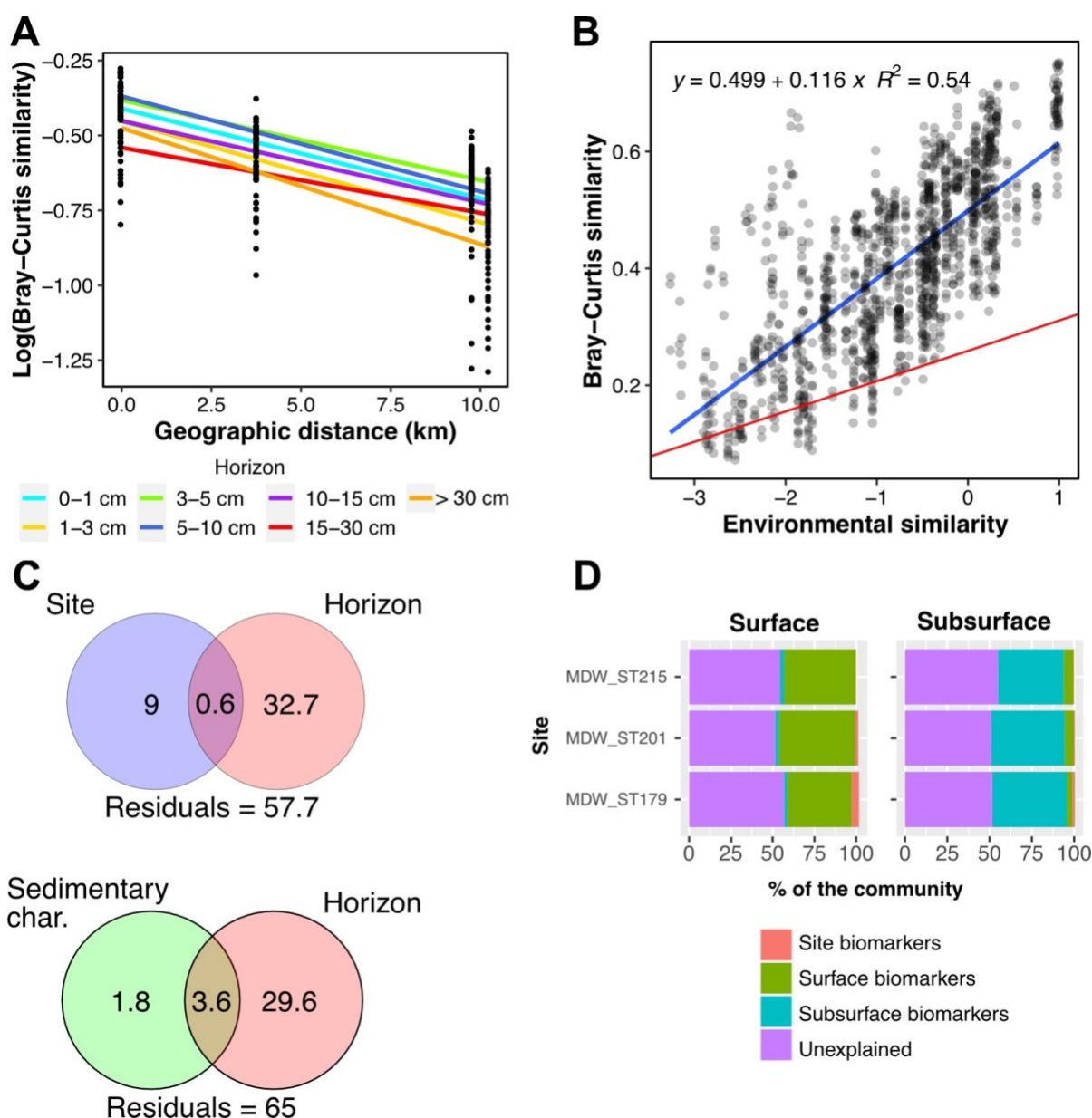


Figure 15: Local biogeographic patterns. Pairwise community similarity between Alborán Sea samples as a function of **(A)** geographic distance (km), modeled by horizon, and **(B)** environmental. In plot **(B)**, the blue line illustrates the linear regression for the Alborán Sea samples (p -value $< 2.2e^{-16}$), and the red line the regression for the complete dataset (p -value $< 2.2e^{-16}$). **(C)** Variation partitioning analysis for the Alborán Sea samples. **(D)** Relative composition of communities of the three sites of the Alborán Sea (Western Mediterranean Sea) based on the five sets of biomarkers identified: overall surface and subsurface biomarkers, and site-specific biomarkers (considering only surface horizons). Horizons between 0 and 10 cmbsf were considered surface horizons, and horizons deeper than 10 cmbsf were considered as subsurface horizons.

Discussion

In this work, using a rigorous, standardized slicing scheme and homogeneous molecular and bioinformatic analyses, we confirm at local, regional, and inter-regional scales the existence of strong biogeographic patterns for prokaryotic communities across the Mediterranean-Atlantic transition. Patterns emerged revealing regional and inter-basin differentiation following a distance-decay relationship for all scales considered. We further investigated present day environmental drivers and main evolutionary forces shaping the composition of prokaryotic communities populating the seafloor. In addition to the longitudinal structuration of communities, we confirm a systematic vertical stratification also reported at different depth scales in previous studies (Durbin and Teske, 2011; Orcutt et al., 2011; Jochum et al., 2017; Petro et al., 2019; Lloyd et al., 2020). We took advantage of the longitudinal extent of our dataset to investigate what processes might be at play in the assembly of microbial communities in and just below the bioturbation zone.

On the importance of water depth: an environmental and biogeographic boundary?

The transition between the upper and lower continental slope, around 800-1000 mbsl, is often associated with sharp local changes in sea bottom and water column conditions. It delineates the upper bathyal zone, found below the mesopelagic waters between 200 and ~1000 mbsl, and the lower bathyal zone (1000 - 3500/4000 mbsl) (Watling et al., 2013; Costello et al., 2017). Our results indicate that this transition is also associated with marked changes in benthic microbial community structure as shown in the multidimensional analysis (Fig. 14B). In terms of alpha-diversity, no decline with horizon depth was observed in the upper bathyal zone, while all samples of the lower bathyal zone exhibited such a trend (Fig. S8B). In terms of beta-diversity, communities present above 800 mbsl and below 1200 mbsl clustered independently of temperature, which was very variable in this study, from the warmer Mediterranean to the colder Atlantic waters. This segregation was reflected in the larger

CHAPTER 2

amount of ASVs shared among shallow sites of the Mediterranean Sea and the Gulf of Cádiz, rather than with geographically closer but deeper sites (data not shown). Several non-mutually exclusive hypotheses could account for such sharp ecological transition, among which i) the transition from piezotolerant to piezophilic microorganisms around 10 MPa, i.e. 1000 mbsl (Fang et al., 2010; Cario et al., 2019; Scoma, 2020), and ii) the nature of organic matter (OM) and its lability. Indeed, even though OM quantity did not vary significantly between depth zones (Fig. S11), more available OM may characterize sites closer to the shoreline in the upper bathyal zone (Seiter et al., 2005; Kallmeyer et al., 2012; Giovannelli et al., 2013).

When excluding the very distinct upper bathyal samples from the analysis, oceanic basin emerged as the second parameter influencing community structure: below 1200 mbsl, communities segregated according to basin origin (Atlantic versus Mediterranean). However, basins and temperature co-vary between the Atlantic and the Mediterranean, their respective contribution to beta-diversity has thus been difficult to disentangle, as illustrated by ordinations in Fig. 14A and 4C. Other available environmental variables partially describing habitats (sediment granulometry, water content, OM) showed similarities between oceanographic basins (Fig. 11B), and minimally contributed to the general beta-diversity variation partitioning analysis (5.2%, Fig. S10C). Finally, a latitudinal effect has been shown in other studies (Friedline et al., 2012). Here, no such correlation emerged, possibly due to the relatively narrow sampling zone, constrained between 20°N and 40°N.

Beyond depth: do historical or contemporary parameters drive community structure?

Given that the ecological processes influencing patterns of microbial community assembly are at play at any given time, it is necessary to consider their effects from a temporal as well as spatial point of view, thus distinguishing between historical and contemporary processes. In an influential review, Martiny et al. (2006) laid out a structured framework to interpret biogeographic data and the respective contribution of both types of processes. The authors defined “microbial habitats” as environments where microbial communities are structured by current ecological niche (defined by a set of biotic and abiotic parameters), while

CHAPTER 2

“microbial provinces” refer to regions that have undergone different historical processes, the legacies of which are visible in the contemporary structure of microbial communities. In the latter case, communities in equivalent niches but different provinces may harbor diverging communities. In our study, we thus used this framework and compared community and environmental similarity matrices to identify the contributions of these processes at different geographic scales.

At regional scale, we observed an important correlation of community similarity with both environmental similarity and geographic distance (Fig. 12A), indicating the presence of both distinct microbial habitats and different microbial provinces. At distances beyond regional scale however, only the distance-decay relationship remained visible, while correlations with environmental similarity were largely lost (Fig. 12B). Here, we cannot rule out the possibility that measurement of additional environmental parameters could lead to an increased link between community and environmental similarity, especially since the contribution of environmental selection to community structure was visible in the clustering patterns correlated with depth and temperature in the ordinations (Fig. 14B and C). It has also been put forward that dormancy and the presence of microbial spores, abundant in marine sediments (Wörmer et al., 2019), can affect the biogeographic patterns observed at the microbial level (Locey et al., 2020). Nevertheless, in spite of this potential “noise” in our data, DDRs remained clearly apparent at both the inter- and intra-regional scales. Overall, our results show that the influence of historical processes such as dispersal limitation and past environmental conditions supersedes contemporary influences at inter-regional scale.

Previous studies (Martiny et al., 2006, 2011; Lecours et al., 2015) have highlighted differences in beta-diversity patterns depending on the scale considered. At local scale (Alborán Sea), both the distance-decay relationship (Fig. 15A) and the link between community and environmental similarity (Fig. 15B) were apparent. Around 30% of ASVs were shared among all Alborán sites, quantitatively representing between 80% and 86% of reads. When focusing on the quantitative variation of these shared ASVs, the correlation between

CHAPTER 2

community and environmental similarity weakened (data not shown). This may reveal an influence of environmental selection mostly visible in the less abundant ASVs specific to each site, and/or dispersal limitation, with larger populations dispersing more easily and making up a core community of shared ASVs (Li et al., 2021). The presence of this biogeographic pattern at a scale of less than 10 km underlines the limited dispersal capability of benthic microorganisms (Zinger et al., 2011, 2014; Bienhold et al., 2016).

Environmental filtering and ecological drift in subsurface community assembly

The important link between sediment horizon and community richness and composition was evident from the sample ordinations (Fig. 14D), taxonomic composition (Fig. S8A), alpha-diversity patterns (Fig. S8B), and PERMANOVA analysis ($F= 5.11$, $p= 0.001$). The clearest changes in relative abundance were observed for Acidobacteria, Chloroflexi, and Bathyarchaeia, which all became more abundant deeper below the seafloor (Bienhold et al., 2016; Hoshino et al., 2020; Jørgensen et al., 2020; Lloyd et al., 2020; Vuillemin et al., 2020). In contrast, Desulfobacterota showed first an increase in relative abundance, before decreasing in deeper horizons in the upper bathyal zone, a pattern also described by Lloyd et al., 2020.

Recently, Petro et al. (2017, 2019) invoked selection of subsurface microorganisms locally from the surface community during burial as an important process driving subsurface community assembly. Here, we tried to quantify the relative contribution of stochastic versus environmental processes at local scale using site-specific biomarker analysis. When comparing three adjacent sites (< 10 km apart), we found that site-specific ASVs only marginally contributed to the total community (< 1.8%, Fig. 15D), whereas horizon-specific biomarkers, considered here as *proxies* for environmental filtering, were predominant (36.9-45.6% of reads). This finding is in line with the hypothesis of a strong environmental influence raised in previous work (Starnawski et al., 2017; Petro et al., 2017, 2019; Kirkpatrick et al., 2019; Marshall et al., 2019; Lloyd et al., 2020), even within the first decimeters of sediment. The observations from Petro et al. (2019) on the depth of influence of the bioturbation zone

CHAPTER 2

are also in line with the detection of macrofauna in a parallel metabarcoding study on these Mediterranean and Atlantic samples (Brandt et al., in prep).

In addition, we observed increasing rates of distance-decay with increasing depth in the sediment throughout the entire transect (Fig. 13, Table S1). In theory, an increase in DDR slope steepness can generally be explained by two processes: selection or lack of dispersal resulting in ecological drift (Hanson et al., 2012). In the case of selection, spatial auto-correlation of environmental parameters (e.g., salinity or temperature gradients) can lead to an increase of beta-diversity with distance. In this study, it is safe to assume that all environmental parameters possibly correlated with geographic distance (temperature, water depth) apply a similar pressure throughout the entire sediment horizons considered. We should thus observe parallel DDR regressions for the different sediment horizons, as community turnover rate would not vary with sediment depth. This is indeed what we observed at local scale (Fig. 15A), suggesting that although the composition of surface communities differed, they changed simultaneously while being buried, most probably due the similar set of environmental conditions encountered. In contrast, we observed a clear increase of the distance-decay rate with sediment depth over the whole dataset (Fig. 13, Table S1). We thus argue that, in this case, decreasing dispersal towards lower sediment horizons, leading to increasing ecological drift, is most probably responsible for the increase in microbial community differences across deep-sea sites with increasing sediment depth.

Conclusion

This study presents a large scale, high definition characterization of the spatial distribution of benthic bacteria and archaea at the transition between two oceanic basins. Overall, we observed strong biogeographic patterns over the transition between the Mediterranean Sea and the Atlantic Ocean that depended on the scale considered. While at local and regional scale, community composition seemed to reflect both the influence of historical processes and of current environmental conditions, at the inter-regional scale the legacy of historical processes appeared more prevalent. Water depth, ocean basin, and water temperature were important environmental drivers of community structure. We found that in addition to environmental filtering, dispersal limitation and ecological drift emerged as influential processes in shaping the evolution of benthic microbial community composition with increasing depth in the sediment.

In the future, the importance of stochastic biogeographic processes in the assembly of early subsurface microbial communities could be further investigated by applying neutral and null-model approaches (Stegen et al., 2013; LaBrie et al., 2020), which might be more adapted to detecting the influence of drift in particular. In addition, part of the unexplained variation detected in our data is probably linked to biotic interactions with organisms not covered in this study (Gralka et al., 2020; Tobias-Hünefeldt et al., 2020), and may thus be further elucidated with metabarcoding data being generated for metazoans and protists in the scope of this project.

CHAPTER 2

Author contributions:

LM and SA-H designed the study. SAH, MB, JP and CB carried out field and laboratory work. CO supplied ship time to conduct the sampling in Alborán Sea, Gulf of Cádiz, Ormonde seamount and Formigas seamount. BT performed the bioinformatic and statistical analyses. BT, J-CA, LM and SA-H contributed to data interpretation. BT, LM and SA-H drafted the manuscript. All authors contributed to the final manuscript.

Data availability statement:

The dataset generated for this study has been submitted to the European Nucleotide Archive (ENA) under the following project: PRJEB33873. Additionally, the full dataset, including raw sequences, processed reads and ASV tables, as well as bioinformatic scripts are available through github (https://github.com/loimai/ABYSS_16S/tree/master/docs).

Acknowledgements:

This work was part of the “Pourquoi Pas les Abysses?” project funded by Ifremer and the project eDNAbyss (AP2016-228) funded by France Génomique (ANR-10-INBS-09) and Genoscope-CEA. It was also co-funded by a grant from the University of Western Brittany (UBO) through the Ecole Doctorale des Sciences de la Mer et du Littoral (EDSML). This study was supported by the European Union’s Horizon 2020 Research and Innovation Program, under the ATLAS project (Grant Agreement No. 678760). This output reflects only the author’s view, and the European Union cannot be held responsible for any use that may be made of the information contained therein.

The samples were collected during research cruises: MEDWAVES, PEACETIME (DOI 10.17600/17000300), AMIGO 2017, CanHROV and ESSNAUT16 (DOI 10.17600/16000500). The authors wish to thank all the participants of the cruises who aided in the sampling efforts for this study, and in particular: the captain and crew of the R/V Atalante and the Nautilie submersible team for their efficiency, as well as the chief scientist, Marie-Anne Cambon-

CHAPTER 2

Bonavita, and scientific parties of ESSNAUT16, Cécile Guieu and Christian Tamburini for the sampling on board the R/V Atalante during cruise PEACETIME, Sophie Arnaud-Haond, François Bonhomme and the crew of R/V Atalante during cruise AMIGO 2017, Marie-Claire Fabri and the crew of R/V Europe and Ariane HROV during cruise CanHROV and Perregrino Cambeiro, Joana R. Boavida, Juancho Movilla, Maria Rakka, Anna Maria Addamo and Meri Bilan for sampling on board R/V Sarmiento de Gamboa during the MEDWAVES cruise.

Supplementary figures and tables

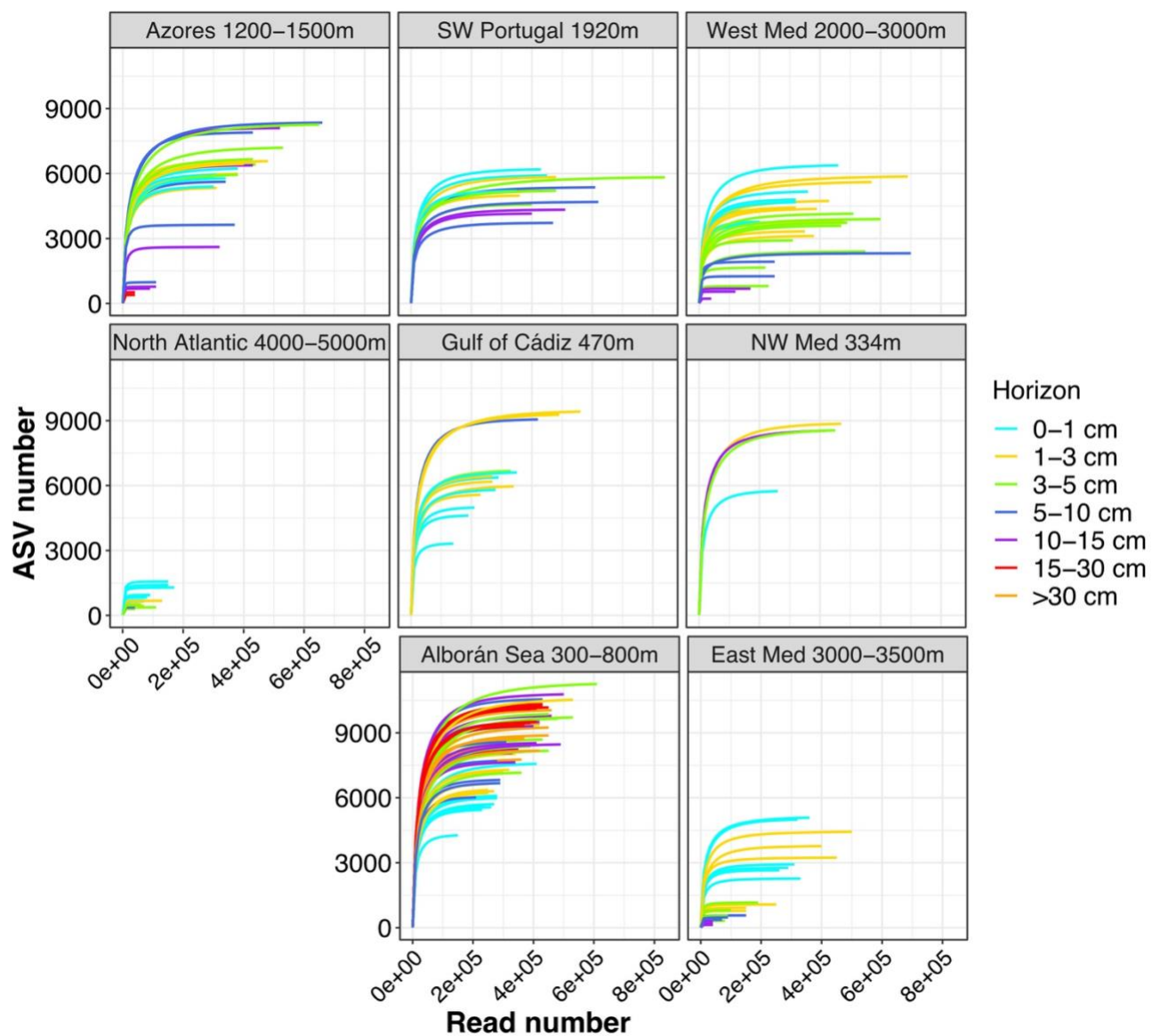
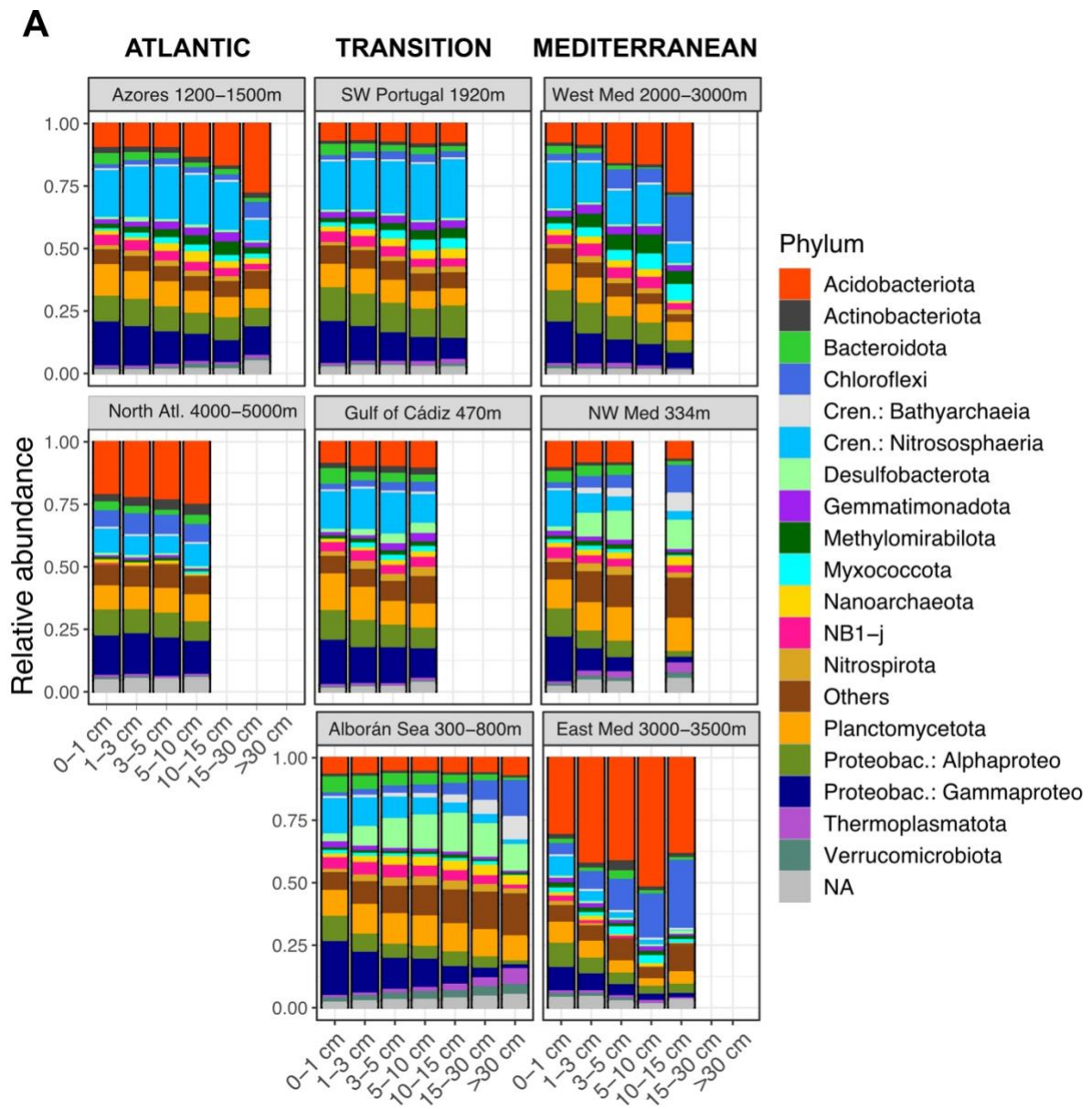


Figure S7 : Rarefaction curves for each sample in the metabarcoding dataset, arranged by sampling location and colored by sediment horizon.



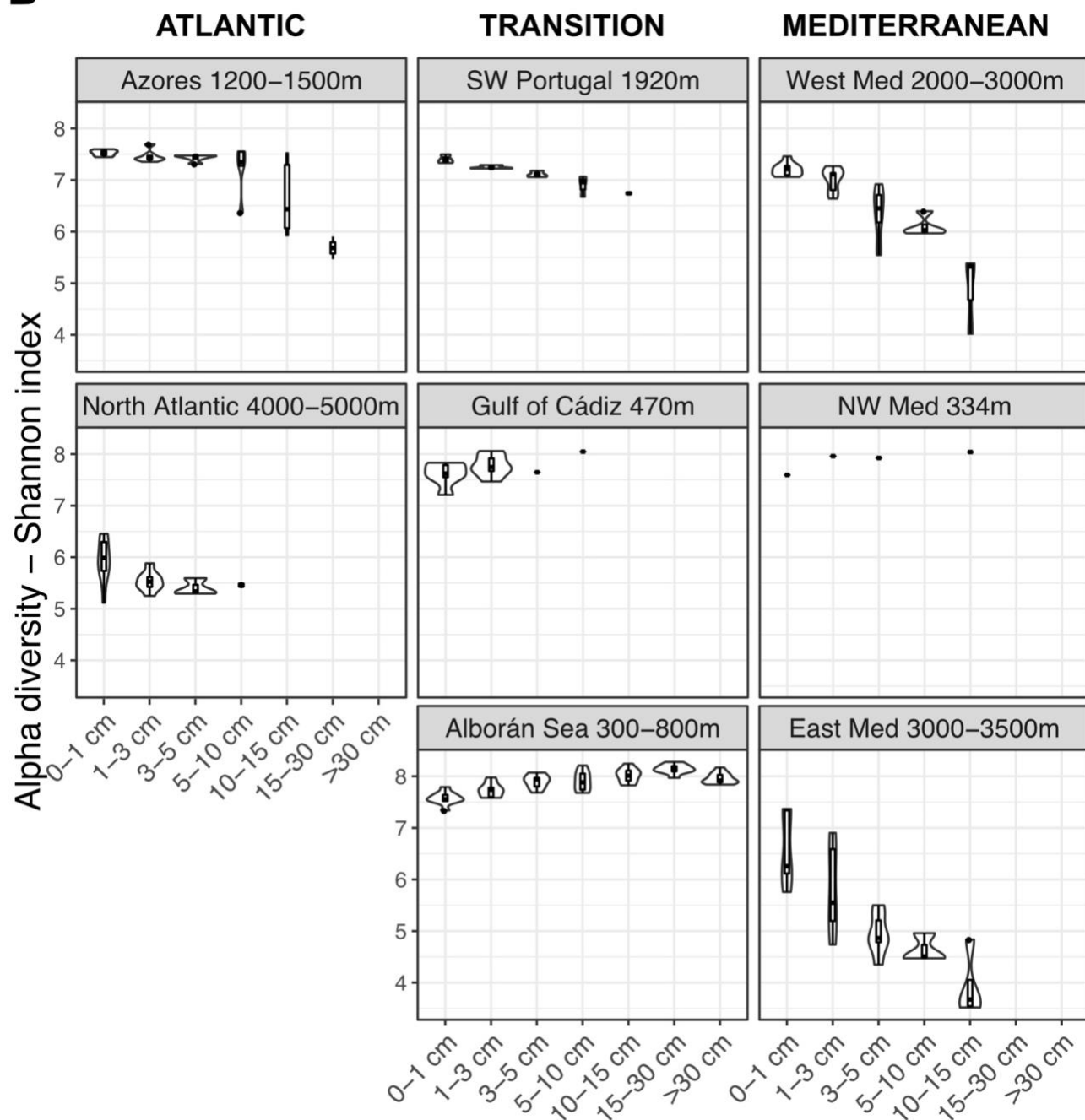
B

Figure S8 : Taxonomic profiles and alpha-diversity estimates. **(A)** Relative abundance profiles of the 16 most abundant phyla in the dataset grouped by sampling location with increasing horizon depth (depth below the seafloor). Please note that Proteobacteria and Crenarchaeota members are identified at the class level for clarity. **(B)** Estimated alpha diversity (Shannon index) in samples grouped by sampling location and ordered by sediment horizon.

CHAPTER 2

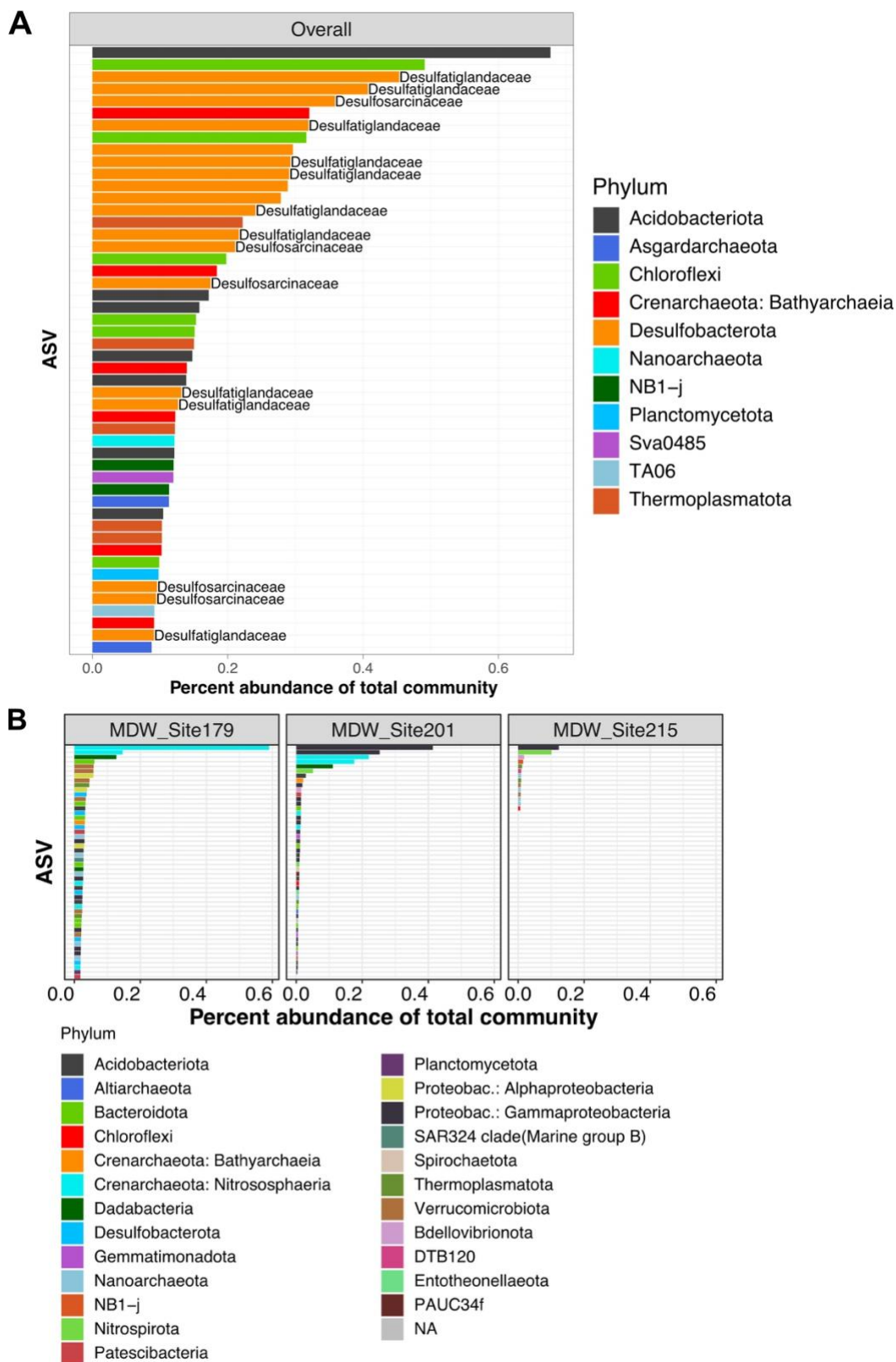


Figure S9: Taxonomy of the 50 most abundant Alborán Sea biomarker ASVs for **(A)** the overall subsurface biomarker set and **(B)** the site-specific biomarker sets

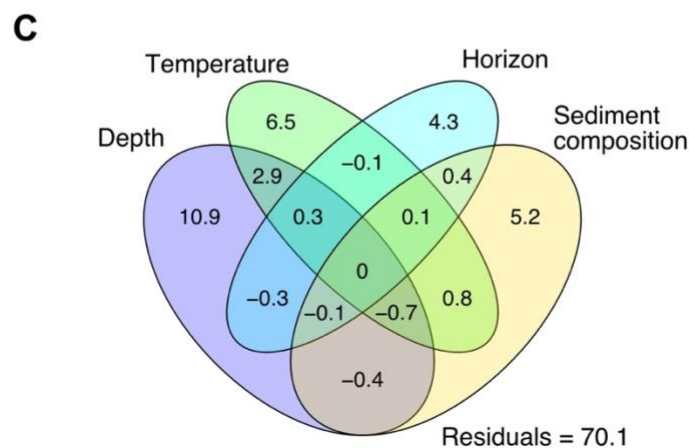
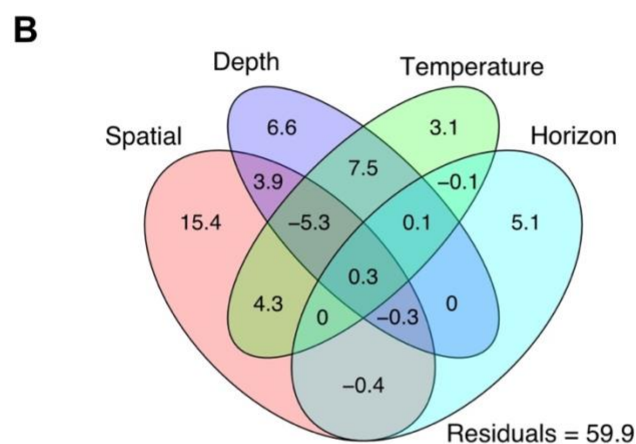
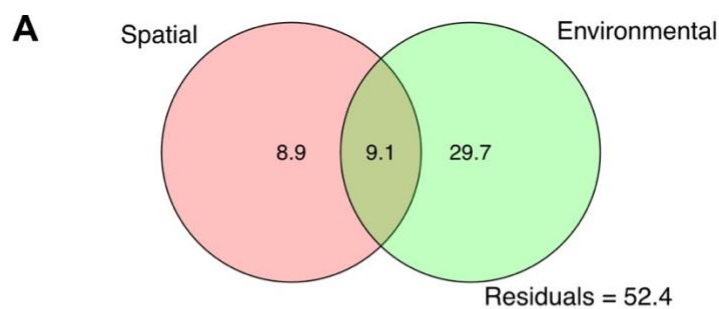


Figure S10: Variation partitioning analysis of the data using Bray-Curtis dissimilarity. Partitioning according to (A) spatial vs environmental components, (B) spatial component, water depth, temperature, and sediment horizon, (C) water depth, temperature, sediment horizon, and sediment composition. The spatial component (A, B) refers to a combination of latitude, longitude and squared latitude. Sediment composition (C) refers to the combination of humidity level, organic matter content, mean granulometry and particle size (heterogeneity). Finally, the environmental component in (A) is calculated using all the sediment variables

CHAPTER 2

mentioned above, combined with water depth, temperature, distance from shore and horizon depth. Fractions are annotated with the obtained adjusted R square for each explanatory variable or matrix, for which significance was tested. Values for the intersections are found by subtracting different models and underline the possible redundancy of the explanatory variables. They cannot be assigned significance, and negative values are a possible artefact of the analysis.

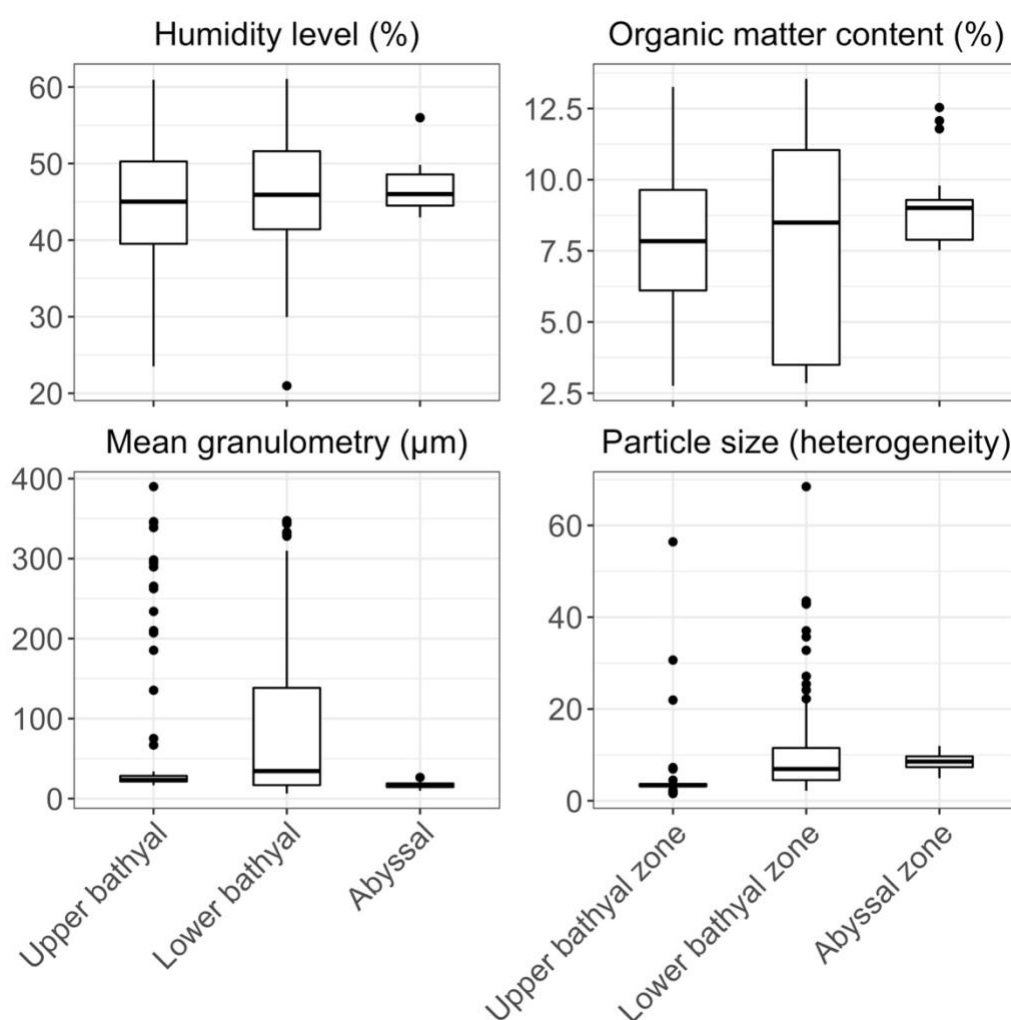


Figure S11: Evolution of sediment characteristics in the three depth zones targeted by the longitudinal sampling scheme. Only mean heterogeneity of particle sizes significantly differed between depth zones.

CHAPTER 2

Table S1: Values of linear regression parameters computed for each sediment horizon on the whole dataset. Formula: $\log(\text{Bray-Curtis similarity}) = f(\text{geographic distance})$

Horizon	Slope	Intercept	R ²	p-value
0 - 1 cmbsf	0.000446	-1.49	0.34	< 2.2e-16
1 - 3 cmbsf	0.000527	-1.68	0.4	< 2.2e-16
3 - 5 cmbsf	0.000499	-2.07	0.24	< 2.2e-16
5 - 10 cmbsf	0.000657	-2.03	0.29	< 2.2e-16
10 - 15 cmbsf	0.000811	-2.43	0.17	2.805e-13
15 - 30 cmbsf	0.00255	-0.680	0.97	< 2.2e-16

Table S2: Values of linear regression parameters computed for the Alborán Sea samples for each sediment horizon. Formula: $\log(\text{Bray-Curtis similarity}) = f(\text{geographic distance})$

Horizon	Slope	Intercept	R ²	p-value
0 - 1 cmbsf	-0.030	-0.411	0.58	4.162e-08
1 - 3 cmbsf	-0.034	-0.453	0.46	3.536e-06
3 - 5 cmbsf	-0.027	-0.382	0.76	2.963e-12
5 - 10 cmbsf	-0.032	-0.370	0.91	< 2.2e-16
10 - 15 cmbsf	-0.027	-0.451	0.64	3.189e-09
15 - 30 cmbsf	-0.022	-0.541	0.37	4.68e-05
30+ cmbsf	-0.039	-0.475	0.42	6.546e-08

CHAPTER 3

Distribution of Archaea and putative associations of members of the Nanoarchaeota phylum in abyssal and hadal surface sediments revealed by network analysis

**Distribution of Archaea and putative associations of members of
the Nanoarchaeota phylum in abyssal and hadal surface sediments
revealed by network analysis**

Blandine Trouche¹, Ferial Boudarka¹, Clemens Schaubberger², Jean-Christophe Auguet³,
Caroline Belser⁴, Julie Poulain⁴, Bo Thamdrup², Patrick Wincker⁴, Ronnie N. Glud², Sophie
Arnaud-Haond³ and Loïs Maignien^{1,5}

¹ Univ Brest, CNRS, IFREMER, Microbiology of Extreme Environments Laboratory (LM2E), F-
29280 Plouzané, France

² Hadal & Nordcee, Department of Biology, University of Southern Denmark, Odense,
Denmark

³ MARBEC, Univ Montpellier, Ifremer, IRD, CNRS, Sète, France

⁴ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Évry,
Université Paris-Saclay, 91057 Evry, France

⁵ Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular
Biology and Evolution, Woods Hole, MA, United States

Draft paper

Résumé de l'article en français

La diversité taxonomique et fonctionnelle présente dans les sédiments marins profonds a encore été peu explorée. De récents efforts de caractérisation de la diversité microbienne présente dans les sédiments benthiques des fosses Kermadec et Atacama ont souligné l'influence marquée de la biogéochimie dans la structuration des communautés. Dans cette étude, nous nous sommes intéressés plus particulièrement à la diversité archéenne détectée dans les mêmes sites en analysant 170 librairies de métabarcoding obtenues à partir d'échantillons de sédiments provenant de sept sites dans l'axe des fosses, et quatre sites situés sur les plaines abyssales adjacentes. Nous avons observé une prédominance de *Nitrososphaeria* (nom proposé pour les Thaumarchées dans SILVA 138), en particulier pour les sites abyssaux et les premières couches de sédiment des sites hadaux. Les *Woesearchaeales* (proposition de nom au niveau de l'ordre pour le phylum candidat *Woesearchaeota* dans SILVA 138) quant à elles sont la lignée principale détectée dans les horizons les plus profonds des sites hadaux et de la pente continentale. Une analyse en réseau de cooccurrence a mis en avant pour ces lignées des schémas de distribution liés à l'habitat (abyssal ou hadal) et à la profondeur dans le sédiment, reflétant sans doute le contexte géochimique, et en particulier les différences de profondeur de pénétration de l'oxygène entre sites. L'analyse de la modularité du réseau a aussi souligné la distinction entre communautés associées aux zones abyssales et hadales, sans ségrégation apparente pour les sites partageant le même habitat mais géographiquement éloignés. En conséquence, nous avons observé une séparation par niche écologique qui pourrait coïncider avec des sous-genres spécifiques pour les deux principales lignées taxonomiques. L'analyse du réseau n'a pas permis de montrer la présence de fortes associations spécifiques entre les membres des *Woesearchaeales* et les autres lignées d'Archées, ce qui semble renforcer l'hypothèse d'une complémentarité métabolique non spécifique.

Abstract

Deep sea sediments hold a wealth of undescribed taxonomic and functional diversity. Recent efforts to characterize the microbial diversity in benthic sediments of the Kermadec and Atacama trenches have highlighted the importance of biogeochemistry in shaping community structure. Here we focused on the archaeal diversity found in the same locations and analyzed 170 metabarcoding libraries obtained from sediment samples from seven trench axis sites and four sites located on the adjacent abyssal plains. We observed a predominance of Nitrososphaeria (proposed name for Thaumarchaeota in SILVA 138), particularly at abyssal sites and in the surface layers of hadal sites. Woesearchaeales (proposed order-level name for candidate phylum Woesearchaeota in SILVA 138) prevailed in deeper horizons of the hadal and continental shelf sites. A co-occurrence network analysis revealed patterns of distribution for these lineages in relation with habitat (hadal or abyssal) and sediment depth, seemingly reflecting geochemical context, and in particular differences in oxygen penetration depth between sites. Modularity analysis also highlighted a clear distinction between communities associated to the abyssal and hadal zones, without any apparent separation between geographically distant sites of the same habitat. As a result, we observed possible ecological niche separation that could coincide with subclades in the two most abundant lineages. We did not detect signs of strong specific associations between Woesearchaeales and other archaeal lineages, strengthening hypotheses of non-specific metabolic complementation.

Introduction

Archaea play crucial roles in marine subsurface sediments and biogeochemical cycles (Biddle et al., 2006; Lipp et al., 2008; Offre et al., 2013; Vuillemin et al., 2019). At the interface between these sediments and pelagic waters, benthic layers are the seat of early diagenesis, the first steps in the remineralization of sinking organic matter (Froelich et al., 1979). As a consequence, microbial communities of these layers are determining players in the partitioning between buried organic matter and nutrients released in the water column (Deming, 1985). In addition, it has been suggested that subsurface communities assemble through selective survival of benthic microorganisms (Petro et al., 2017), making the description of benthic Archaea an important step toward understanding deep subsurface life. However, the contribution and biogeochemical role of Archaea to benthic ecosystem functioning is still sparsely described in sediments of abyssal plains and hadal trenches.

A vast majority of archaeal diversity is only known through molecular approaches such as metabarcoding, metagenomics or single-cell genomics (reviewed in Adam et al., 2017; Spang et al., 2017; Baker et al., 2020). Indeed, Next Generation Sequencing has been instrumental in the continual improvement of our knowledge of Archaea. Among other discoveries, the DPANN superphylum was proposed in 2013 (Rinke et al), originally including phyla Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota. Since then, further studies have described new candidate phyla seemingly belonging to this monophyletic and deep-branching clade: Altiarchaeota (Probst et al., 2013), Huberarchaeota (Probst et al., 2018), Micrarchaeota (Comolli et al., 2009) and Pacearchaeota (Takai and Horikoshi, 1999). Most members of the DPANN exhibit small cell sizes and reduced genomes (Dombrowski et al., 2019), and the few microorganisms that were successfully enriched in co-culture were shown to be obligate symbionts of archaeal hosts (Huber et al., 2002a; Podar et al., 2013; Wurch et al., 2016; Golyshina et al., 2017; Krause et al., 2017; St. John et al., 2019). The first of these to be cultivated was *Nanoarchaeum equitans*, an ectosymbiont of *Ignicoccus*

CHAPTER 3

hospitalis (Huber et al., 2002a). However, the hypothesis that the DPANN superphylum is dominated by symbionts has recently been questioned, in favor of a highly communal lifestyle and fermentative metabolism (Beam et al., 2020).

Candidate phylum Woesearchaeota, classified as order Woesearchaeales in the latest version of SILVA (v138, Quast et al., 2013) and initially referred to as Deep-sea Hydrothermal Vent Group 6, is rather widely distributed in diverse environments (e.g. Castelle et al., 2015; Ortiz-Alvarez and Casamayor, 2016; Han et al., 2017; Xu et al., 2017). In fact, a recent meta-analysis reported a wide diversity of Woesearchaeales subgroups distributed in different biotopes, yet with a majority of anoxic environments (Liu et al., 2018). Co-occurrence analysis suggested a potential syntrophic relationship between Woesearchaeales and anaerobic methanogenic archaea in inland ecosystems. Due to their recent addition to databases, these small-celled archaea are often partially missed in metabarcode-based diversity inventories (Eloe-Fadrosh et al., 2016; Bahram et al., 2019, Chapter 1.4). They were yet recently found to be abundant members of archaeal communities of the deeper layers of benthic hadal sediments (Peoples et al., 2019; Hiraoka et al., 2020; Schaubberger et al., 2021b).

In the oxic surface layers of hadal sediments, as well as throughout the sediment column of abyssal plains, Thaumarchaeota, reclassified as class Nitrososphaeria in the latest version of the SILVA database (v138), dominate archaeal communities (Peoples et al., 2019; Vuillemin et al., 2019; Hiraoka et al., 2020; Hoshino et al., 2020; Schaubberger et al., 2021b). These archaea are also widely distributed in pelagic environments (Francis et al., 2005), and all described ammonia-oxidizing archaea (AOA) belong to this lineage (Alves et al., 2018). As a consequence, they are expected to be important contributors to the nitrogen and carbon cycle in marine sediments.

Hadal trenches are the deepest marine points on Earth, situated on the subducting borders of tectonic plates (Bruun, 1956; Jamieson et al., 2010). Among other parameters, they are distinguishable from adjacent abyssal plains by the increased hydrostatic pressure experienced at the bottom of the trench, as well as increased input of organic matter due to topographical funneling and mass-wasting events (Danovaro et al., 2003; Turnewitsch et al.,

CHAPTER 3

2014; Kioka et al., 2019). Because of this heightened influx of substrate, hadal trenches are hotspots of biogeochemical activity in the deep sea (Ichino et al., 2015; Glud et al., 2021). Indeed, oxygen consumption rates, often used as a measure of biological activity and benthic carbon mineralization, have been estimated to be on the order of a few centimeters to ~10 cm in the Kermadec and Atacama trenches, while in adjacent abyssal plains depletion of oxygen was not reached at 20 cm (Glud et al., 2021). A study of the microbial communities found in the benthic sediments of these trenches has highlighted phylum-level patterns of similarity along the trench axes and strong variations in community composition with sediment depth in conjunction with redox stratification (Schauberger et al., 2021b). Similar results were also observed in other Pacific trenches, including the distinctness between abyssal and hadal communities (Peoples et al., 2019; Hiraoka et al., 2020).

Here, we aimed at characterizing the archaeal communities found in benthic sediments of the Kermadec and Atacama trenches and adjacent abyssal plains. Using co-occurrence networks, we investigated the patterns of distribution, and in particular the possible distinctness or connectivity between trenches and with the shallower sites. Finally, we attempted to outline putative associations of members of the Woesearchaeales with other archaeal lineages.

Material & Methods

1. Sample collection and processing

Samples for this study were collected during the two South Pacific HADES-ERC cruises described in Chapter 1.4: a first cruise to the Kermadec trench in November 2017 and a second cruise to the Atacama trench in March 2018 (Fig. 16A). During the first one, sites K6 (in the trench axis) and K7 (on the adjacent abyssal plain) were sampled using a multicorer and a boxcorer respectively, recovering one core from each site (Fig. 16B). The second set of samples collected during the cruise to the Atacama trench were all obtained using a multicorer. Triplicate cores were recovered from six sites in the trench axis (A2, A3, A4, A5, A6, A10), one bathyal (A1) and one abyssal site (A9) located between the trench and the coastline, and one abyssal site (A7) lying under the open ocean, west of the trench (Fig. 16C).

As in chapter 1.4, the recovered sediment cores were sliced immediately after getting onboard in a 3°C cold room using equipment previously bleached and rinsed with ethanol and nanopure water. Each core was sliced into depth layers following a standard scheme (0-1 cm, 1-3 cm, 3-5 cm, 5-10 cm, 10-15 cm, and 15-30 cm), with always at least 1 cm left on the extruder to avoid contamination. Slicing was performed using spatulas also bleached and rinsed with nanopure water before each use. Samples were then transferred into zip-lock bags, homogenized, and frozen at -80°C on board before being shipped on dry ice to the laboratory where they were also kept at -80°C. At station A9 and A1, empty bags were also conditioned on board to be later used as field controls.

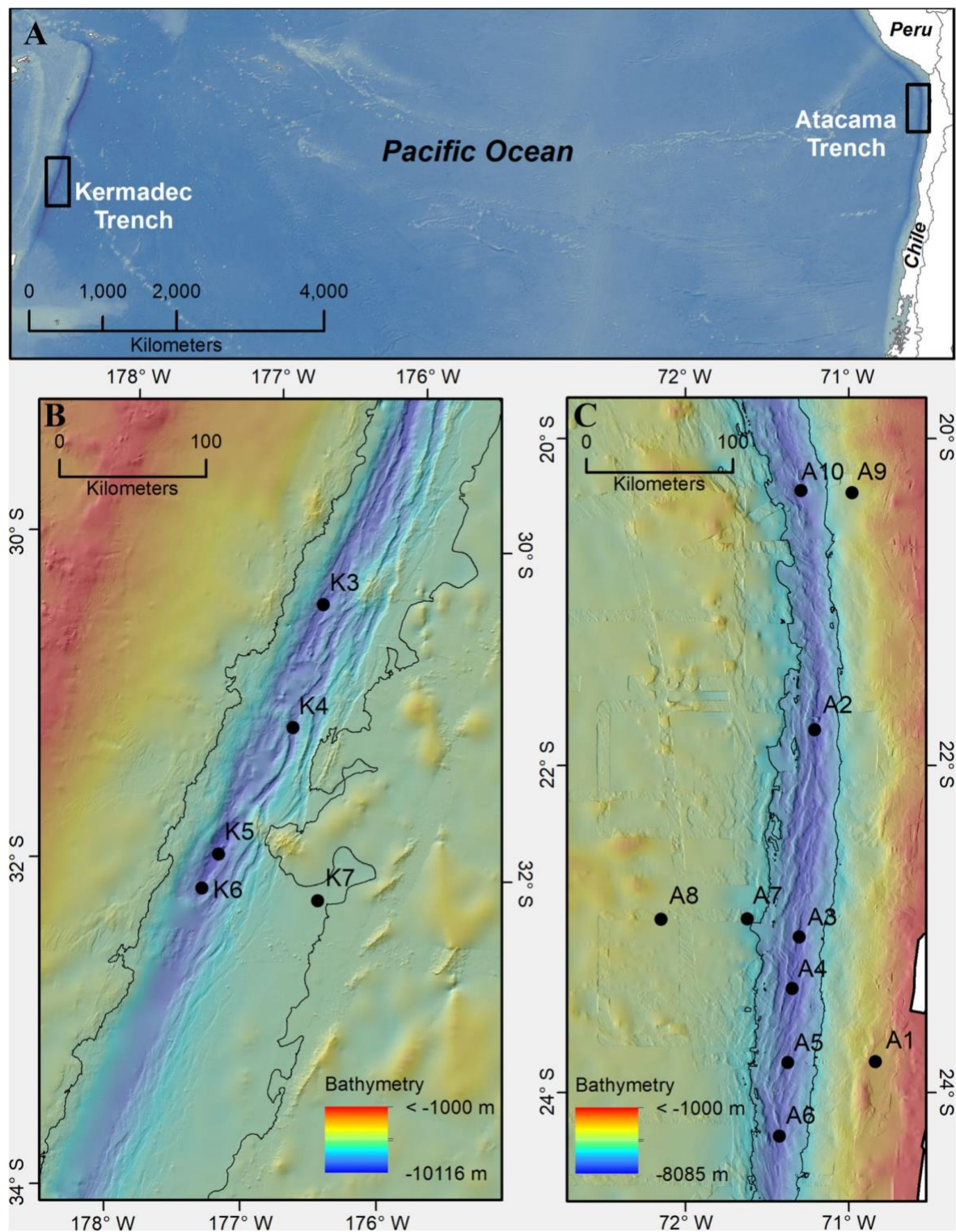


Figure 16: Map of the study areas in the South Pacific Ocean (A). Bathymetric maps with sampling sites in the Kermadec Trench (B) and the Atacama Trench (C). Extracted from (Schauberg et al., 2021). All bathymetry data were obtained from the Global Multi-Resolution Topography Synthesis (Ryan et al., 2009). Black lines in B and C indicate the 6000 m depth contour.

CHAPTER 3

2. DNA extraction, library construction and sequencing

DNA extractions were performed using 10g of sediment as presented in chapters 1.4 and 2. For this study, the universal primer pair by Parada et al. (2015) targeting the V4V5 region of the 16S rRNA gene was used (515F-926R), based on the results of chapter 1.4. Library preparation and sequencing were carried out at Génoscope (Evry, France) as detailed in chapter 1.4.

3. Bioinformatic processing

Bioinformatic processing of the metabarcoding dataset was performed using the standardized pipeline implemented and tested in chapter 1.3 (Brandt et al., 2021), available on Gitlab (<https://gitlab.ifremer.fr/abyss-project/abyss-pipeline>), on a home-based cluster (DATARMOR, Ifremer).

This processing was carried out as detailed in chapter 1.4, with the exception that after decontamination of the data, the sample with the lowest number of sequences (24,019) was removed from the dataset. All other libraries contained more than 250,000 sequences. The library with the maximum number of reads was also an outlier, totaling over 3 million reads, and was randomly subsampled to 1,125,000, a number similar to the second highest number of reads.

The reads were taxonomically assigned in DADA2 v1.10 with the RDP naive Bayesian classifier (Wang et al., 2007), using the SILVA v138 reference database (Quast et al., 2013) and a bootstrap threshold of 80. Subsequent analyses were run only on ASVs identified as Archaea.

4. Statistical analysis

Subsequent statistical analyses were done mostly in R v3.6.1, using phyloseq (v1.28.0, McMurdie and Holmes, 2013), vegan (v2.5.7, Oksanen et al., 2015) and ggplot2 (v3.3.0, Wickham, 2016) packages to compute alpha diversity, and produce taxonomy barplots.

The archaeal network was computed using pairwise covariance with SPIEC-EASI (Kurtz et al., 2015) on a filtered phyloseq object containing only the most abundant ASVs, with more than 50 sequences in at least three samples. This filtered object represented 7.1% of archaeal ASVs but 90.4% of their abundance. Edges with positive weights were then used to determine modules of highly interconnected ASVs using function *cluster_louvain* and *make_clusters* of igraph (Csardi, 2013), with the Louvain algorithm (Blondel et al., 2008). The network of the main correlations (> 0.5 edge weight) was visualized in Gephi (Bastian et al., 2009) with Force Atlas 2 layout algorithm. The following analysis of module composition and distribution was done in R v3.6.1 with the same packages as stated above.

Results

1. Dataset description

A total of 171 sequence libraries were obtained after amplification with the universal primers from Parada et al. (2015) and sequencing produced 155,713,315 raw sequences of the 16S rRNA gene V4-V5 region, with a mean of 627,876 reads by library. An additional 19,951,385 reads were recovered from the 25 control libraries that were constructed and sequenced simultaneously, including sampling (empty bags of storage conditioned on-board research vessels at the end of some of the sampling sessions), extraction (empty kit processed through all extraction steps together with the samples) and PCR (nanopure water) controls.

After processing with DADA2 the dataset included 106,675,420 sequences distributed among 176,216 ASVs. Of those, 2040 (about 1.2%) were found in control libraries, with 1337 ASVs exclusive to these libraries. A specific ASV accounted for 99% of all negative control libraries. This ASV, affiliated with partial 16S sequences of *Sphingobium* strains, has been recognized by the manufacturer as a pervasive contaminant of Taq-Phusion reagents (Salter et al., 2014), a contamination that occurred in all commercial kits up to 2019.

After decontamination and removal of eukaryotic, mitochondrial and chloroplast sequences, the dataset comprised 170 sample libraries with 85,272,666 sequences representing 172,107 ASVs. 18.8% of these ASVs (32,388 ASVs adding up to 15.8% of sequences) were assigned to domain Archaea.

2. Overview of archaeal diversity

The mean ratio of sequences identified as Archaea in the samples considered was 16%, with a wide variation of estimates, between 5.03% and 37.5% (Fig. 17). The lowest value was

CHAPTER 3

obtained from the Atacama hadal site A10 in horizon 3-5 cm, and the highest value in horizon 10-15 cm of open ocean abyssal site A7 (adjacent to the Atacama trench). Overall, the percentage of archaeal sequences was higher in abyssal samples on the open ocean side of the trench (A7 and K7) than in hadal samples (Kruskal-Wallis rank sum test, p -value = 1.181×10^{-13}), and at the Atacama abyssal site this percentage increased with depth in the sediments.

In terms of taxonomic diversity, the most abundant class detected was Nitrososphaeria (SILVA 138 classification for Thaumarchaeota), with 59.3% of archaeal sequences, but interestingly only 3.8% of ASVs (Fig. 18). Conversely, the second most abundant class, Nanoarchaeia, totaled 33.1% of archaeal sequences but 81.8% of ASVs. All of these ASVs were classified at order level as Woesearchaeales (SILVA 138 classification for *Ca.* Woesearchaeota).

Nitrososphaeria dominated surface horizons at all the sampled sites, and the open ocean abyssal communities up to 30 cm (deepest horizon sampled). In sediments from the trench axis and the continental plate, there was an increase in diversity with increasing sediment depth, mostly illustrated by an increase in the relative abundance of Woesearchaeales (Fig. 18). This was reflected in alpha diversity patterns, with higher estimates in deeper sediments of these sites, and the highest alpha diversity computed for the bathyal site (Fig. S12). At this bathyal site (A1), a higher relative abundance of Thermoplasmata and Bathyarchaeia was detected. Lokiarchaeia were also important members of archaeal communities at both landward sites, bathyal and abyssal, and 27% of unassigned sequences at class level were actually members of the Asgardarchaeota (Fig. 18). Finally, at the Kermadec trench site Woesearchaeales dominated the archaeal community below 10 cm. At the Atacama trench sites more taxonomic diversity was detected below 5 cm, with varying relative contributions of Bathyarchaeia, Hydrothermarchaeia and Lokiarchaeia depending on the site (Fig. S13).

CHAPTER 3

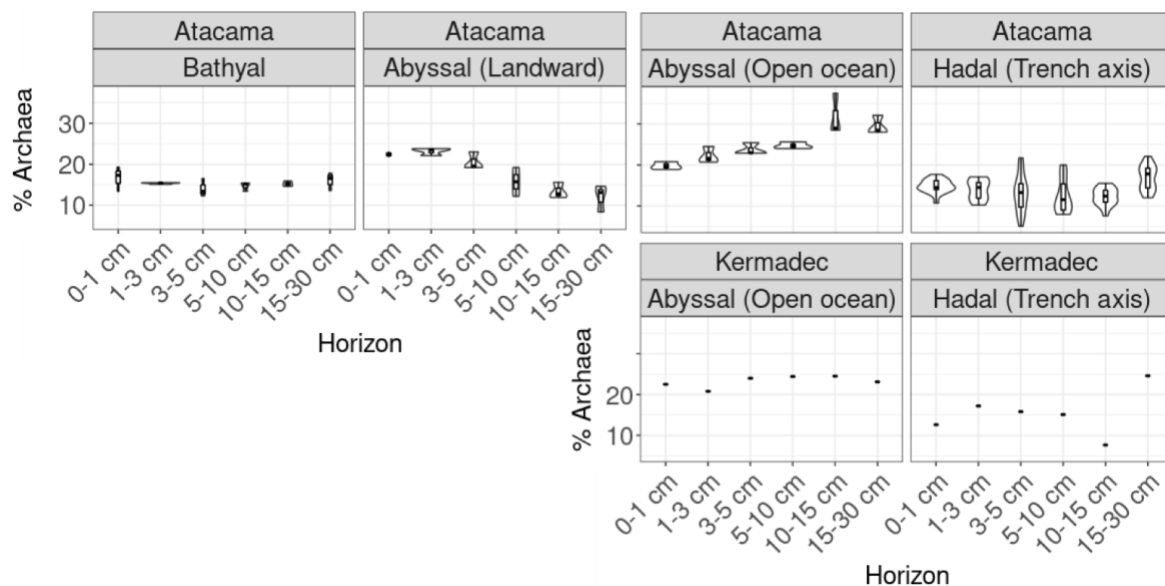


Figure 17: Percentage of sequences identified as Archaea in each sample, grouped by trench, zone, and horizon depth.

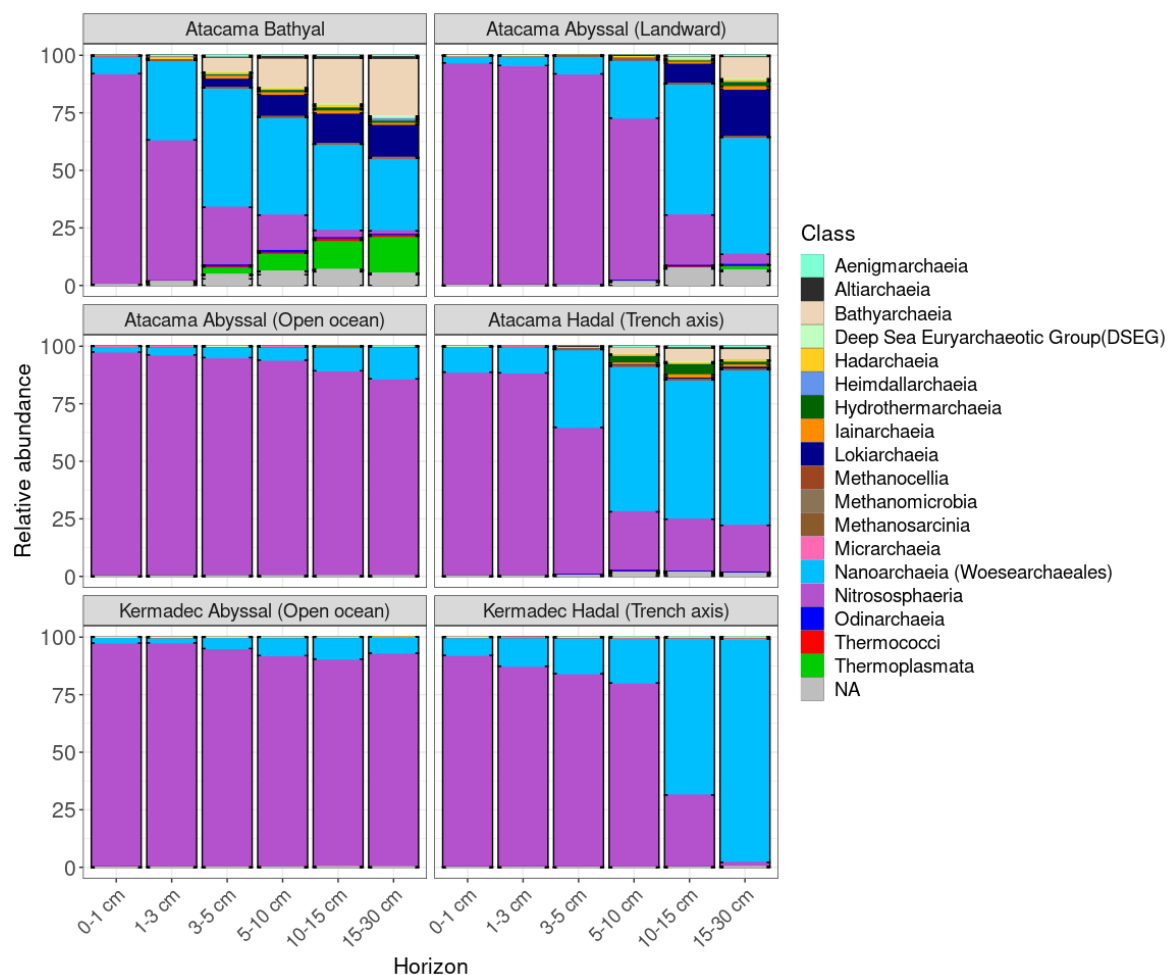


Figure 18: Archaeal taxonomic profiles at class level, grouped by trench, zone, and horizon depth. Taxonomic assignment was based on SILVA 138.

3. Archaeal co-occurrence network

The co-occurrence network obtained from the archaeal ASVs presenting more than 50 sequences in over three samples consisted of 2314 nodes (ASVs) and 51,720 edges (links). Modularity analysis with the Louvain algorithm revealed seven modules (Fig. 19A), with clear patterns of distribution depending on habitat type and horizon depth (Fig. 19B). Interestingly, no module appeared to be specific to a trench region when the same type of habitat was sampled in both Kermadec and Atacama regions.

Modules 5, 6 and 7 were almost exclusively present in the landward sites: module 7 in the surface of the bathyal site, module 6 in the deeper horizons of both sites and module 5 in the deeper horizons of the abyssal site (Fig. 19B). These modules thus reflected the taxonomic diversity described above at these sites (Fig. 20).

Additionally, module 1 made up a large part of the surface communities of the Atacama landward abyssal site, and dominated open ocean abyssal sites (Fig. 19B). This module was composed almost entirely of Nitrososphaeria (Fig. 20). When visualizing the network with edges of weight over 0.5, module 1 seemed to tend towards a split in two sub-modules (Fig. 19A). Closer inspection of the distribution of ASVs assigned to this module revealed two principal modes of distribution: decrease or increase in abundance with sediment depth (Fig. S14). ASVs becoming more abundant with sediment depth were shared between abyssal open ocean sites of both trenches, but seemed very scarce in samples from the landward abyssal site.

Module 2 was also mostly composed of Nitrososphaeria (Fig. 20), and largely dominated hadal sites surface layers both in Atacama and Kermadec (Fig. 19B). Its contribution decreased with sediment depth, starting at horizon 3-5 cm overall in Atacama, and 10-15 cm in Kermadec. Module 2 was also detected at abyssal sites to a lesser extent, with a similar pattern of decrease in abundance with sediment depth.

CHAPTER 3

Finally, modules 3 and 4 were characteristic of deeper horizons of hadal sites, and mostly composed of Woesearchaeales (Fig. 19B, Fig. 20). In the Kermadec trench hadal site, both modules increased in abundance in parallel, while in Atacama hadal sites, module 4 increased up to 10 cm depth, before decreasing, and module 3 became dominant below 10 cm.

Network visualization using Force Atlas 2 layout algorithm opposed modules characterizing hadal sites on the bottom right of the figure (modules 2, 3 and 4), and the other modules on the top left (Fig. 19A). The node with highest betweenness centrality at the interface between these two big clusters was classified as belonging to phylum Asgardarchaeota but was not assigned further. A blast search yielded two identical sequences of uncultured archaeon from other benthic studies of cold seeps and asphaltic hydrocarbon emissions (Knittel et al., 2005).

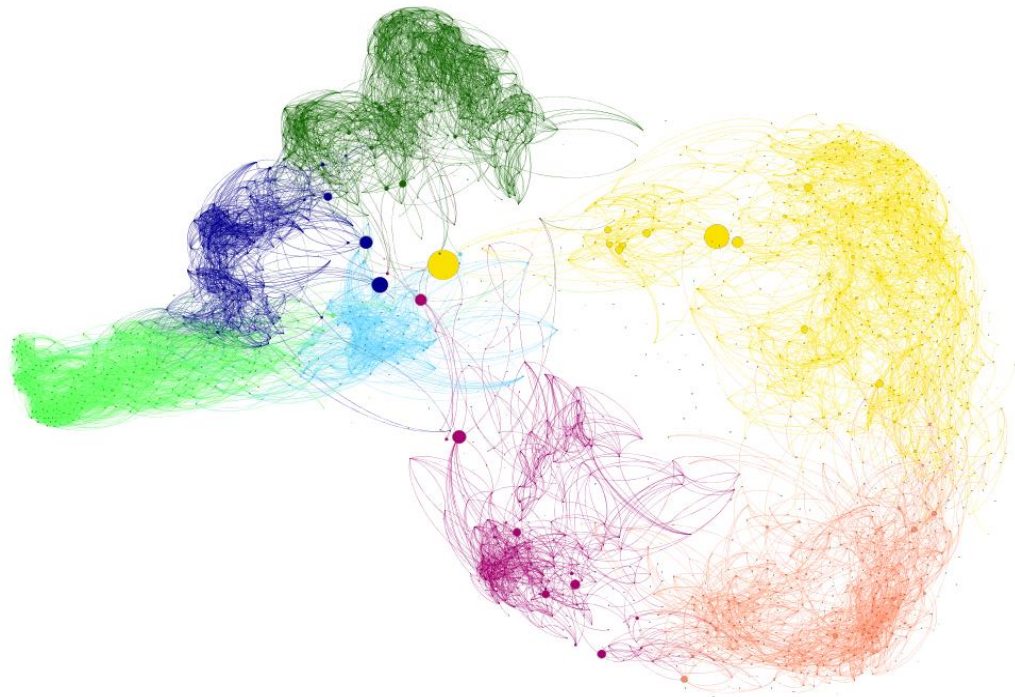
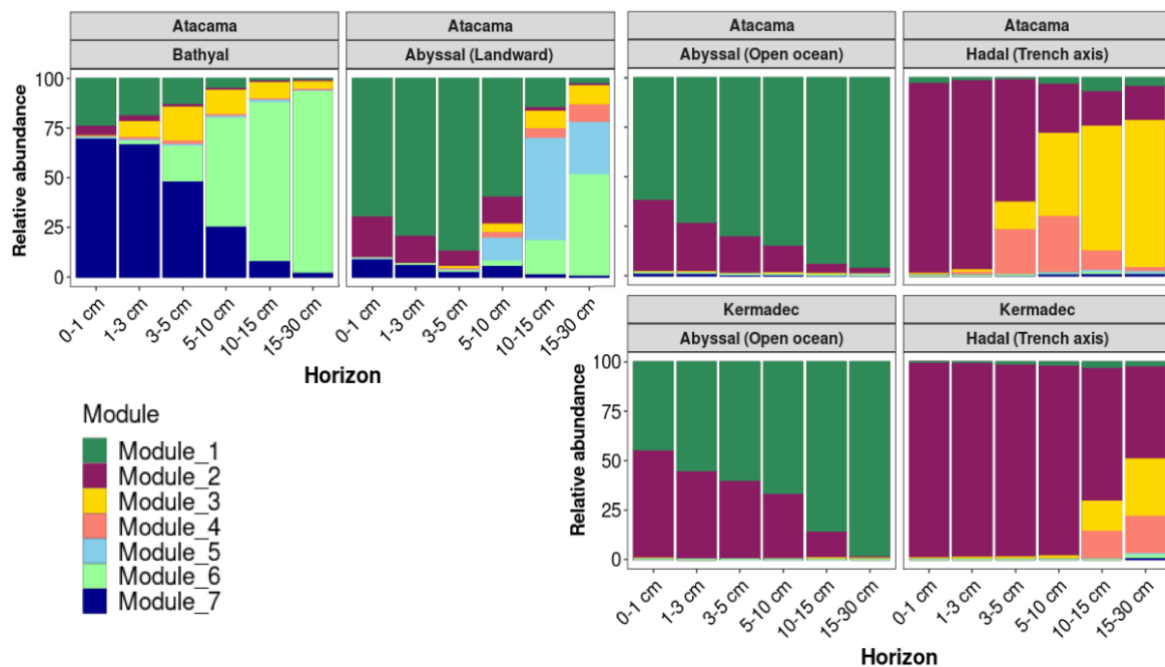
A**B**

Figure 19: Co-occurrence network and modularity analysis **(A)** Visualization of the co-occurrence network of Archaea computed with Spiec-easi, using Force Atlas 2 layout algorithm. Only edges with a weight over 0.5 are represented. Nodes and edges are colored according to the module they were assigned to, and node size illustrates betweenness centrality. **(B)** Distribution profiles of the archaeal modules in samples organized by trench,

CHAPTER 3

zone and increasing horizon depth. The relative abundance illustrated is based only on the ASVs used to compute the network.

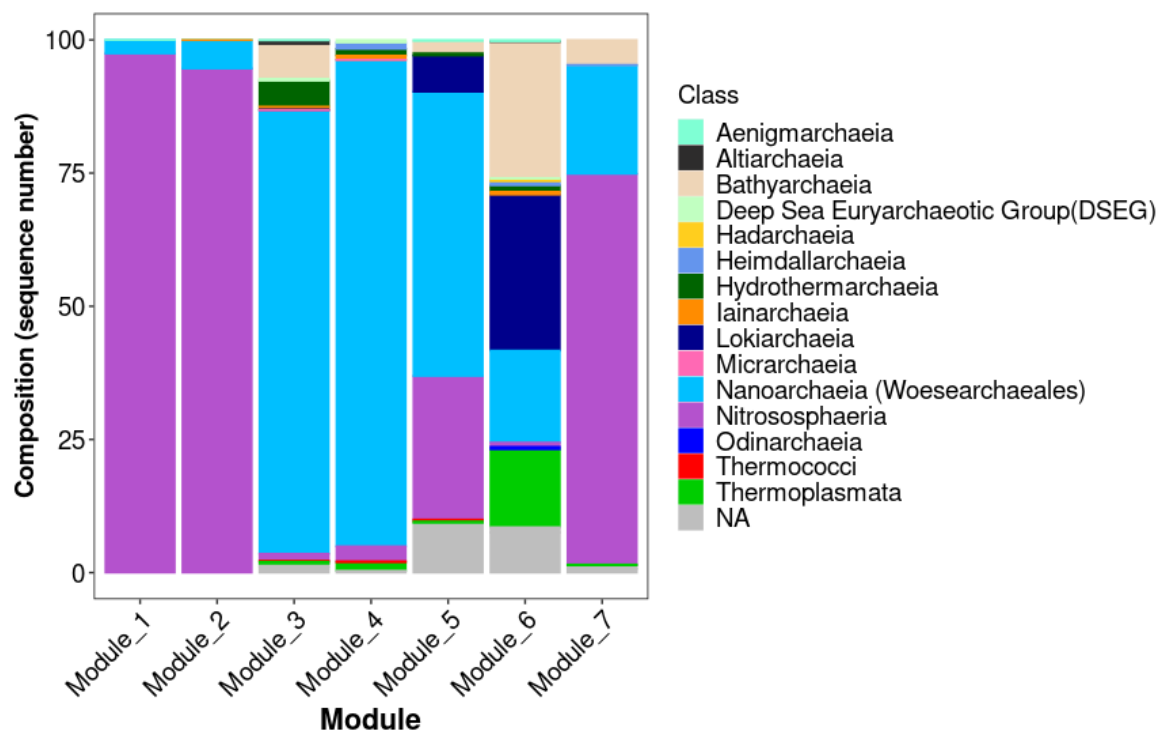


Figure 20: Taxonomic composition of the modules at class level. Relative proportion of each class is computed ASV sequence numbers.

4. Putative associations of Woesearchaeales

Based on the prevalence of Woesearchaeales in the deep horizons of hadal sites, we visualized the sub-networks corresponding to modules 3 and 4, restricting them to strong correlations (edges of weight superior to 0.7) (Fig. 21). In both cases, most correlations were found between Woesearchaeales.

In module 3, there were links between Woesearchaeales and nodes belonging to a number of classes: Bathyarchaeia, Hydrothermarchaeia, Nitrososphaeria, Lokiarchaeia and Altiarchaeia. Three of the unassigned nodes (244, 738, 2047) belonged to phylum Asgardarchaeota, and the last one (2155) was only classified at domain level. The closest relative according to a blast search was an uncultured archaeon, 81.9% similar, identified in a study of geothermal

CHAPTER 3

springs (Kormas et al., 2009). In module 4, edges linked nodes belonging to Woesearchaeales with Hydrothermarchaeia and Thermoplasmata members.

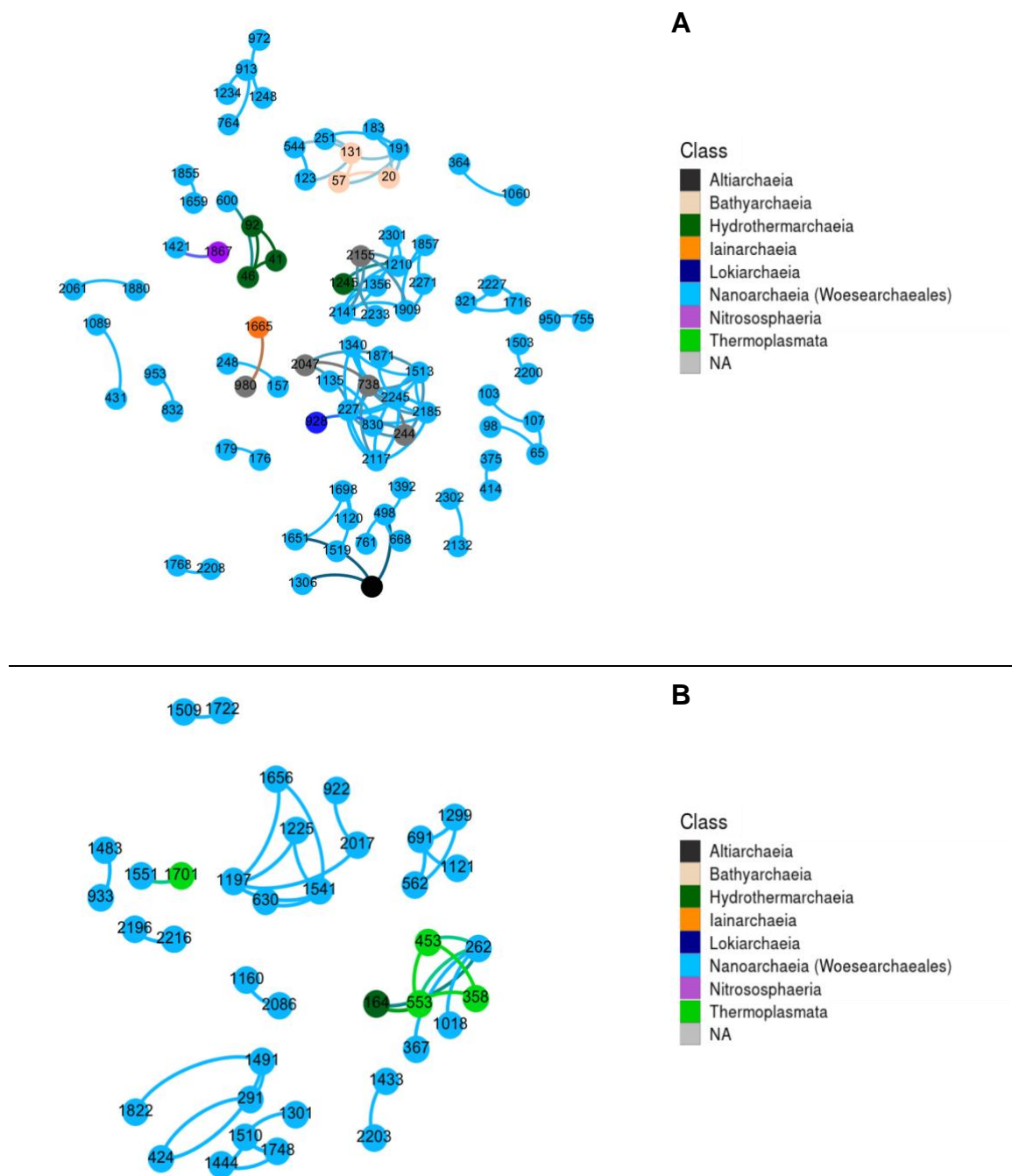


Figure 21: Visualization of the edges of weight over 0.7 in modules 3 (**A**) and 4 (**B**). The Open Ord layout algorithm was used to better distinguish clusters. Nodes and edges are colored according to taxonomy of the nodes (class level, based on SILVA 138).

Discussion

1. Influence of habitat, sediment depth and trench of origin on archaeal benthic communities

In this study, we investigated archaeal communities of abyssal and hadal benthic sediments using metabarcoding data generated from standardized samples from 11 sites of the Kermadec and Atacama trench regions. The PCR primers used for this analysis were previously evaluated against metagenomic data to ensure appropriate coverage of the diversity specific to these habitats.

In agreement with previous studies (Peoples et al., 2019; Vuillemin et al., 2019; Hiraoka et al., 2020; Hoshino et al., 2020; Kerou et al., 2021), the most abundant clades in our samples were Nitrososphaeria and Woesearchaeales (Fig. 18). The overall ratio of archaeal to bacterial sequences was higher in abyssal samples from sites on the oceanward side of the trenches (Fig. 17), but did not exceed 37.5% of sequences, suggesting that if Archaea dominate seafloor environments (Biddle et al., 2006; Lipp et al., 2008; Vuillemin et al., 2019), the shift in relative abundance happens deeper than 30 cm.

Co-occurrence networks can be used to explore spatial and ecological associations in complex community datasets (reviewed in Delmas et al., 2019; Espinoza et al., 2020). Here we conducted a co-occurrence and modularity analysis to identify groups of microorganisms with possible metabolic interactions.

Distribution of the seven modules identified (Fig. 19B) strongly resembled taxonomic profiles (Fig. 18) in that both showed patterns linked with sediment depth and habitat (bathyal, abyssal or hadal sediments). This pattern of community change with sediment depth is commonly observed in surface sediments and is expected to be linked with early diagenesis and the succession of terminal electron acceptors used in organic matter remineralization (Durbin and

CHAPTER 3

Teske, 2011). The rate of remineralization is directly influenced by organic matter input, both in terms of quantity and lability. This influx in substrate sinking through the water column is impacted by surface primary productivity, ocean depth and topographic layout. Indeed, hadal trenches have been shown to accumulate organic matter faster due to funneling effects and seismic activity resulting in landslides (Oguri et al., 2013; Turnewitsch et al., 2014).

This hadal increase in organic matter input and resulting higher rate of microbial activity has been assessed by Glud et al. (2021) at the same sites considered in this study through measurement of oxygen penetration depth. It was measured between 2.6 and 4.1 cm for hadal sites of the Atacama trench, and at 11.5 for the hadal site of the Kermadec trench considered here. At open ocean abyssal sites, oxygen penetrated deeper than 20 cm, however at the abyssal landward site (A9), it was depleted below 6.2 cm.

These thresholds of oxygen penetration and resulting geochemical zones have been shown to impact overall microbial community composition of these sediments by Schaubberger et al. (2021b). Here we see that they are also reflected in the shift towards dominance of Woesearchaeales in deeper sediments of hadal sites and landward abyssal site A9 (Fig. 18), and in the patterns of module distribution at these sites as well.

Interestingly, though two modules were specific to hadal sites, there was no trench-specific module. Indeed, module distribution profiles were remarkably similar between trenches, in the hadal axis as well as at the adjacent abyssal sites. This seems to indicate that the signal of archaeal community structure linked with the distinct environmental conditions encountered in abyssal and hadal environments is stronger than possible geographic isolation, at least in what concerns the most abundant ASVs. This limited endemism of trench communities was also observed on the total microbial community by Schaubberger et al. (2021b), and between the Mariana and Kermadec trenches by Peoples et al. (2019).

2. Distribution and modules of Nitrososphaeria

As expected from previous studies (Peoples et al., 2019; Vuillemin et al., 2019; Hiraoka et al., 2020; Hoshino et al., 2020; Kerou et al., 2021), archaeal communities of the first layers of sediments from all sites were dominated by Nitrososphaeria. While they stayed prevalent throughout the horizons sampled from open ocean abyssal sites, their relative abundance gradually decreased with sediment depth at all other sites (Fig. 18). As stated above, this pattern matched the oxygen penetration depths assessed by Glud et al. (2021). Given that most of the Nitrososphaeria ASVs detected in this dataset were assigned to family *Nitrosopumilaceae*, considered to be aerobic oxidizers of ammonia (Könneke et al., 2005; Walker et al., 2010; Stahl and de la Torre, 2012), this differential abundance is congruent with an involvement of at least part of the ASVs detected here in nitrification in the oxic zone.

Three of the seven modules identified from our co-occurrence network, modules 1, 2 and 7, were mostly composed of Nitrososphaeria nodes, with module 7 being specific to the Atacama bathyal site, module 1 detected in all abyssal sites, and module 2 mostly present in hadal and open ocean abyssal sites (Fig. 19B).

The abyssal Nitrososphaeria module (module 1) presented both ASVs abundant in surface horizons and ASVs rising to prominence with increasing sediment depth. This clear distinction could be due to subgroups of Nitrososphaeria (Sintes et al., 2013; Nunoura et al., 2016), though further assignment of these subgroups should rely on marker genes such as the *amoA* gene (Alves et al., 2018). Interestingly, this module was detected in all abyssal sites, even when separated by the Atacama trench or high geographic distances, but it was mostly absent in samples of the hadal sites (Fig. 19B).

Indeed, Nitrososphaeria members shared between hadal and abyssal sites were grouped in module 2, though they exhibited the same pattern of decrease with sediment depth as some of the ASVs found in module 1. This clear distinction between “abyssal” and “hadal” populations could be explained by a number of processes, including selection due to

CHAPTER 3

environmental conditions specific to the trenches (hydrostatic pressure, type of organic matter content), or lack of connectivity between deep hadal trenches and surrounding abyssal plains.

3. Putative associations of Woesearchaeales

Woesearchaeales, also referred to as *Ca. Woesearchaeota*, belong to the DPANN superphylum, have reduced genomes and are possible episymbionts of other Archaea (Dombrowski et al., 2019). We investigated the strong correlations involving ASVs assigned to this lineage in our dataset to assess the possibility of interactions with other Archaea. We focused on the hadal zone because continental plate sites were single representatives of their environmental conditions, which could influence the patterns of association.

Unlike a previous study that showed a potential syntrophic relationship between Woesearchaeales and methanogenic archaea in anoxic inland ecosystems (Liu et al., 2018), we did not detect a strong signal of specific association here. On the contrary, sub-networks showed a high diversity of potential archaeal partners for Woesearchaeales.

There was a notable difference in putative partners depending on the module considered, possibly reflecting different ecological niches and subclades. These putative niches were also visible in the distribution patterns of these modules in the Atacama trench sites: module 4 showed first an increase in relative abundance between 3 and 10 cm, before decreasing and almost disappearing in the deepest horizon, which were dominated by module 3 (Fig. 19B).

This pattern could be linked with geochemical zonation. Below the oxic zone, the following most favorable terminal electron acceptor is nitrate. Nitrate penetration depth has been estimated between 6 and 8 cm at Atacama hadal sites, and 15 cm at Kermadec hadal site K6 (Schauberger et al., 2021b, supp). However, the vertical resolution of our dataset is not high enough to be able to observe a close match between distribution pattern and geochemical profile. In any case, module 3 did not seem affected by the depletion of nitrate in Atacama

CHAPTER 3

sites (Fig. 19B) and became abundant below its depletion threshold, where onboard incubations documented active sulfate reduction (Glud et al., 2021).

Woesearchaeales have been shown to be very diverse (Liu et al., 2021) and possibly involved in diverse biogeochemical cycles: nitrogen cycling in wastewater treatment plants (Liu et al., 2021), sulfur cycle at hydrothermal vents (Cai et al., 2021) and potentially methanogenesis through their hypothesized symbiosis with methanogens (Liu et al., 2018). This wide variety of possible metabolism and thus association is not reflected in the taxonomy assigned to members of this lineage, limiting the conclusions to be drawn from these results without functional or experimental data. It has also been observed that Woesearchaeales abundances in high-altitude lakes is positively correlated with phylogenetic diversity of bacterial communities (Ortiz-Alvarez and Casamayor, 2016), and that Woesearchaeales tend to encode for proteins capable of binding and degrading bacterial cell walls (Castelle et al., 2021), suggesting that subsequent studies of Woesearchaeales metabolic associations should account for Bacteria.

Overall, these results highlight the ubiquitousness of Woesearchaeales in the anoxic part of the sediment cores characterized here, and provide spatial and taxonomic targets for further investigation of their putative associations.

Conclusion and perspectives

In summary, this study confirmed the domination of hadal and abyssal archaeal communities by Nitrososphaeria and Woesearchaeales. As reported for the general microbial community (Schauberger et al., 2021b), distribution of these lineages was tightly linked with both habitat and sediment depth. Co-occurrence network analysis highlighted the differences between abyssal and hadal communities, while also showing strong resemblance in module composition between geographically distant sites of the same habitat, underlining the environmental influence on community structure.

The abyssal and hadal Nitrososphaeria modules showed patterns of distribution possibly linked with subclades of AOA that will be investigated further in the next chapter using marker genes and metagenomic data. Finally, exploration of the strong correlations between Woesearchaeales ASVs and other Archaea did not show patterns of specific syntrophic relations, but provided targets for further examination of possible metabolic associations. Complementary information could be obtained through reconstruction of genomes from metagenomic data and inference of metabolic dependencies, and experimental visualization of physical links with CARD-FISH microscopy, for which samples were prepared onboard alongside the dataset analyzed here.

Supplementary figures

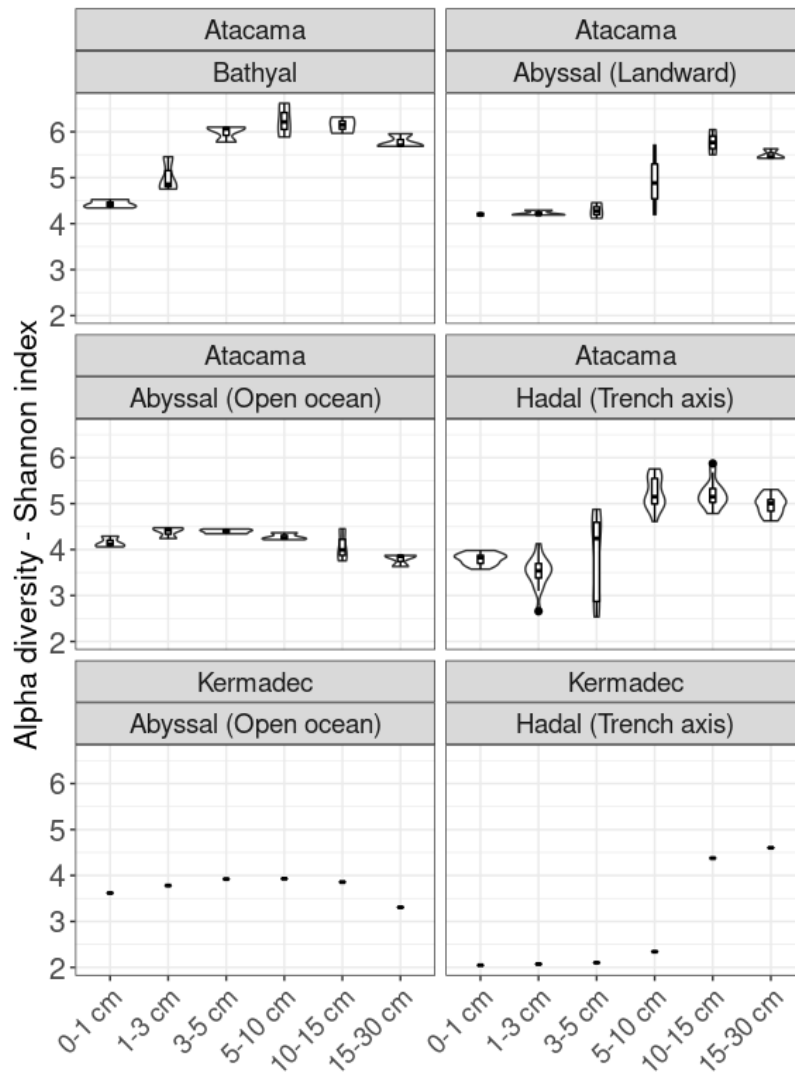


Figure S12: Archaeal alpha diversity estimates with Shannon index organized by trench, zone, and horizon depth.

CHAPTER 3

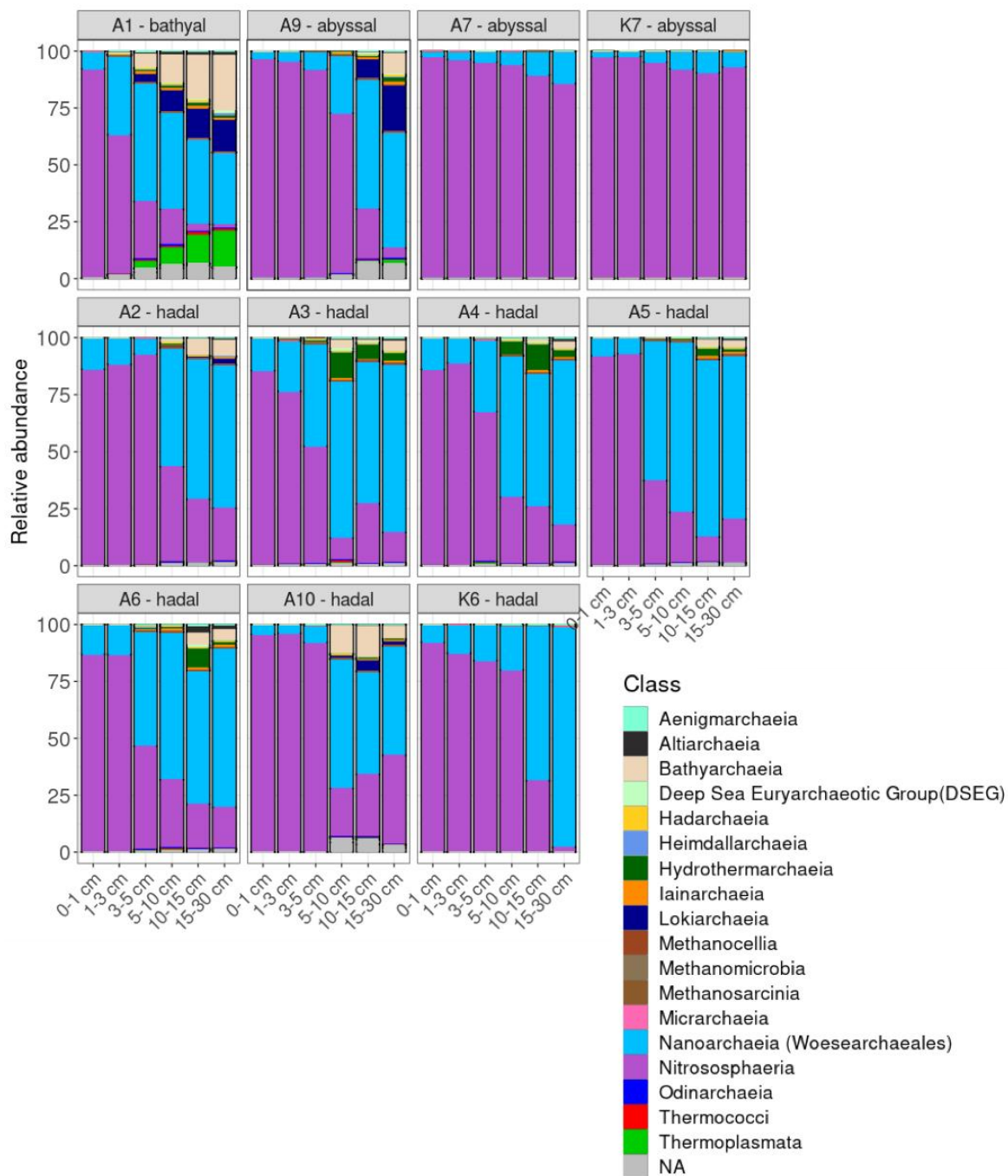


Figure S13: Archaeal taxonomic profiles at class level, grouped by sampling site, and horizon depth. Taxonomic assignment was based on SILVA 138.

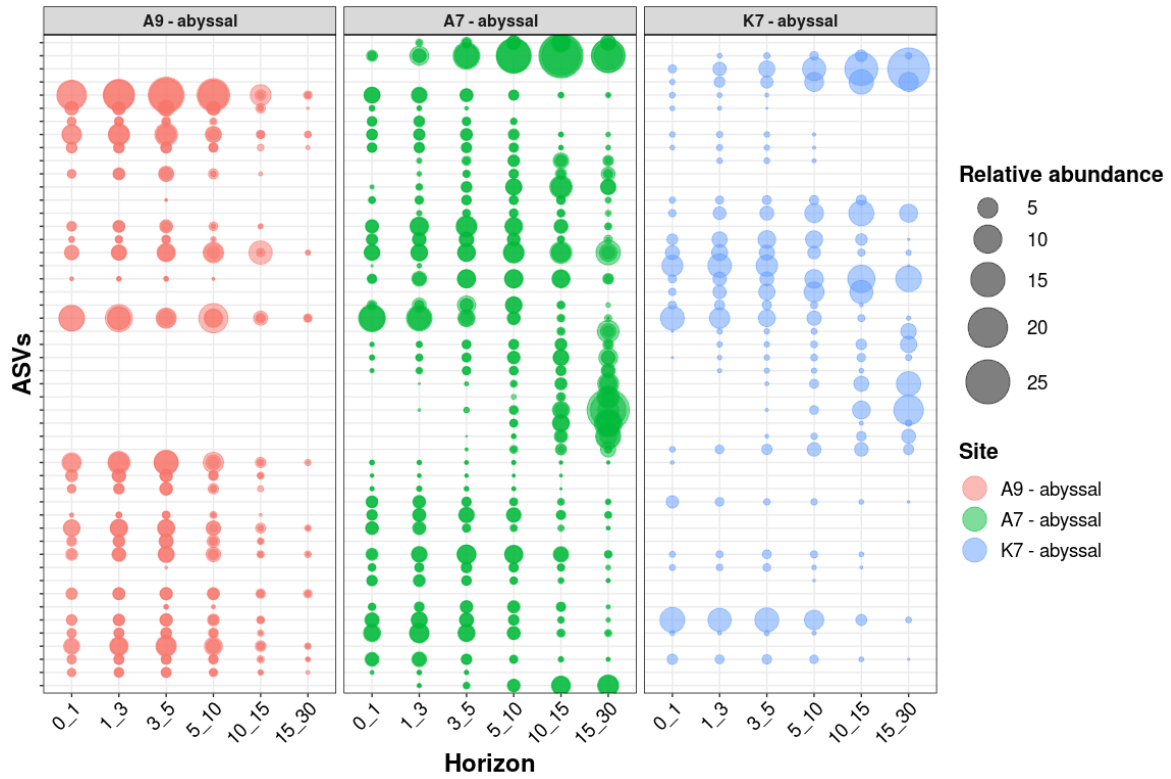


Figure S14: Relative abundance of the 50 most abundant nodes of module 1 in abyssal sites (A9, A7, K7). Each horizontal line represents an ASV, and color refers to site. All ASVs belonged to class Nitrososphaeria.

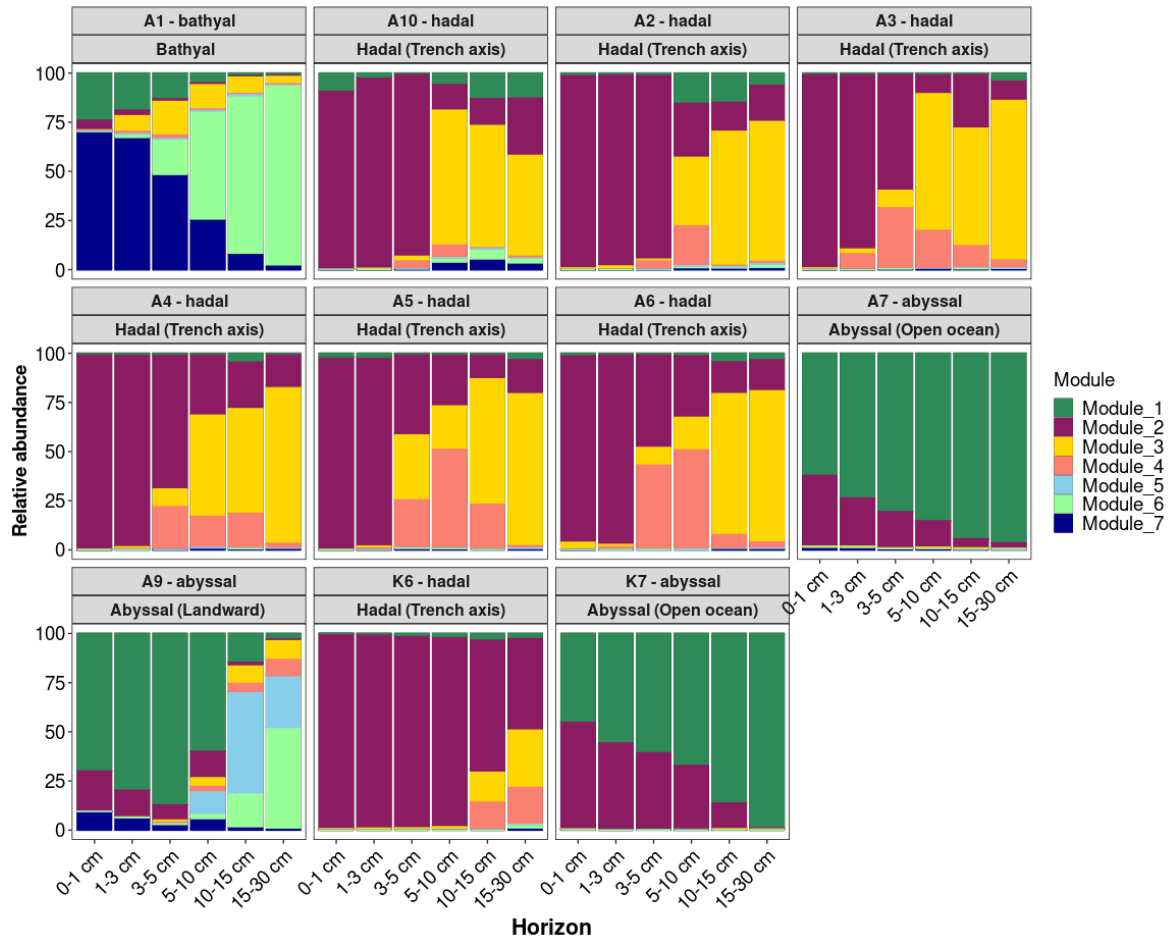


Figure S15: Distribution profiles of the archaeal modules in samples organized by site and increasing horizon depth. The relative abundance illustrated is based only on the ASVs used to compute the network.

CHAPTER 4

Clade distribution and genomic
variation of ammonia oxidizing
Archaea in abyssal and hadal
surface sediments

**Clade distribution and genomic variation of ammonia oxidizing
Archaea in abyssal and hadal surface sediments**

Blandine Trouche¹, Feriel Bouderkka¹, Clemens Schauburger², Jean-Christophe Auguet³,
Caroline Belser⁴, Julie Poulain⁴, Bo Thamdrup², Patrick Wincker⁴, Ronnie N. Glud², Sophie
Arnaud-Haond³ and Loïs Maignien^{1,5}

¹ Univ Brest, CNRS, IFREMER, Microbiology of Extreme Environments Laboratory (LM2E), F-
29280 Plouzané, France

² Hadal & Nordcee, Department of Biology, University of Southern Denmark, Odense,
Denmark

³ MARBEC, Univ Montpellier, Ifremer, IRD, CNRS, Sète, France

⁴ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Évry,
Université Paris-Saclay, 91057 Evry, France

⁵ Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular
Biology and Evolution, Woods Hole, MA, United States

Draft paper

Résumé de l'article en français

Les Thaumarchées (Nitrososphaeria dans la dernière version de la base de données SILVA) sont les micro-organismes les plus abondants dans les eaux océaniques profondes et les sédiments oxiqes de subsurface. Les Archées capables de réaliser l'oxydation de l'ammoniac (AOA) sont des membres importants de ce groupe qui jouent un rôle crucial dans les cycles biogéochimiques de l'azote et du carbone dans les sédiments benthiques, à l'interface entre écosystèmes pélagiques et sédiments profonds. Leur distribution a été étudiées dans les eaux hadopélagiques et les sédiments abyssaux, mais peu d'informations sont disponibles sur la connectivité entre populations présentes dans les sédiments de surface abyssaux et hadaux. Dans ce chapitre nous nous sommes intéressés à la distribution des gènes *amoA* spécifiques des différents clades d'AOA, extraits à partir de 56 métagénomomes issus de 6 sites des fosses Atacama et Kermadec. En outre, nous avons reconstruit des génomes partiels (MAGs) affiliés à six de ces clades et une lignée parente énigmatique n'oxydant probablement pas l'ammoniac. Pour les quatre clades les plus abondants, *amoA*-NP-gamma-2.1, gamma-2.2, theta et delta, nous avons mis en avant des profils de variabilité génomique liés à des différences de niche écologique, pouvant résulter de pressions de sélection opposées. Les *amoA*-NP-gamma, caractéristiques des eaux profondes, pourraient ainsi être bien adaptées à leur environnement et être soumise à une sélection purifiante, tandis que les clades *amoA*-NP-theta et delta, détectés dans les couches plus profondes de sédiments abyssaux, pourraient être sujets à une sélection positive, un scénario possible d'adaptation aux environnements très limités en énergie que sont les sédiments de subsurface.

Abstract

Thaumarchaeota (Nitrososphaeria in the latest version of the SILVA database) is the most abundant archaeal lineage in deep ocean waters and oxic subsurface sediments. Ammonia oxidizing archaea are important members of this group that play a crucial role in the biogeochemical cycling of nitrogen and carbon in benthic sediments, at the interface between pelagic and deep sedimentary ecosystems. Though their distribution has been studied in hadopelagic waters and abyssal sediments, information is lacking on the comparability of populations between hadal and abyssal surface sediments. Here, we investigated the clade-level distribution of *amoA* genes extracted from 56 metagenomes in 6 sites from the Atacama and Kermadec trench. Additionally, we reconstructed MAGs from 6 AOA clades and an enigmatic non ammonia oxidizing sister lineage. Focusing on the four most abundant clades, *amoA*-NP-gamma-2.1, gamma-2.2, theta and delta, we highlighted different patterns of genomic variability linked with differences in ecological niche, that could be the result of opposing selective pressures. Deep-ocean dwelling *amoA*-NP-gamma populations could be well adapted to their environment and thus be under purifying selection, while *amoA*-NP-theta and delta clades, typical of deeper sediment layers, could be experiencing positive selection as an adaptive scenario to the energy-limited subsurface environment.

Introduction

The existence of ammonia oxidizing archaea (AOA) was first evidenced by the isolation of *Nitrosopumilus maritimus* SCM1 from a marine tropical fish tank (Könneke et al., 2005). Since this discovery, their high abundance has been repeatedly reported in marine environments where they appear to be crucial contributors to the nitrogen and carbon cycles (Francis et al., 2005). Similarly to ammonia oxidizing bacteria (AOB), they are capable of chemolithoautotrophic life by aerobically oxidizing ammonia to nitrite, the first step in the nitrification process (Hooper et al., 1997; Könneke et al., 2005).

AOA are found throughout the water column as well as in deep sea sediments, where they are usually more abundant than AOB, due to their greater affinity for ammonia and possibly copper (Stahl and de la Torre, 2012; Shafiee et al., 2021). Studies of their distribution in these environments have shown the presence of distinct clades (Sintes et al., 2013; Alves et al., 2018), most of them part of order Nitrospumilales (NP), reclassified as family *Nitrosopumilaceae* in the most recent version of the SILVA database (v138, Quast et al., 2013), inherited from the Genome Taxonomy Database (GTDB) efforts to produce a rank-normalized archaeal taxonomy (Parks et al., 2020; Rinke et al., 2021). They are usually studied using either 16S rRNA or *amoA* (ammonia monooxygenase subunit A) as a marker gene.

Based on the clade definition by Alves et al. (2018) with the *amoA* marker gene, hadopelagic waters of North Pacific trench systems are populated by clades *amoA*-NP-alpha and *amoA*-NP-gamma (Nunoura et al., 2016, 2018; Wang et al., 2018; Zhong et al., 2020). The *amoA*-NP-gamma clade seems to be omnipresent in oceans, irrespective of water depth, and in fact *Nitrosopumilus maritimus* SCM1 falls in this clade. *amoA*-NP-alpha populations seem most highly abundant in bathy- and abyssopelagic environments (1000 to 6000 m) (Nunoura et al., 2016; Zhao et al., 2020; Zhong et al., 2020). Contrastingly, the dominant clades in sediments of abyssal plains are *amoA*-NP-theta and *amoA*-NP-delta (Zhao et al., 2020; Kerou et al.,

CHAPTER 4

2021), however data is scarce regarding hadal sedimentary clades, especially in settings allowing for the comparison of their nature and occurrence with those in adjacent abyssal environments.

All AOA are part of class Nitrososphaeria (proposed classification of phylum Thaumarchaeota in the GTDB and SILVA 138 databases), though not all Nitrososphaeria can oxidize ammonia (Aylward and Santoro, 2020). Nitrososphaeria is the most abundant archaeal lineage in hadopelagic waters and oxic seafloor sediments (Peoples et al., 2018; Vuillemin et al., 2019; Hiraoka et al., 2020; Hoshino et al., 2020). In the previous chapter, Nitrososphaeria were the dominant archaeal class detected in the surface sediments of the Atacama and Kermadec trench, and persisted in deeper horizons of abyssal sites, seemingly linked with a higher oxygen penetration depth (Glud et al., 2021; Schaubberger et al., 2021b). In addition, a co-occurrence network analysis revealed a segregation of Nitrososphaeria into three sub-networks (modules) with sediment depth and habitat (hadal or abyssal). In particular, we identified a distinctly abyssal module, and a “hadal” module composed of all hadal Nitrososphaeria, with some members also detected at abyssal depth. These two modules thus highlighted differing module memberships for Nitrososphaeria ASVs present at abyssal depths and exhibiting a similar decrease in abundance with sediment depth.

However, while metabarcoding data provided a first insight into the distribution of Nitrososphaeria in these sediments, limitations associated with short barcode sequences prevented a more thorough study of closely related variant distribution, and did not allow linking marker genes with the functional repertoire of the associated genomes.

To address these questions, we thus used 56 metagenomes generated from the same DNA extracts as the metabarcoding dataset to further study the distribution of Nitrososphaeria in these sediments at higher taxonomic resolution. We first applied a gene-based classification by extracting *amoA* gene sequences from metagenome assemblies and characterized the clade diversity present in these samples. Furthermore, we reconstructed genomes affiliated

CHAPTER 4

with Nitrososphaeria, and studied the genomic variability of these metagenome-assembled genomes (MAGs), with the aim of getting insights into the niche separation of these populations.

Material & Methods

1. Sampling sites, slicing scheme and DNA extraction

Samples for this study were collected during two cruises to the Atacama and Kermadec trenches in the South Pacific Ocean, as previously described in chapter 1.4 and chapter 3. Here, as in chapter 1.4, we considered samples from two sites from the Kermadec trench, one site in the trench axis (K6, 9555 m) and one site on the adjacent abyssal plain (K7, 6080 m), and four sites from the Atacama trench, two trench axis sites (A3 and A10, 7915 and 7770 m) and two abyssal sites (A9, landward, 4050m and A7, oceanward, 5500 m) (Fig. 16).

Triplicate sediment cores were recovered for Atacama sites A3 and A7, and single cores for the other sites. All cores were sliced into standardized depth layers as follows: 0-1 cm, 1-3 cm, 3-5 cm, 5-10 cm, 10-15 cm and 15-30 cm.

DNA extractions were performed using 10g of sediment from each of these layers, and library preparation and sequencing were carried out at Génoscope (Evry, France) as described in the methods for chapter 1.4.

2. Assembly and binning

The quality filtration of the demultiplexed metagenomic raw reads was carried out with Illumina-Utils python scripts (Eren et al., 2013b) following recommendations by Minoche et al. (2011). Metagenomes were then split into ten co-assembly groups based on *de novo* comparison of the unassembled metagenomes using *k*-mer counts (Simka, Benoit et al., 2016). Composition of the co-assembly groups can be found in Table S3. Most of the following steps were performed with the help of the Snakemake workflows (Köster and Rahmann, 2012) available with Anvi'o (v7, Shaiber et al., 2020; Eren et al., 2021).

We co-assembled the samples using Megahit (v 1.1, Li et al., 2015) with preset meta-sensitive and minimum contig length of 1000 bp. Identification of Open Reading Frames (ORFs) in the

CHAPTER 4

contigs was run with Prodigal (Hyatt et al., 2010) and functional annotation obtained using KOfamscan (Aramaki et al., 2020) and the COG database (2020 release, Galperin et al., 2021).

After mapping of the short reads on the resulting contigs (Langmead et al., 2009; Danecek et al., 2021), automatic binning was performed with Concoct (Alneberg et al., 2014), restricting the number of bins to half the number of predicted bacterial genomes to prevent fragmentation errors. Archaeal bins were then inspected and refined manually twice using Anvi'o's interactive interface (Eren et al., 2015). Completeness and redundancy were estimated by Anvi'o based on single-copy core gene collections (Lee, 2019).

Reconstructed MAGs were dereplicated based on pyANI with a minimum alignment fraction of 0.5, and a similarity threshold of 0.95. They were then once again run through the mapping steps of the workflow to obtain final coverage values.

3. Reference genomes

To add context to our results, we included MAGs from other deep-sea studies to the phylogenetic placement sections of our analysis. In particular, we downloaded 4 MAGs reconstructed by Zhong et al. (2020) from the water column of the Mariana Trench (MTA1, MTA4, MTA5, MTA6) and 9 MAGs reconstructed by Kerou et al. (2021) from marine sediments at abyssal depths (NPMR_NP_delta_1 to 3, NPMR_NP_theta_1 to 5 and NPMR_NP_iota_1). They were run through the same steps as described above for gene calling and functional annotation.

4. Phylogenetic placement of *amoA* reconstructed genes

To identify the clades of ammonia-oxidizing archaea present in our metagenomes and MAGs, we extracted from our co-assemblies the sequences of genes annotated as ammonia monooxygenase subunit A (*amoA*) by KOfam. We dereplicated them using CD-Hit (Fu et al., 2012) with 100% identity and we also obtained the sequences for this gene from the reference MAGs detailed above. We used a blastn search against the non redundant NCBI nucleotide

CHAPTER 4

collection to confirm gene assignment and determine domain-level taxonomy (Zhang et al., 2000). We then aligned the sequences matching archaeal *amoA* genes using MAFFT with default parameters (v7.273, Katoh, 2002) and placed them in the reference tree by Alves et al. (2018) using EPA-ng (Barbera et al., 2018). We visualized this tree in R (v3.6.1) using packages *ggtree* and *treeio* (Yu et al., 2016; Wang et al., 2020).

We obtained the coverage information for the *amoA* genes present in our co-assemblies and visualized it in R with package *ggplot2* (v3.3.0, Wickham, 2016), filtering out coverage values for which detection was below 0.9 (probable non-specific mapping).

5. Taxonomic placement of MAGs

Taxonomic assignment of the MAGs was performed using GTDB-tk *classify_wf* workflow (Chaumeil et al., 2020) with the GTDB database (Parks et al., 2018, 2020).

We used the marker protein sequence alignment generated by the GTDB-tk *align* command, after masking of positions with over 0.5 gap frequency, to reconstruct a maximum likelihood tree in IQTREE (v2.0.3, Hoang et al., 2018; Minh et al., 2020) under the model WAG with 1000 ultrafast bootstrap replicates.

This method was used to generate a tree containing our MAGs, the reference MAGs and the GTDB representative sequences for order Nitrososphaerales (also referred to as Nitrosopumilales). The tree was rooted by choosing class Korarchaeia as an outgroup. We applied this approach again to generate a tree of only our MAGs. Given that not all MAGs reconstructed contained an *amoA* gene, when possible, *amoA* clade affiliation of MAGs was extrapolated from these phylogenomic trees.

6. Single nucleotide and single amino acid variant analyses

During the last mapping of the short reads on the contigs grouped into our MAGs, we performed single nucleotide variants (SNVs) and single amino acid variants (SAAVs) calling to investigate genomic variability using the *anvi_profile* function from Anvi'o, with flags `--skip-`

CHAPTER 4

SNV-profiling = false, and --profile-SCVs = true. We then chose representative MAGs based on their completeness and abundance from each well-distributed *amoA* clade (theta, delta, gamma) and used command *anvi-gen-variability-profile* with engine NT (nucleotide) and AA (amino acid) to compute tables listing all variable positions in samples where the MAG had a coverage over 10x. Additionally, we required reported positions to have a coverage above 10x in every sample considered and a minimum departure from consensus of 0.1 (total number of reads not matching the consensus divided by the total number of mapped reads).

We then introduced the resulting tables in R to compute additional information on sequence variability for each representative MAG. We calculated the number of SNVs by kbp in each sample by dividing the number of identified positions by the total length of the sequences considered. For each MAG we also obtained the ratio of SAAV to SNV in each sample, and for each gene. Finally, we linked variable positions and predicted function of the genes to visualize the variability of genes depending on function.

Results and discussion

1. Distribution of AOA clades in abyssal and hadal benthic sediments

In order to better constrain AOA distribution in these abyssal and hadal sediments, we first used the *amoA* marker genes retrieved from our metagenome assemblies and placed them in the reference tree proposed by Alves et al. (2018). In addition, we completed this tree with *amoA* sequences extracted from the 13 external abyssal and hadopelagic MAGs from previous studies (Zhong et al., 2020; Kerou et al., 2021). Out of the 166 sequences extracted from our co-assemblies based on KOfam functional assignment, 105 matched archaeal sequences from the NCBI non redundant nucleotide collection, 27 had bacterial matches, and 31 returned inconclusive results. The archaeal sequences matched a number of clades, but most were related to *amoA*-NP-gamma, *amoA*-NP-theta and *amoA*-NP-delta (Fig. 22) with 14 monophyletic sequences placed on the same *amoA*-NP-gamma-2.2 node. Two sequences were placed in the *amoA*-NP-alpha and *amoA*-NP-iota clades respectively (Alves et al., 2018; Kerou et al., 2021).

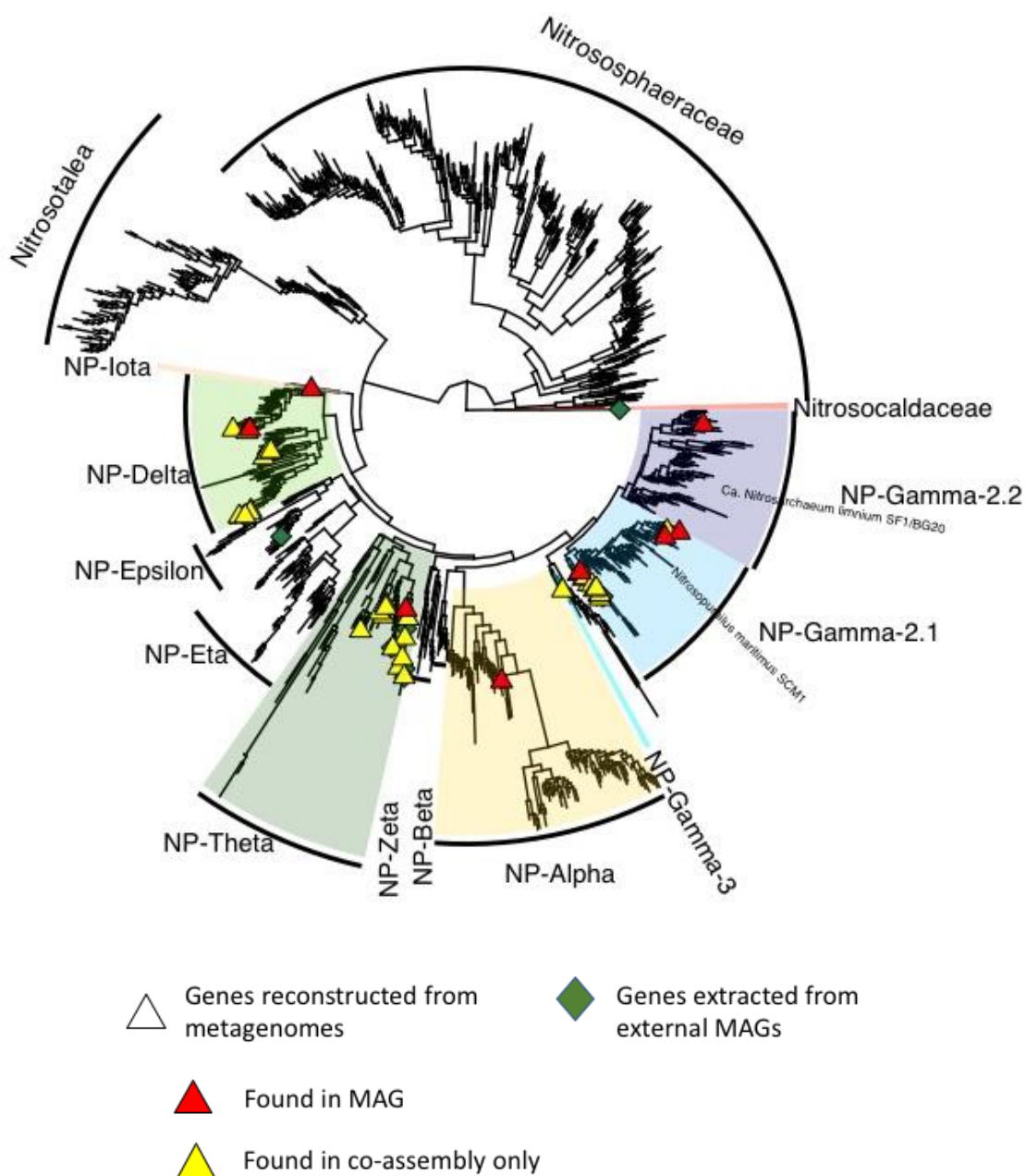


Figure 22: Global phylogenetic tree of *amoA* genes obtained from Alves et al. (2018), with placement of *amoA* genes identified from co-assemblies and extracted from reference MAGs. The shape of the points denotes the origin of the sequence, and the color of the triangles illustrates the status of the sequence (identified in a MAG of this study or only in the co-assembly results). Shading highlights *amoA* clades.

CHAPTER 4

Members of the *amoA*-NP-theta, delta and gamma clades dominated the samples as indicated by their higher coverage (Fig. 23).

amoA-NP-gamma genes were detected in the surface sediments of all sites (Fig. 23), however we found that they had distinct distribution patterns. For *amoA*-NP-gamma-2.1, 15 variants were shared between at least one abyssal and hadal site, and 10 variants were specific to abyssal sites. We also found several hadal-specific variants, 6 for the *amoA*-NP-gamma-2.1 clade including the variant dominating the K6 site, and 7 *amoA*-NP-gamma-2.2 variants.

Additionally, though all sequences matching the *amoA*-NP-gamma-2.2 clade were placed on the same node of the reference tree (Fig. 22), they apparently formed two distinct clusters with very similar sequences (Fig. S16B). The first relatively abundant cluster was shared between abyssal and hadal samples, and the second one was specific to the hadal zone.

The distinctness of the abundant sequences characteristic of the abyssal and hadal sites, particularly for clade NP-gamma-2.1, seemed to reflect the distribution patterns of the modules defined in chapter 3. These modules highlighted a partial split between abyssal and hadal Nitrososphaeria ASVs exhibiting a similar decrease in abundance with sediment depth. The results presented here (Fig. 23) show that this split is not explained by the presence of distinct clades with a similar pattern of abundance, but rather by similar variations in abundance for distinct abyssal and hadal members of the same clade. This suggests an influence of the habitat (abyssal *versus* hadal), that may have distinct contemporary or historical origins stemming from differences in hydrostatic pressure and organic matter input and quality, or stronger geographic isolation between the Atacama and the Kermadec trenches than between abyssal sites.

amoA-NP-gamma clades were the dominant clades detected in the hadopelagic waters of the Mariana trench (Nunoura et al., 2018; Zhong et al., 2020). Shallow waters are typically dominated by AOA associated with the cultivated genus *Nitrosopumilus*, a member of the *amoA*-NP-gamma-2.1 clade as well. However, based on phylogenomic placement in the reference tree, it seemed that benthic hadal members of this clade form a distinct cluster, most probably due to the very different set of conditions characterizing this environment (Fig. 22).

CHAPTER 4

amoA-NP-delta and *amoA*-NP-theta were abyssal clades, and in addition to the influence of water depth, we also observed distinct distribution patterns with sediment horizons. Members of the *amoA*-NP-theta clade increased in coverage with sediment depth and are most likely adapted to low oxygen environments. This could also explain their absence from hadal sites where the oxic zone is much shallower and the low oxygen niche could be absent (Glud et al., 2021). Conversely, members of *amoA*-NP-gamma-2.1 were more abundant in surface sediments and decreased with horizon depth, suggesting an adaptation to higher oxygen levels than their *amoA*-NP-theta counterparts. *amoA*-NP-gamma-2.2 genes had higher coverage at the Atacama sites and exhibited an intriguing distribution pattern in hadal sites of the Atacama trench, with relatively high coverage in top and bottom sediment layers, and no detection between 5 and 10 cm. It was found at low coverage in surface horizons of Kermadec sites (K6 and K7).

These patterns of distribution of the different clades are in agreement with previous studies that showed that *amoA*-NP-theta *Nitrosopumilaceae* dominate deep subsurface AOA communities, while *amoA*-NP-gammas tend to be more abundant in the upper sediment layers than in deeper sediment, and no striking variation is observed for the *amoA*-NP-delta (Vuillemin et al., 2019; Zhao et al., 2020; Kerou et al., 2021).

Finally, we also detected two rare clades that were represented by a unique *amoA* sequence in the metagenome assemblies, and were restricted to very specific sediment layers: surface layers of abyssal site A7 for *amoA*-NP-alpha, and deeper layers of the same site for *amoA*-NP-iota (data not shown). *amoA*-NP-alpha has been detected at high relative abundances in the deep water column, with a peak in abundance between 2000 and 4000 m (Zhong et al., 2020). *amoA*-NP-iota, similarly to NP-theta and delta, is a sediment-dwelling clade (Kerou et al., 2021).

CHAPTER 4

Comparing overall coverage values for the archaeal and bacterial *amoA* sequences extracted from the co-assemblies, we found that bacterial sequences made up only 5.5% of mapped reads. Though reconstruction of genes from metagenomes is a less reliable way than qPCR to assess the relative contribution of ammonia oxidizing bacteria and archaea to biogeochemical cycles, our results suggest that the benthic sediments considered here are dominated by AOA (Schleper, 2010; Wang et al., 2017).

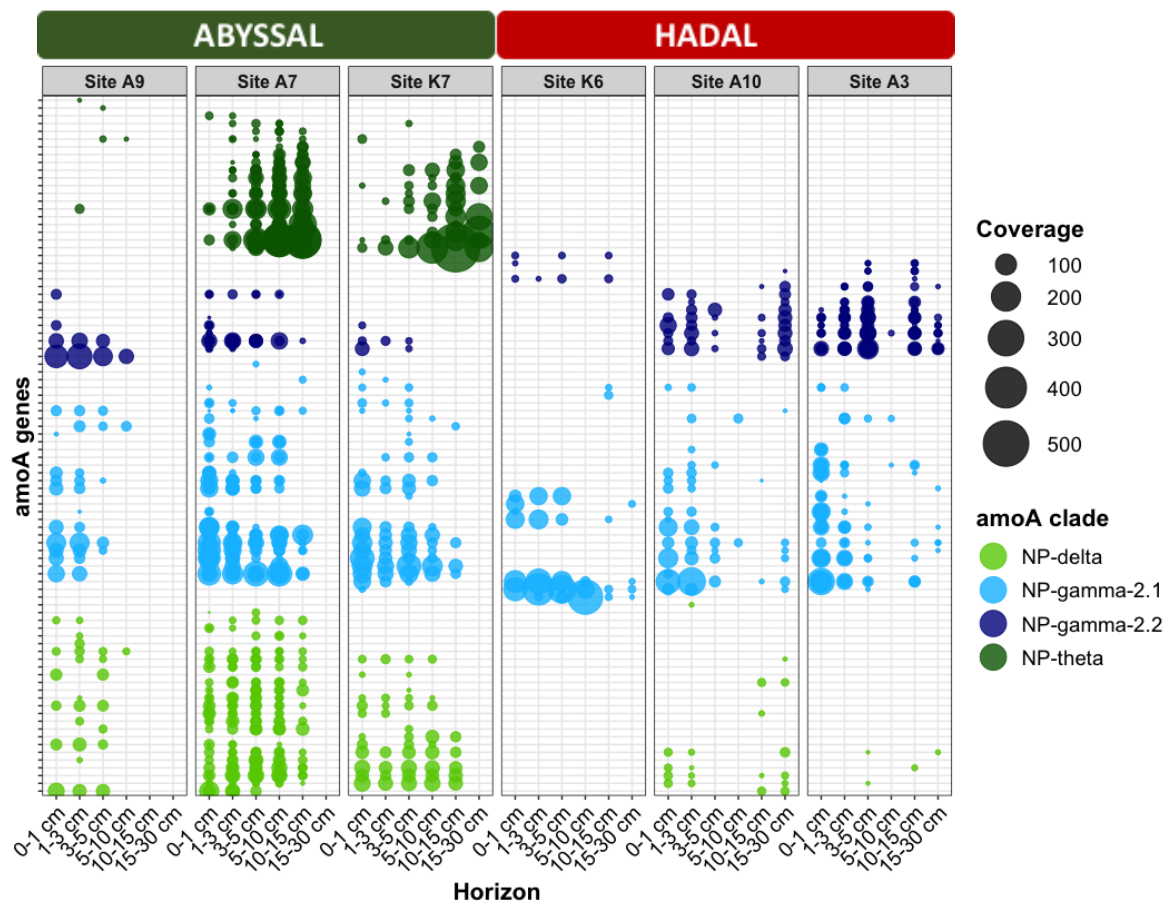


Figure 23: Coverage of *amoA* genes identified in co-assemblies (detection over 0.9). Point size is proportional to coverage in each sample, points are colored according to clades assigned previously by phylogenomic tree placement. Samples are ordered by site and horizon depth increases along the horizontal axis for each site.

2. Phylogenomic placement and distribution of MAGs affiliated to class Nitrososphaeria

To further examine distribution of AOA, we conducted contig binning from our co-assemblies to form metagenome assembled genomes (MAGs). After manual refinement of all bins affiliated with class Nitrososphaeria, we obtained a total of 53 Nitrososphaeria MAGs. Despite high sequencing depths, only one of these MAGs was of high quality, with completeness over 90% and redundancy under 5% (Fig. 25, Table S4). 12 other MAGs had completeness over 70% and redundancy under 7%. The remaining MAGs had lower completeness estimates but always redundancy under 10.53% (Table S4). We identified the *amoA* gene in 10 of our MAGs and overall, their repartition in the GTDB tree, as well as that of the reference MAGs, matched the *amoA* clade results (Fig. 24). Based on this observation, we inferred clade affiliation for the MAGs lacking the *amoA* gene, possibly because of lower completeness. The genome size of the 13 most complete MAGs varied between 0.98 and 1.4 Mbp, in accordance with previous reports of marine free-living Nitrososphaeria (Zhang et al., 2019). Mean GC content among MAGs identified as members of the Nitrosopumilaceae family was 33.4%. MAGs placed in genus *Nitrosopumilus* were not closely related to the cultivated or enriched strains, which were recovered from shallower environments (Könneke et al., 2005; Mosier et al., 2012; Park et al., 2012a, 2012b; Bayer et al., 2019). MAGs HAS_Bin_00039 and HKT_Bin_00022 also clustered apart from the representative *Nitrosoarchaeum* genomes, though they belonged to the same *amoA* clade.

10 MAGs from our study clustered together with the NP-delta MAGs reconstructed by Kerou et al. (2021) and one GTDB representative genome of genus *CSP1-1*. This representative genome was reconstructed from an aquifer sediment sample at 5 m depth (Hug et al., 2016b). Finally, 14 MAGs formed a cluster without any reference GTDB genome, but clustered with all Kerou et al. (2021) NP-theta MAGs.

CHAPTER 4

In addition, we also reconstructed five MAGs with GC content between 43.48 and 46.38% (Fig. 25, Table S4) that clustered together in the phylogenomic tree (Fig. 25), and were placed outside of *Nitrosopumilaceae* in the GTDB reference tree (Fig. 24). They are related to families *UBA141* and *UBA57*, knowledge of which exclusively relies on genomic data. *UBA141* in particular is based on only one MAG (GenBank accession number DAEN00000000.1), reconstructed by Parks et al. (2017) from a meta-analysis of database metagenomes. Its distribution and metabolic potential have thus not been studied and will need further characterization based on additional data generated in the current study. Members of the family *UBA57* were reconstructed by Aylward and Santoro (2020) from deep ocean water samples and described as a sister lineage to AOA with small genomes, a putatively heterotrophic lifestyle, and lacking the ability to oxidize ammonia. They were also found to be widely distributed in the ocean. Here, our divergent MAGs also did not possess the *amoA* gene but had very low coverage levels indicating low abundance in these abyssal and hadal sediments (Fig. 25).

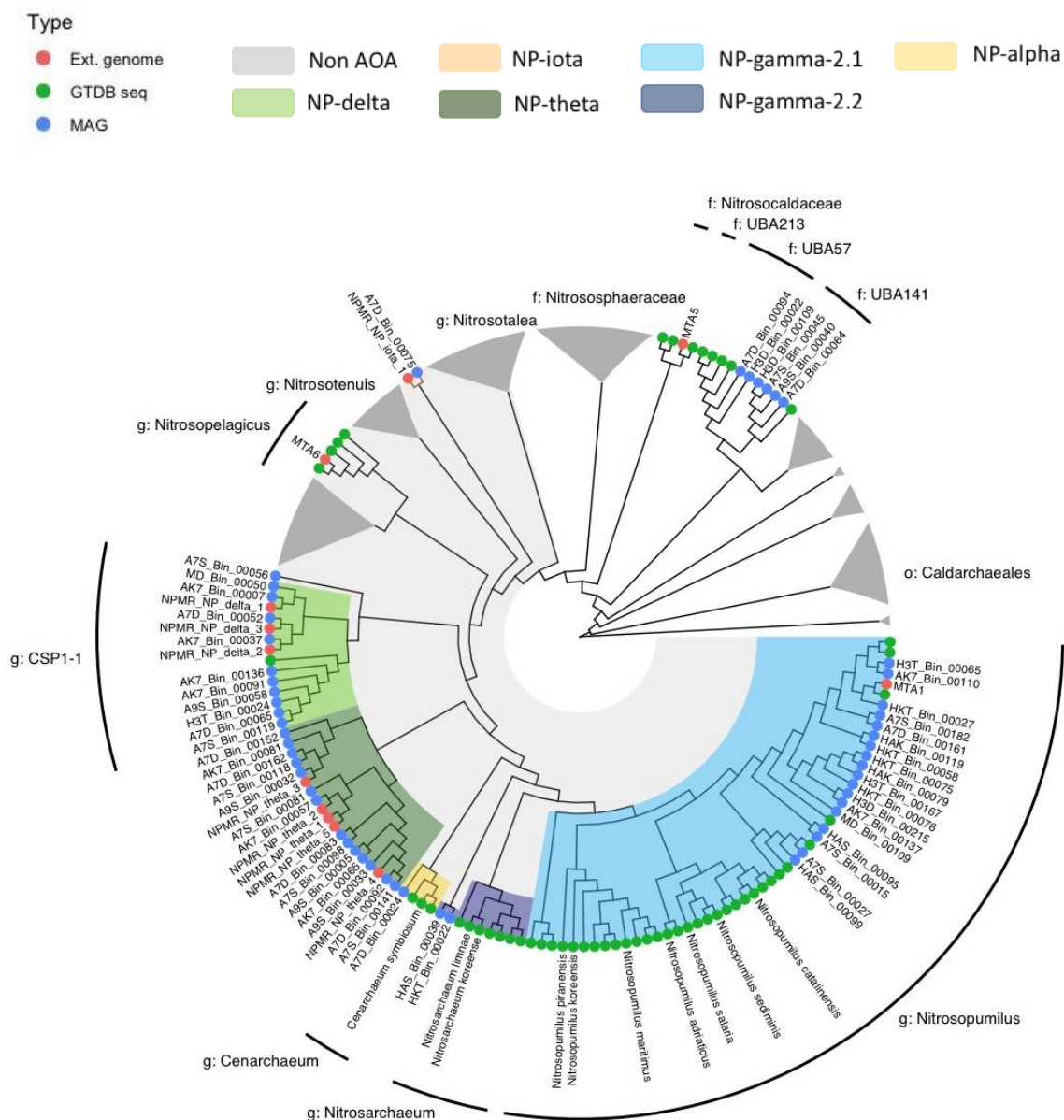


Figure 24: Phylogenomic tree of order Nitrososphaerales built from the GTDB database with addition of our MAGs and 13 reference MAGs from two other studies. This order in the GTDB database corresponds to class Incertae Sedis of the Thaumarchaeota according to NCBI taxonomy. The grey shading highlights family Nitrosopumilaceae and tip point color refers to the type of MAG/genome. Colored shading highlights *amoA*-NP-clades as defined in Fig. 22 and in Alves et al.(2018).

CHAPTER 4

In general, mean coverage of the MAGs in our samples was congruent with those of corresponding clades defined by *amoA* genes described in the previous section (Fig. 25). We observed a relatively high abundance of *amoA*-NP-delta MAGs in abyssal samples, except for the deeper samples of site A9. All *amoA*-NP-theta MAGs exhibited an overall expected pattern of increasing coverage with sediment depth, though some populations were fairly rare while others dominated the deep horizons of sites A7 and K7. *amoA*-NP-gamma-2.1 MAGs were detected in the surface horizons of all sites and *amoA*-NP-gamma-2.2 MAGs presented the same intriguing pattern as *amoA* genes of a drop in coverage between 5 and 10 cm.

In this study, we used MAG reconstruction with the aim of examining AOA genomic population structure and distribution (next section) as defined by patterns of nucleotide and amino acid variants. However, comparative genomics between closely related MAGs with distinct ecological distribution will be an obvious next step to conduct for the comprehension of evolutionary processes that have led to such a diversity and fine scale biogeography.

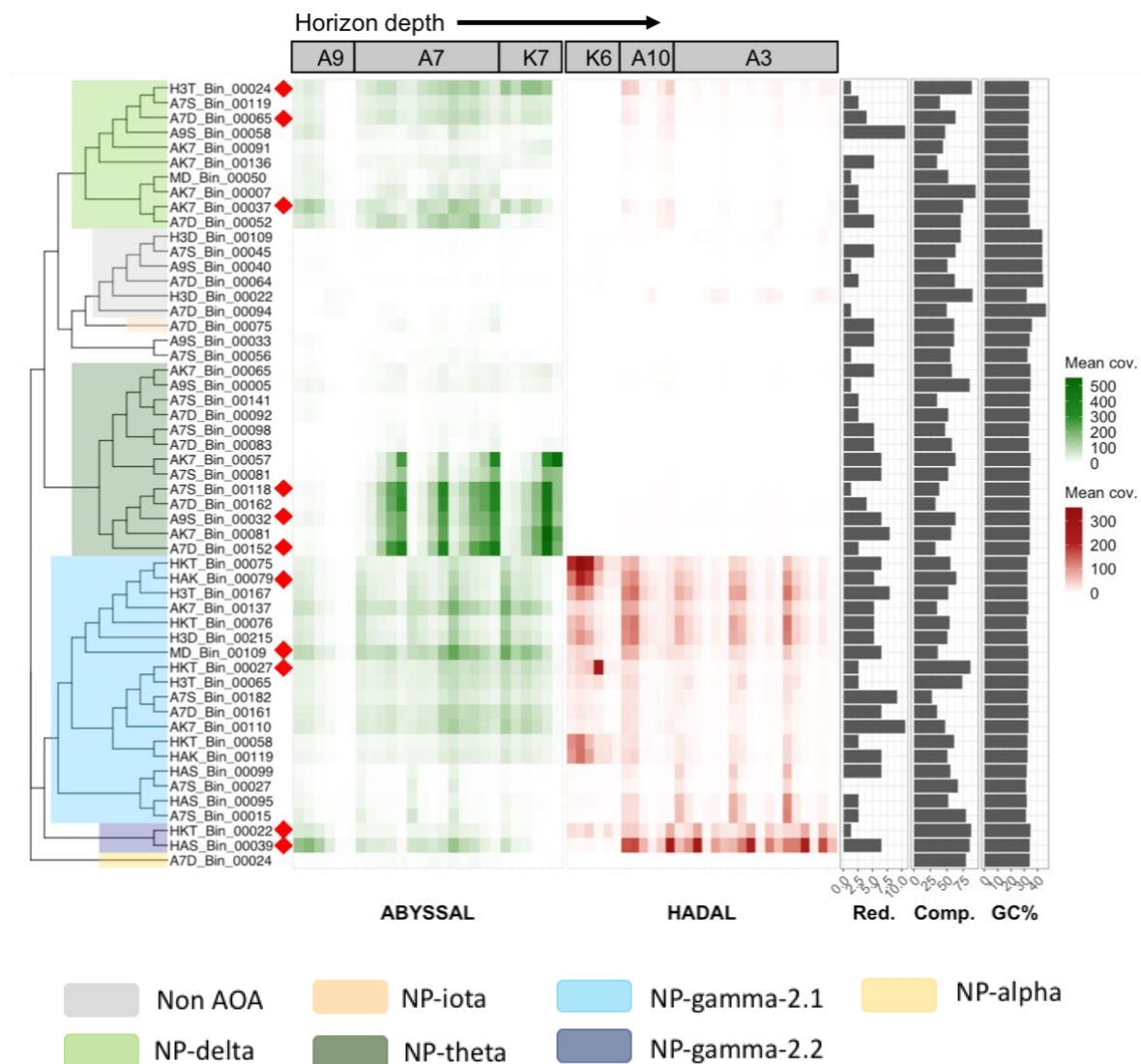


Figure 25: Mean coverage of Nitrososphaeria MAGs in our samples. Samples are ordered by site and increasing horizon depth, with the green-colored heatmap illustrating abyssal samples and the red one hadal samples. The left hand panel depicts the phylogenomic tree of MAGs generated using GTDB toolkit, with *amoA* clades colored according to the results from Fig. 22 and 24. Right hand panels represent the redundancy, completion and GC content of each MAG. Red diamonds signal the MAGs used later for genomic variation analysis.

CHAPTER 4

3. Sequence variability of AOA MAGs highlights differences in selective pressure

We focused on 3 MAGs from the *amoA*-NP-theta, delta and gamma-2.1 clades, and the 2 *amoA*-NP-gamma-2.2 MAGs to conduct a genomic variation analysis. We chose these MAGs based on completion and abundance, since we then analyzed genome variation solely in samples where the mean coverage of the MAG was over 10x.

Single nucleotide variants (SNVs) refer to the positions in assembled contigs where mapping of short reads highlights a disagreement between the reference nucleotide and the identity of the corresponding base in mapped reads. These positions are a minority compared to stable frequencies, and arise because contigs represent a consensus obtained through assembly. They can illustrate the heterogeneity of a population, inside as well as between metagenomes. Additionally, single amino acid variants (SAAVs) are defined in a similar way as SNVs except they consider the amino acid space. As a consequence, the relative abundance of SAAVs compared to SNVs in a MAG is an indication of the proportion of genetic variants that affect the amino acid sequences and thus potentially the phenotype of a population. It can thus provide hints regarding the strength and type of selection applying to a population.

Here, SNV density was highly variable between MAGs and, for a given MAG, between samples (Fig. 26), but the estimates were comparable to observations from two hydrothermal vent fields (Anderson et al., 2017). Only positions where the divergence from the consensus nucleotide or amino acid was over 10%, and coverage of the position in all considered samples was over 10x were kept. Mean coverage values of MAGs in samples were variable but there was no clear correlation between coverage and SNV density (Fig. S17).

There was no clear distinction in SNV density depending on clade, however, the pattern of SAAV to SNV ratio was significantly linked with *amoA* clade (Kruskal-Wallis rank sum test, $p < 2.2e-16$). It was higher for *amoA*-NP-delta and theta clades, with a mean of 0.31 overall. Conversely, the mean for the NP-gamma clades was 0.25.

CHAPTER 4

This result seems to indicate that the *amoA*-NP-gamma populations characterized here are affected by purifying selection, where mutations are disfavored because they are deleterious, and/or that *amoA*-NP-theta and delta experience positive selection, wherein novel mutations are favored because they confer a selective advantage (Hedge and Wilson, 2016).

amoA-NP-gamma (16S-NP-alpha) populations have been shown to be less genomically diverse than other clades in the deep waters of the Mariana and Ogasawara trenches (Wang et al., 2018). Given that this clade dominates AOA communities of hadopelagic waters (Nunoura et al., 2016; Zhong et al., 2020), it is probable that the populations found in surface sediments are at least partly inherited from the water column. Thus, this clade of ammonia oxidizing archaea could already be well adapted to the environmental conditions found at great depths in the dark ocean (Nunoura et al., 2018). Conversely, the high ratio of SAAVs to SNVs for clades *amoA*-NP-theta and delta could reflect positive selection as an adaptation scenario to the limited resources available in deep subsurface sediments, since these clades are found to be abundant in the deeper layers of abyssal sediments. This scenario of positive selection, associated with gene expansion, has been proposed as a way for Bacteria to adapt to oligotrophic conditions in the water cooling system of a nuclear research reactor (Props et al., 2019).

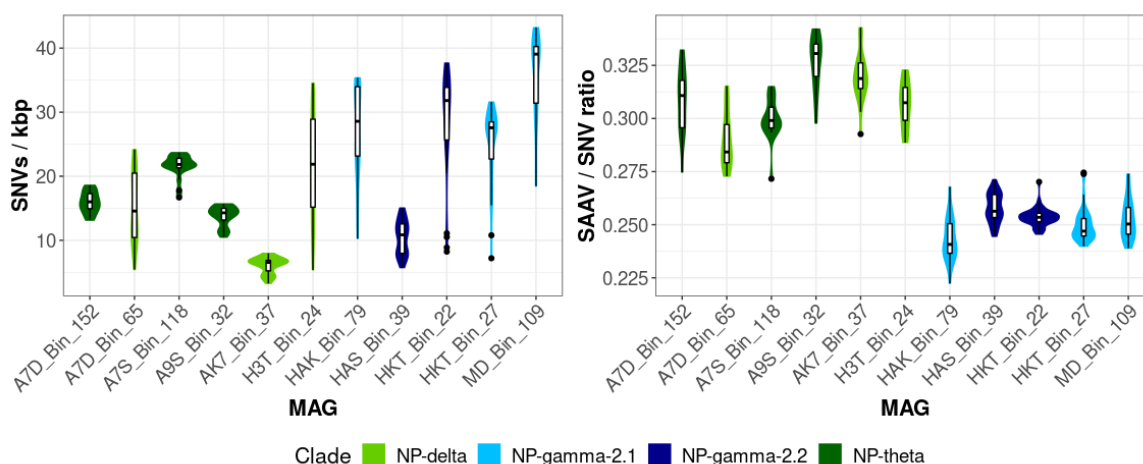


Figure 26: Violin plots showing the estimation of the number of SNVs by kbp and the ratio of SAAV to SNV in each sample for 11 MAGs from four different *amoA* clades.

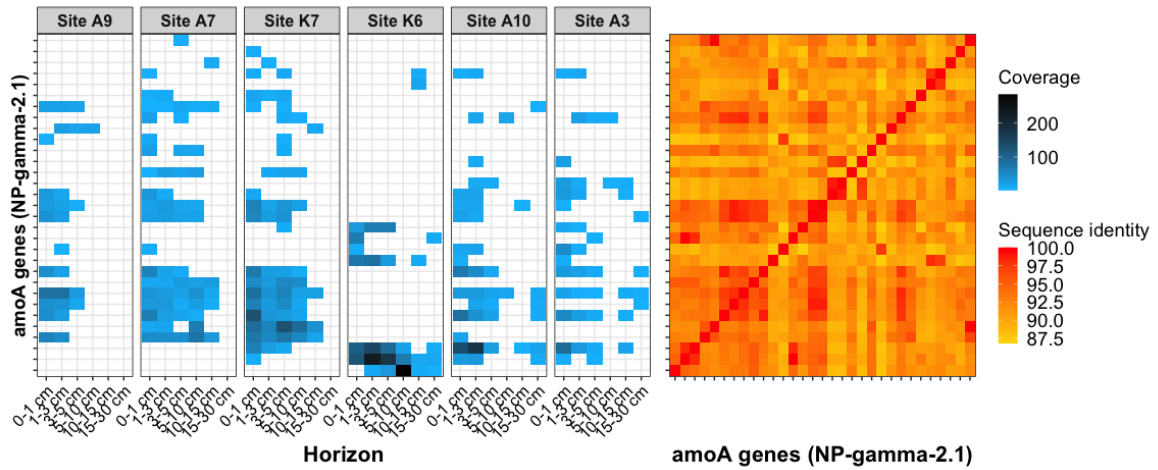
Conclusion

In this study, we endeavored to characterize the fine-scale diversity and genomic variability of Nitrososphaeria, the dominating archaeal lineage in deep sea benthic sediments. We found that communities were very structured, with a predominance of *amoA*-NP-gamma clades in the surface layers, possibly inherited from the deep waters where they are abundant, and *amoA*-NP-theta and delta clades in deeper sediments of open ocean abyssal sites, where oxygen penetrates deeper. We reconstructed 13 good quality MAGs, and 40 additional bins of varying completeness, among which populations related to the yet understudied family *UBA141*, to be further investigated.

The genomic variability of 11 of these MAGs highlighted a pattern of higher ratio of SAAV to SNV for *amoA*-NP-theta and delta clades compared to the gamma clades. Coupled with the fact that *amoA* gene distribution showed a cluster of *amoA*-NP-gamma-2.1 sequences shared between abyssal sites and differing from hadal sequences, these results suggest that these populations, abundant in deep waters, are adapted to the deep pelagic environments where they are found, and follow distinct evolutionary pathways possibly involving purifying selection. Results also suggest positive selection in NP-theta and delta populations that may explain their differentiation in the severely energy- and nutrient-limited environments of subsurface sediments. Future steps of this study will focus on exploring the genomic and functional differences underpinning MAGs differential distribution and distinct variability profiles.

Supplementary figures

A



B



Figure S16: Heatmaps of coverage and sequence identity for *amoA* clades NP-gamma-2.1 (A) and NP-gamma-2.2 (B).

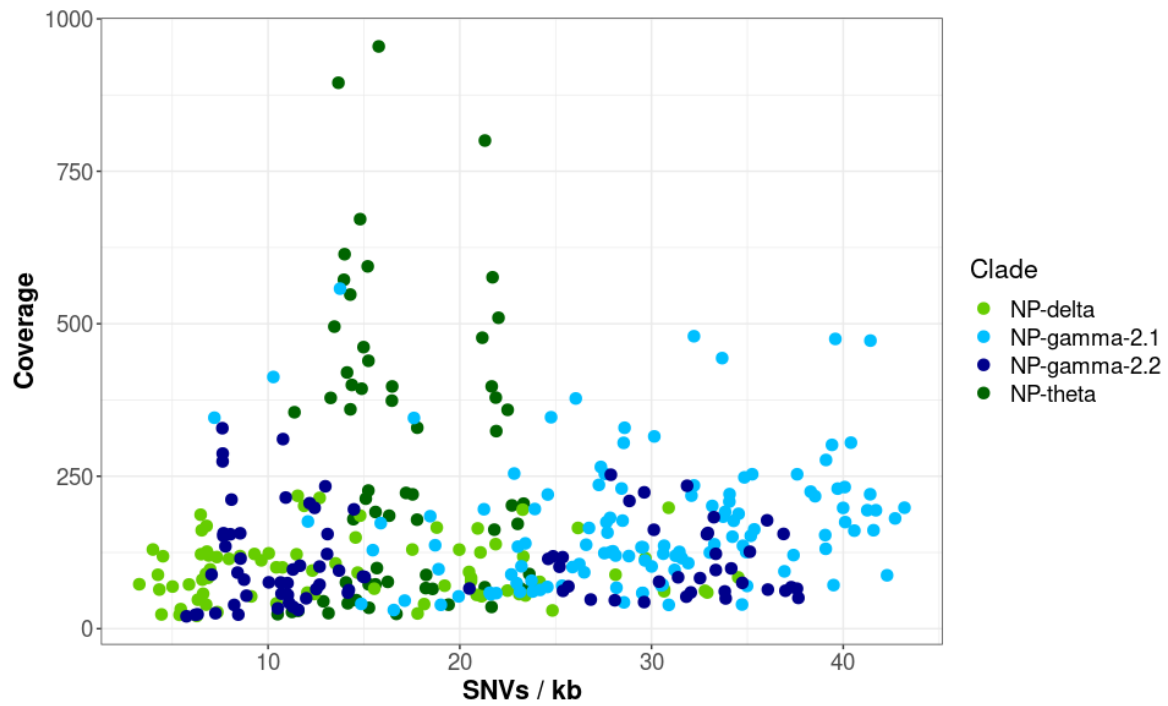


Figure S17: Evolution of the density of SNVs in a MAG with mean coverage of this MAG in samples. The 11 MAGs considered here are the ones presented in Fig. 26. Data is presented only for samples in which coverage of the MAG was over 10x.

GENERAL DISCUSSION

1. Large-scale ecological study of the deep ocean seafloor in the age of NGS

1.1. Accessing archaeal diversity with Next Generation Sequencing

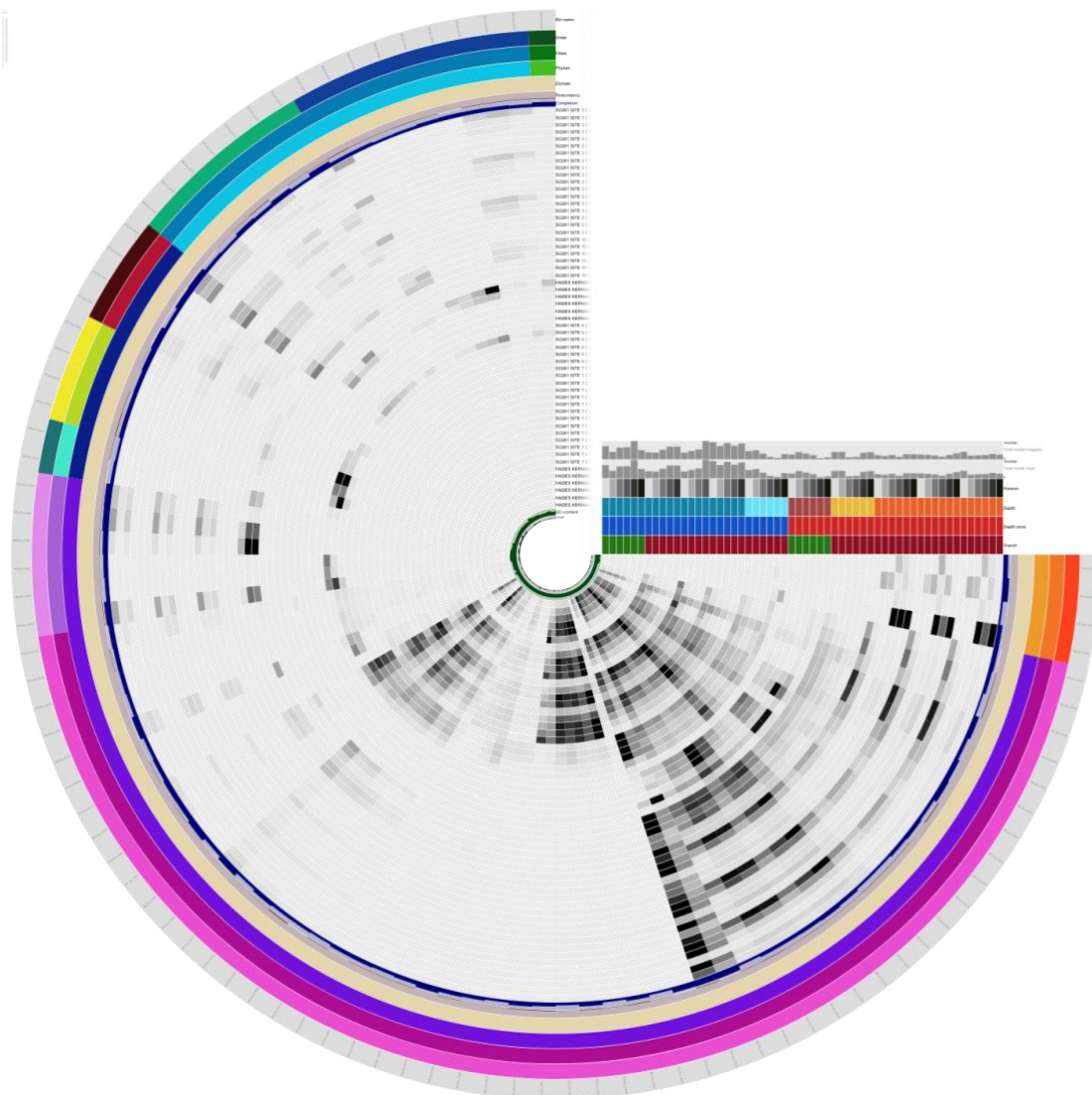
The advent of molecular methods for the description of microbial diversity in the environment opened the door to large-scale explorations of microbial life in deep sea ecosystems. While metabarcoding gives access to the taxonomic diversity of Bacteria and Archaea in a cost-effective manner, both in terms of price and computing resources, whole genome sequencing makes it possible to also characterize the functional diversity of a community. These approaches have been widely used in studies of the deep ocean seafloor, and have been the means of discovering a wide array of lineages, particularly in Archaea (e.g. Vetriani et al., 1999; Huber et al., 2002a; Wang et al., 2005; Durbin and Teske, 2010; Probst et al., 2018; Farag et al., 2020; Kerou et al., 2021).

In this project, we took advantage of these cultivation-independent methods based on the sequencing of environmental DNA to study the microbial diversity of seafloor sediments in abyssal areas and hadal trenches. We also characterized the diversity and distribution of Archaea in benthic hadal sediments based on metabarcoding data, and on 53 Nitrososphaeria (Thaumarchaeota) MAGs reconstructed from metagenomes collected from the Kermadec and Atacama trenches and adjacent abyssal plains. Among these MAGs, four were related to family *UBA141* in the GTDB tree, a probably non ammonia oxidizing lineage defined based on a single genome obtained from a massive effort of genome reconstruction from database available metagenomes (Parks et al., 2017). As such, this lineage had not yet been characterized, either in terms of environmental distribution or metabolic capacities. The addition of the results of our reconstruction efforts will make it possible to take a closer look at this enigmatic lineage.

GENERAL DISCUSSION

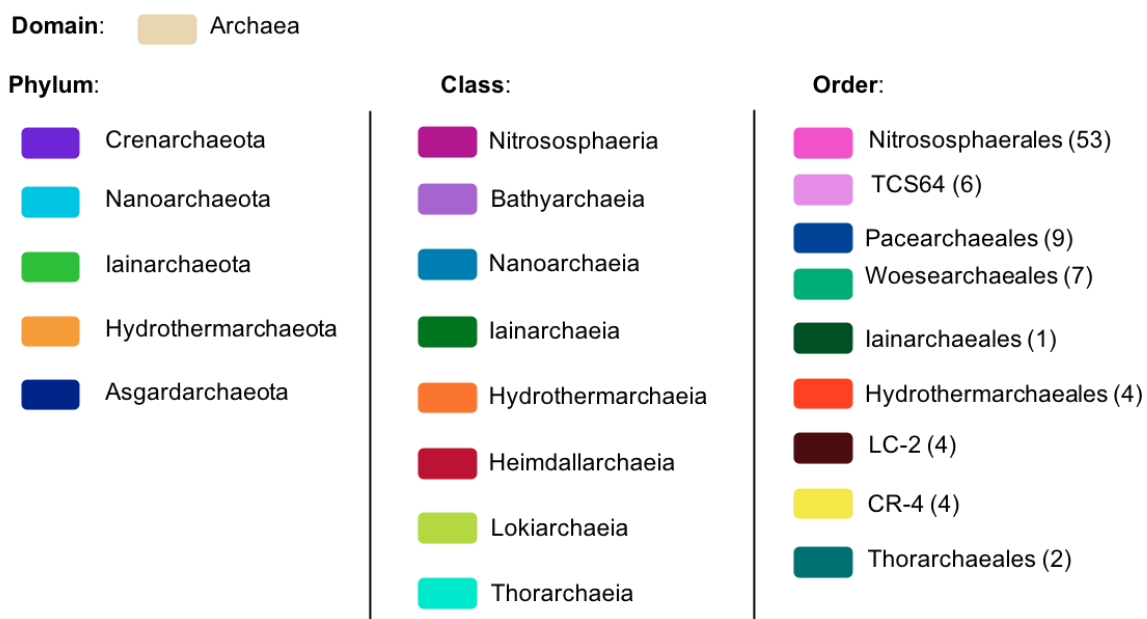
In order to expand our current knowledge of the DPANN superphylum and Asgardarchaeota, of particular ecological and evolutionary interest (Spang et al., 2017; Zaremba-Niedzwiedzka et al., 2017; Dombrowski et al., 2020), 37 additional MAGs from five archaeal phyla were recovered during this project, and should be the object of further studies as well (Fig. 27).

These results, focused on domain Archaea and encompassing five phyla-level lineages, highlight the ongoing potential of whole genome sequencing to better unravel both taxonomic and putative functional diversity in understudied environments.



GENERAL DISCUSSION

Taxonomy legend (outer layers):



Sample legend:

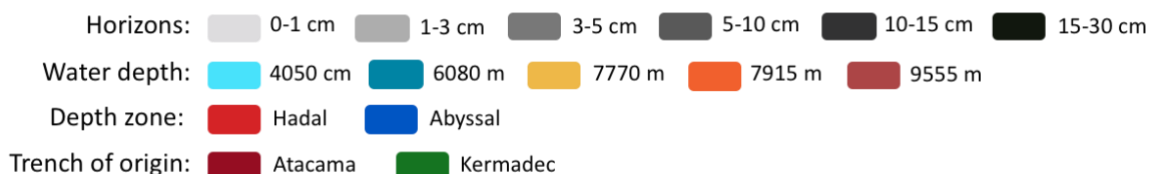


Figure 27: Mean coverage by sample of the 90 archaeal MAGs reconstructed during this project. Each layer of the cladogram represents a sample, or additionally taxonomic information for the outside layers. Each leaf of the cladogram is a MAG. MAGs are organized by taxonomy, which was obtained by placement in the GTDB (Parks et al., 2018).

1.2. Expanding the database of metagenomes to study the global distribution of Bacteria and Archaea

In addition to their usefulness in accessing new diversity, published metagenomes are important resources for microbial ecologists to study the occurrence of functional genes or specific lineages over a wide variety of samples or environments (Delmont et al., 2011). As an

GENERAL DISCUSSION

example, the extensive TARA Oceans dataset, generated from water samples from oceans around the globe, has been used by Delmont et al. (2019) to study the population genetics of a SAR11 clade across oceans. By profiling single-amino acid variants, they were able to link allele frequencies and temperature trends of large-scale ocean currents, as well as quantify the effects of purifying selection on specific protein sequences of these populations. Pereira et al. (2019) employed TARA oceans metagenomes to investigate the distribution of archaeal Marine Group II (MGII) and were able to link MGII subgroups and gene-coding sequences through co-occurrence networks to define possible ecological niches for this lineage.

During this doctoral work, I performed the bioinformatic analyses based on a subsample of 56 metagenomes recovered from 6 sampling sites in two hadal trenches and adjacent abyssal plains. This dataset is destined to be supplemented with 153 additional hadal and abyssal metagenomes with higher vertical resolution that have been produced and assembled using the same pipelines. It will hopefully become a useful tool for deep marine studies in expanding the database of available reference metagenomes.

1.3. A matter of scales

In Chapter 2, we investigated the biogeographic patterns and environmental drivers of the benthic microbial communities at the transition between the Mediterranean Sea and the Atlantic Ocean. This study highlighted the importance of spatial scale when exploring biogeography, with contemporary environmental effects being superseded at inter-regional scale by the legacies of historical processes. Because of the limited dispersion of benthic populations, horizontal spatial scale is linked with temporal scale. Additionally, the slow accumulation rates of sediments in abyssal plains entails long time scales across even relatively short vertical distance (Jahnke, 1996; Roy et al., 2012; Vuillemin et al., 2020).

GENERAL DISCUSSION

This property of deep sea sediments, concurrently with the possible long-term preservation of extracellular DNA by adsorption onto the sediment matrix or complexation with organic compounds, can result in marine subsurface sediments acting as genetic archives (reviewed in Torti et al., 2015). This conservation of ancient DNA has not been evidenced for Bacteria and Archaea (Torti et al., 2018). Although such ancient DNA has allowed for the study of some eukaryotic phyla over thousands of years (Lejzerowicz et al., 2013; De Schepper et al., 2019), the molecular pilot study presented in Chapter 1.2 suggested a negligible influence of long term storage at the scale of eukaryotic communities, at least in the superficial layers of sediment (Brandt et al., 2020). Our results were at the root of the choice of an extraction method allowing to target both intracellular and extracellular DNA with a limited impact, if any, on the inventories of contemporary communities.

In addition to spatial and temporal scales, taxonomic scales should also be taken into account. Because deep sea benthic microbial communities are hitherto sparsely characterized, and composed of lineages still understudied in terms of taxonomic and functional diversity, many studies focus on higher taxonomic-level patterns (e.g. Bienhold et al., 2016; Hoshino et al., 2020), Chapter 2). Co-occurrence networks and biomarker analysis are interesting methods to detect patterns at the “ecological unit” scale, be it OTU or ASV (e.g. Peoples et al., 2019; Hiraoka et al., 2020). With the rise of NGS and whole genome sequencing, it is now possible to focus on single populations, reconstructed or isolated, and study single-nucleotide variants and non-synonymous mutation patterns to infer fine-scale biogeographic processes and evolutionary drivers of diversity (Anderson et al., 2017; Delmont et al., 2019; Crits-Christoph et al., 2020).

To sum up, studying microbial diversity and ecology at multiple spatial, temporal and taxonomic scales is a challenging endeavor, that is nevertheless essential to obtain a comprehensive understanding of an ecosystem and has been rendered accessible by the improvement of sequencing technologies.

GENERAL DISCUSSION

1.4. The importance of a holistic approach for ecosystem characterization

In addition to the scales discussed above, implementing a holistic approach is another crucial parameter when conducting large-scale ecological studies (Karsenti et al., 2011).

In benthic sediments, bioturbation and bioirrigation can have an important influence on oxygen exchanges at the sediment-water interface (Pischedda et al., 2008). Food webs also have important impacts on biogeochemical cycles such as the carbon cycle (Goody, 1994; Schratzberger and Ingels, 2018). For these reasons, the methodological studies presented in Chapter 1 aimed at defining standardized methods of benthic sediment sampling, environmental DNA extraction, and bioinformatic processing of the resulting amplicon datasets, so that the results obtained for each biological compartment could be reliably and easily combined to investigate ecological processes.

We report in Chapters 2 to 4 a consistent pattern of strong vertical stratification of microbial communities. This pattern is typical of sedimentary communities and is linked with the cascade of available electron donors and acceptors (Froelich et al., 1979; Durbin and Teske, 2011; Schaubberger et al., 2021b). This highlights a loop between microbial communities and geochemistry: on the one hand, available substrate and electron acceptors partly determine the thriving lineages, while on the other hand, microbial metabolism impacts the rates of reaction and the chemical gradients. Thus, associating geochemical analyses of benthic sediments with microbiological studies is central to the formulation of more precise hypotheses as to ecosystem functioning.

As an example, biogeochemical characterization of hadal sediments through *in situ* measurements and lab incubations led to the detection of strikingly high rates of anammox in hadal sediments of the Atacama trench, correlated with peaks in the abundance of *Ca. Scalindua* (Thamdrup et al., in prep). With this insight into nitrogen cycling, future studies could make use of the available metagenomes to reconstruct the microbial populations involved in this process.

2. Molecular approaches to uncover new diversity: limits and perspectives

2.1. Challenges in linking 16S and metagenomic inventories of diversity

As stated above, this project relied solely on molecular approaches for the characterization of deep sea benthic microbial communities. Despite allowing to circumvent the obstacle of the challenging and time-consuming cultivation of deep-sea lineages, these approaches come with their own set of challenges.

Chapters 1, 2 and 3 were based on metabarcoding analysis, a widely used technique in microbial ecology over the last decades to produce taxonomic inventories of microbial diversity. The more and more comprehensive results offered by metabarcoding studies have widely expanded our vision of microbial diversity and allowed large-scale studies of bacterial and archaeal distribution patterns to infer their abiotic, biotic, and historical drivers.

In recent years however, there has been a shift towards metagenomics for the exploration of understudied environments, due to the possible characterization of functional diversity it affords. Unfortunately, the most used marker gene for metabarcoding studies, 16S rRNA, is often absent from metagenome-assembled genomes because of the difficulty in the *de novo* assembly of genes including conserved regions. This makes it challenging to bridge the gap between genome-scale studies and 16S-based surveys.

One possibility is to use correlations based on distribution patterns, however, reconstructed MAGs do not often match exactly with a single ASV or OTU. Alternatively, marker genes extraction from metagenomes is an interesting technique to characterize taxonomy since they can more easily be linked back to functional diversity.

In Chapter 1.4, the tested single-copy core genes did not yield diversity estimates comparable to metabarcoding or miTAG results. However, this could be due to the high diversity and low abundance of the lineages considered. Clade specific marker genes such as the *amoA* gene

GENERAL DISCUSSION

used in Chapter 4 are also interesting candidates since they can be directly linked to the functional potential of a community, though they are only applicable to previously characterized lineages. For example, this clade level characterization based on functional genes was not applicable to the widespread Woesearchaeales order observed in Chapter 3. Additionally, recent advances in long-read sequencing technologies (Adewale, 2020; Karst et al., 2021) could eventually allow overcoming these obstacles, through easier reconstruction of full-length 16S sequences and higher completion MAGs, effectively linking metabarcoding and metagenomic based inventories (Jeong et al., 2021).

2.2. Limitations due to lack of completeness in the databases

As explored in Chapter 1.4, lack of sequence representation in databases can lead to biases in PCR primer design. To alleviate some of this bias and more easily design primers adapted to the target environment, McNichol et al. (2021) have proposed a workflow to base primer design on 16S rRNA short sequences extracted from metagenomes (miTAGs).

In our study, by comparing the results of universal and archaea-specific primer sets with miTAG results, we highlighted some important gaps in the coverage of the archaeal primers initially planned for the project, not always predictable through *in silico* analysis. It would thus be interesting to take advantage of the metagenomes presented here and additional database metagenomes to delineate a new set of primers more widely applicable to obtain an accurate diversity coverage.

2.3. Discussions around archaeal taxonomy

As discussed in the introduction, there is yet no standard to name or include reconstructed genomes in official taxonomy databases (Hugenholtz et al., 2021). Some arguments against this addition are the errors that could be introduced in databases through the complicated and

GENERAL DISCUSSION

error-prone assembly processes, and the lack of observed metabolic traits. However, single-cell or long-read sequencing now make it easier to reconstruct near complete genomes (Fullerton and Moyer, 2016; Arikawa et al., 2021) providing valuable context to new results.

For deep-sea Archaea in particular, culturing efforts are often very challenging (Imachi et al., 2020; Hu et al., 2021). On the other hand, an important diversity of archaeal genomes is reconstructed from environmental samples and given putative names (Probst et al., 2018; Dombrowski et al., 2020; Farag et al., 2021). Without a set of guidelines or standard practices in naming conventions or phylogenomic relationship reconstruction, it can become challenging to gather the appropriate data to replace one's results in a wider context.

In this respect, the efforts of the team behind the Genome Taxonomy Database (GTDB) are valuable in providing easier access to genomic data and integration of new discoveries (Parks et al., 2018). In their efforts to advance toward standardizing bacterial and archaeal taxonomy, they reconstructed a phylogenomic tree from 120 ubiquitous single-copy proteins for Bacteria and 122 single-copy proteins for Archaea (Parks et al., 2020; Rinke et al., 2021).

Given that the diversity of Archaea has been gradually uncovered since the 1990s, mostly through molecular approaches, and new lineages are still being described (De Anda et al., 2021), the phylogenomic organization of the domain is regularly updated (Fig. 5). In the past, these updates have come with the proposition of new names for a number of lineages, for example DHVE-5 and 6 became Woesearchaeales and Pacearchaeales (or *Ca.* Woesearchaeota and *Ca.* Pacearchaeota) (partly reviewed in Dombrowski et al., 2019).

The standardization and rank normalization work of Parks et al. and Rinke et al. also led to their proposing new names for a number of clades, such as the Thaumarchaeota phylum, reclassified at class level under the name Nitrososphaeria. Though their work is indeed valuable, the reconstruction of phylogenomic trees is a challenging task, and often the subject of debates. In addition, renaming such clades as the Thaumarchaeota, that has been well established since its proposal in 2008 (Brochier-Armanet et al.) and extensively studied in diverse environments due to the capability of some of its members to aerobically oxidize

GENERAL DISCUSSION

ammonia, will probably lead to some confusion, depending on its degree of adoption by the scientific community (Sanford et al., 2021).

Indeed, united with the aforementioned challenges in comparing metagenomic results and 16S surveys, this lack of taxonomic reference dataframe can lead to additional difficulties in comparing new results to previous studies for a given lineage. This problem is already clear through the lack of congruence between the SILVA v138 release from 2019 and the GTDB taxonomy it integrated for curation. Indeed, SILVA v138 is currently using Crenarchaeota as the name for the phylum uniting classes Bathyarchaeia (phylum Bathyarchaeota in v132) and Nitrososphaeria (phylum Thaumarchaeota in v132), even though Rinke et al. first used this name as a placeholder and since replaced it with Thermoproteota in the GTDB (Rinke et al., 2021). In any case, it will be crucial to keep a record of these proposed taxonomy updates for easy integration of past and future studies.

3. Perspectives for deep sea research

3.1. Importance of experimental evidence to complement molecular results

The previous discussion of the challenges in molecular ecological studies leads us to highlight the well-known importance of experimental evidence in support of molecular-based studies, when realistic. Indeed, molecular approaches are truly valuable in bringing insight into the structure of understudied and hard-to-reach microbial communities, as well as allowing for large-scale explorations of ecosystems populated by uncultivated taxa. However, the hypotheses and considerations they generate need to be supplemented by experimental evidence. Hopefully, the results presented in this work can be the basis for easier experimental methods application.

GENERAL DISCUSSION

In Chapter 3, we investigated the distribution of Woesearchaeales in the deeper benthic horizons of hadal sediments and tried to identify putative association targets. No strong pattern of specific association emerged, but it seemed that Woesearchaeales might form non-specific associations or metabolic consortia with a diversity of other archaeal lineages. It would be interesting to backup these observations and hypotheses using CARD-FISH microscopy to visualize this lineage (Huber et al., 2002a; Wurch et al., 2016; Schwank et al., 2019).

Additionally, single-cell genomics is a promising pathway recently used to infer cultivation conditions appropriate to the isolation of *Nanopusillus acidilobi*, an ectosymbiont of *Acidilobus* cells (Wurch et al., 2016). Cultivation efforts of deep sea Archaea could thus be aided by metagenomic results, as seen in other environments (Lugli et al., 2019).

3.2. Establishment of long-term observatories

Finally, the studies presented here, and most recent microbial work in hadal trenches and abyssal plains, only present a snapshot of the benthic communities and potential ecosystem functioning. Compared to pelagic and coastal environments such as the Bay of Brest, where clear seasonal cycles are visible (Fig. 28), deep-sea sediments are expected to be relatively stable environments. However, deep-sea hydrothermal ecosystems below 1500 m have been shown to be influenced by tidal rhythms (Cuvelier et al., 2017; Mat et al., 2020).

Additionally, anthropogenic impacts to the environment are the subject of strong interest, both with the aim of resource exploitation (e.g. deep-sea mining) and for conservation efforts (Ramirez-Llodra, 2020). In order to assess these longer term rhythms or impacts, the next step in abyssal and hadal benthic research would be the establishment of long-term observatories, as is happening in the deep Arctic Ocean at the LTER Observatory HAUSGARTEN (Soltwedel et al., 2016), and at mid-Atlantic ridge hydrothermal sites with the EMSO-Azores observatory (Rommevaux et al., 2019).

GENERAL DISCUSSION

These observatories deploy geochemical sensors, cameras and other equipment capable of recording and transmitting data regularly, with *in situ* measurements. On the other, biological studies usually rely on annual cruises for sample collection, due to the technical limitations in accessing these sites and retrieving samples. Future technological developments might make it possible to remotely deploy tools for *in situ* fixation and storage of microbiological samples to be collected during annual maintenance cruises, or even *in situ* extraction and sequencing of environmental DNA. Such revolutionary advances would make it possible to obtain time-series data from deep-sea benthic ecosystems, and monitor ecosystem functioning and biogeochemical cycling, exciting integrative perspectives for this field of research.

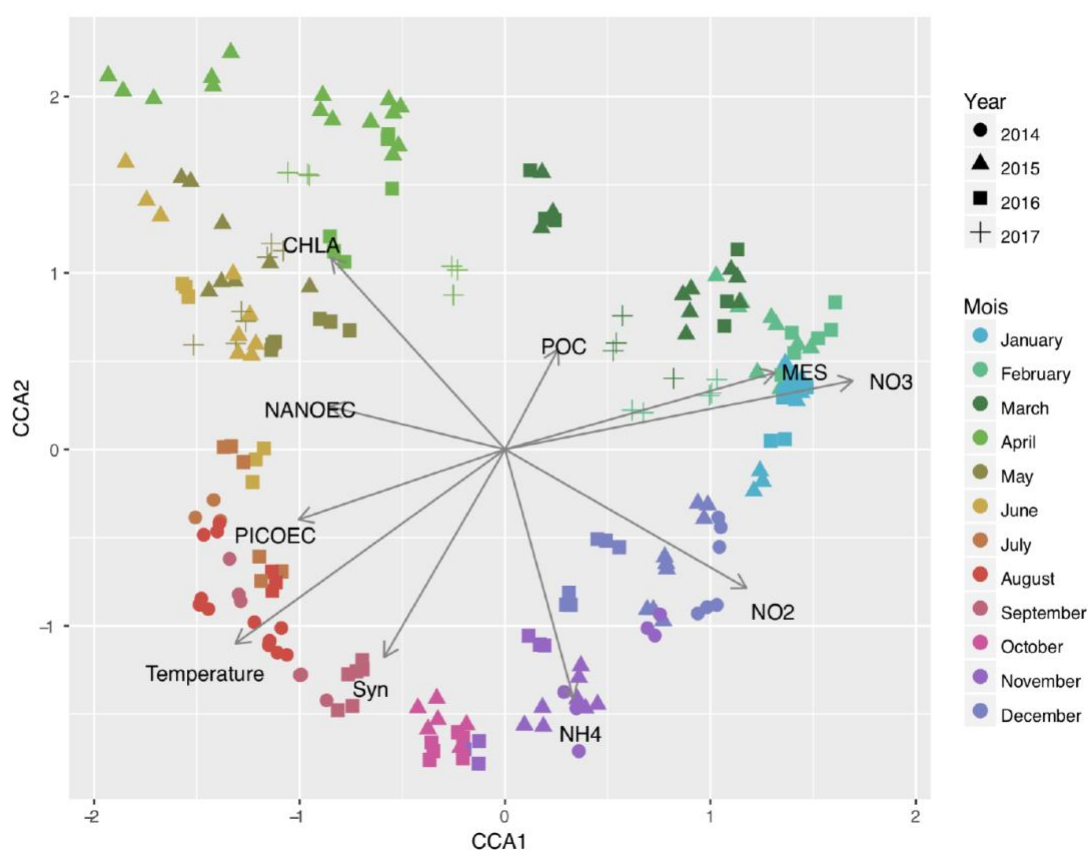


Figure 28: Seasonality of the bacterioplankton communities of the Bay of Brest (Lemonnier, 2019). Canonical Correspondence Analysis (CCA) of the bacterial community composition across the samples and in relation with environmental parameters: chlorophyll a (CHLA), particulate organic carbon (POC), suspended material (MES), nitrates (NO₃), nitrites (NO₂), ammonium (NH₄), temperature and cytometric counts of *Synechococcus* (Syn), Picoeucaryotes (PICOEC) and Nanoeucaryotes (NANOEC).

CONCLUSION AND PERSPECTIVES

CONCLUSION AND PERSPECTIVES

Benthic sediments of the dark ocean below 1000 m represent a vast habitat that plays an important role in global geochemical cycles through the early remineralization of organic matter deposited to the seafloor. However, it is expected to harbor low activity microbial communities, especially benthic sediments of abyssal plains, and has been much more sparsely described than ecosystems where microbial life is stimulated by fluid flow such as hydrothermal vents or cold seeps. In this manuscript, we report the first results of a large-scale survey of microbial communities of the deep ocean seafloor using NGS-based approaches.

Biogeographic study of the transition between Mediterranean Sea and Atlantic Ocean showed a variation in patterns depending on the spatial scale considered, with the historic influence of dispersal limitation and drift superseding environmental selection at larger scales (*i.e.* > 2000 km). Investigating environmental drivers of community structure, we observed a clear depth threshold between 800 and 1200 meters, consistent with the sharp turn over proposed between upper and lower bathyal zones (Watling et al., 2013; Costello et al., 2017). Below this limit, it was proposed that the depth threshold between the lower bathyal zone and abyssal plains did not represent an ecological delineation (Costello and Breyer, 2017). Here the sampling scheme did not make it possible to test this hypothesis, but the planned future addition of new samples to this dataset could bring further insights into the biogeography of abyssal benthic sediments.

Focusing on archaeal communities of two South Pacific hadal trenches, we found that they segregated by habitat (abyssal or hadal environment) as well as sediment depth, most probably reflecting changes in geochemical context and contrasting regimes of organic matter input, as reported for the overall microbial community (Glud et al., 2021; Schaubberger et al., 2021b).

We complemented this analysis by studying the finer-scale clade-level patterns of distribution and genomic variability of ammonia oxidizing archaea in hadal and abyssal sediments using metagenomic data. These results will need to be expanded in the future by identifying possible

CONCLUSION AND PERSPECTIVES

strain diversity and investigating the genes and functions underlying the difference in variability profiles between clades.

Overall, these results bring new insights into the taxonomic diversity of the understudied deep-sea benthic microbial communities. They lay the foundations for continued investigation of the functional diversity and adaptation of hadal sediment archaeal communities. Future steps will also include integrating the taxonomic and functional diversity results with the abiotic geochemical context, as well as in the larger biotic framework through combination with metabarcoding inventories on metazoans and protists.

Version française

Les sédiments benthiques localisés à plus 1000 m de profondeur représentent un vaste habitat qui joue un rôle essentiel dans les cycles géochimiques planétaires car ils sont le siège des premières étapes de la reminéralisation de la matière organique déposée sur les fonds marins. Cependant, les communautés microbiennes benthiques montrent plutôt de faibles taux d'activité, particulièrement dans les sédiments des plaines abyssales, et ont été beaucoup moins décrites que les écosystèmes où la vie microbienne est stimulée par des écoulements de fluides, tels que les cheminées hydrothermales ou les suintements de méthane. Nous avons présenté dans ce manuscrit de thèse les premiers résultats d'une étude à grande échelle des communautés microbiennes présentes dans les sédiments des grands fonds marins, à l'aide d'approches basées sur le séquençage nouvelle génération.

L'étude biogéographique de la transition entre Méditerranée et Atlantique a montré des variations dans les schémas observés suivant l'échelle spatiale considérée, avec une influence historique de la limitation de dispersion et de la dérive écologique plus forte que la sélection environnementale contemporaine sur de longues distances (> 2000 km). Nous avons par ailleurs mis en lumière un seuil de profondeur entre 800 et 1200 mètres, correspondant à une transition dans la structure des communautés microbiennes au niveau du passage de la zone bathyale supérieure à la zone bathyale inférieure (Watling et al., 2013; Costello et al., 2017). Plus bas, il a été suggéré que la limite de profondeur entre zone bathyale inférieure et plaines abyssales ne correspond pas à une délimitation écologique effective (Costello and Breyer, 2017). Le schéma d'échantillonnage considéré ici n'a pas permis de tester cette hypothèse, mais l'ajout futur de nouveaux échantillons à ce jeu de données pourrait permettre d'en savoir plus sur la biogéographie des sédiments abyssaux benthiques. Concernant les communautés archées de deux fosses hadales du Pacifique Sud, nous avons observé une ségrégation par habitat (environnement abyssal ou hadal) et par profondeur de sédiment, reflétant probablement l'évolution du contexte géochimique et les régimes contrastés d'apport en matière organique caractérisant ces deux environnements, comme

CONCLUSION AND PERSPECTIVES

montré pour la communauté microbienne globale (Glud et al., 2021; Schauberger et al., 2021b).

Nous avons complété cette analyse par l'étude à plus haute résolution de la distribution et de la variabilité génomique des clades d'Archées oxydant l'ammoniac dans ces sédiments, à l'aide de données métagénomiques. Ces résultats seront complétés à l'avenir par l'identification de possibles spécificités de souches, de gènes ou de fonctions permettant d'expliquer les différences de profils de variabilité observés pour les différents clades.

Dans l'ensemble, ces résultats apportent de nouvelles connaissances quant à la diversité taxonomique des communautés microbiennes benthiques des grands fonds marins. Ils permettent de poser de solides fondations qui conduiront à l'analyse approfondie de la diversité fonctionnelle et des adaptations aux sédiments hadaux des communautés d'archées. Les prochaines étapes de cette caractérisation nécessiteront une intégration des résultats de diversité taxonomique et fonctionnelle avec le contexte géochimique abiotique, ainsi que le contexte biotique plus large à travers l'ajout d'observations sur les métazoaires et les protistes.

REFERENCES

REFERENCES

- Adam, P.S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. (2017). The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* 11, 2407–2425.
- Adeyale, B.A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *Afr. J. Lab. Med.* 9.
- Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146.
- Alves, R.J.E., Minh, B.Q., and Urich, T. (2018). Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on *amoA* genes. *Nat. Commun.* 17.
- Anderson, R.E., Reveillaud, J., Reddington, E., Delmont, T.O., Eren, A.M., McDermott, J.M., Seewald, J.S., and Huber, J.A. (2017). Genomic variation in microbial populations inhabiting the marine seafloor at deep-sea hydrothermal vents. *Nat. Commun.* 8.
- Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36, 2251–2252.
- Arikawa, K., Ide, K., Kogawa, M., Saeki, T., Yoda, T., Matsushashi, A., Takeyama, H., and Hosokawa, M. (2021). Recovery of high-quality assembled genomes via single-cell genome-guided binning of metagenome assembly. 26.
- Arístegui, J., Gasol, J.M., Duarte, C.M., and Herndl, G.J. (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnol. Oceanogr.* 54, 1501–1529.
- Arndt, S., Jørgensen, B.B., LaRowe, D.E., Middelburg, J.J., Pancost, R.D., and Regnier, P. (2013). Quantifying the degradation of organic matter in marine sediments: A review and synthesis. *Earth-Sci. Rev.* 123, 53–86.
- Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. 17.
- Astorga, A., Oksanen, J., Luoto, M., Soininen, J., Virtanen, R., and Muotka, T. (2012). Distance decay of similarity in freshwater communities: do macro- and microorganisms follow the same rules?: Decay of similarity in freshwater communities. *Glob. Ecol. Biogeogr.* 21, 365–375.
- Aylward, F.O., and Santoro, A.E. (2020). Heterotrophic Thaumarchaea with Small Genomes Are Widespread in the Dark Ocean. *MSystems* 5, 20.
- Bahram, M., Anslan, S., Hildebrand, F., Bork, P., and Tedersoo, L. (2019). Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment: New metabarcoding primers for archaea. *Environ. Microbiol. Rep.* 11, 487–494.
- Baker, B.J., Anda, V.D., Seitz, K.W., Dombrowski, N., Santoro, A.E., and Lloyd, K.G. (2020). Diversity, ecology and evolution of Archaea. *Nat. Microbiol.* 1–14.
- Baker, G.C., Smith, J.J., and Cowan, D.A. (2003). Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* 55, 541–555.
- Ballance, P.F., Ablav, A.G., Pushchin, I.K., Pletnev, S.P., Biryulina, M.G., Itaya, T., Follas,

REFERENCES

- H.A., and Gibson, G.W. (1999). Morphology and history of the Kermadec trench–arc–backarc basin–remnant arc system at 30 to 32°S: geophysical profile, microfossil and K–Ar data. *Mar. Geol.* *159*, 35–62.
- Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2018). EPA-ng: massively parallel evolutionary placement of genetic sequences. *Syst. Bol.* *68*(2), 365–369.
- Barns, S.M., Fundyga, R.E., Jeffries, M.W., and Pace, N.R. (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc. Natl. Acad. Sci.* *91*.
- Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *IcwsM* *8*, 361–362.
- Bayer, B., Vojvoda, J., Reinthaler, T., Reyes, C., Pinto, M., and Herndl, G.J. (2019). *Nitrosopumilus adriaticus* sp. nov. and *Nitrosopumilus piranensis* sp. nov., two ammonia-oxidizing archaea from the Adriatic Sea and members of the class Nitrososphaeria. *Int. J. Syst. Evol. Microbiol.* *69*, 1892–1902.
- Beam, J.P., Becraft, E.D., Brown, J.M., Schulz, F., Jarett, J.K., Bezuidt, O., Poulton, N.J., Clark, K., Dunfield, P.F., Ravin, N.V., et al. (2020). Ancestral Absence of Electron Transport Chains in Patescibacteria and DPANN. *Front. Microbiol.* *11*.
- Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., and Lemaitre, C. (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Comput. Sci.* *2*, e94.
- Biddle, J.F., Lipp, J.S., Lever, M.A., Lloyd, K.G., Sorensen, K.B., Anderson, R., Fredricks, H.F., Elvert, M., Kelly, T.J., Schrag, D.P., et al. (2006). Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proc. Natl. Acad. Sci.* *103*, 3846–3851.
- Bienhold, C., Zinger, L., Boetius, A., and Ramette, A. (2016). Diversity and Biogeography of Bathyal and Abyssal Seafloor Bacteria. *PLOS ONE* *11*, e0148016.
- Bisgaard, M., Christensen, H., Clermont, D., Dijkshoorn, L., Janda, J.M., Moore, E.R.B., Nemec, A., Nørskov-Lauritsen, N., Overmann, J., and Reubsæet, F.A.G. (2019). The use of genomic DNA sequences as type material for valid publication of bacterial species names will have severe implications for clinical microbiology and related disciplines. *Diagn. Microbiol. Infect. Dis.* *95*, 102–103.
- Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* *2008*, P10008.
- Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara Oceans studies plankton at planetary scale. *Science* *348*, 873–873.
- Brandt, M.I., Trouche, B., Henry, N., Liautard-Haag, C., Maignien, L., de Vargas, C., Wincker, P., Poulain, J., Zeppilli, D., and Arnaud-Haond, S. (2020). An Assessment of Environmental Metabarcoding Protocols Aiming at Favoring Contemporary Biodiversity in Inventories of Deep-Sea Communities. *Front. Mar. Sci.* *7*, 234.
- Brandt, M.I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., and Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Mol. Ecol. Resour.*

REFERENCES

- Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* 6, 245–252.
- Bru, D., Martin-Laurent, F., and Philippot, L. (2008). Quantification of the Detrimental Effect of a Single Primer-Template Mismatch by Real-Time PCR Using the 16S rRNA Gene as an Example. *Appl. Environ. Microbiol.* 74, 1660–1663.
- Bruun, A.F.R. (1956). The abyssal fauna: its ecology, distribution and origin. *Nature*.
- Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60.
- Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., et al. (1996). Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*. *Science* 273, 1058–1073.
- Bushnell, B. (2014). BBMap.
- Buttigieg, P.L., and Ramette, A. (2015). Biogeographic patterns of bacterial microdiversity in Arctic deep-sea sediments (HAUSGARTEN, Fram Strait). *Front. Microbiol.* 5.
- Cai, R., Zhang, J., Liu, R., and Sun, C. (2021). Metagenomic Insights into the Metabolic and Ecological Functions of Abundant Deep-Sea Hydrothermal Vent DPANN Archaea. *Appl. Environ. Microbiol.* 87, e03009-20, /aem/87/9/AEM.03009-20.atom.
- Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583.
- Callahan, B.J., McMurdie, P.J., and Holmes, S.P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643.
- Canfield, D.E. (1994). Factors influencing organic carbon preservation in marine sediments. *Chem. Geol.* 114, 315–329.
- Canfield, D.E., and Thamdrup, B. (2009). Towards a consistent classification scheme for geochemical environments, or, why we wish the term ‘suboxic’ would go away. *Geobiology* 7, 385–392.
- Cario, A., Oliver, G.C., and Rogers, K.L. (2019). Exploring the Deep Marine Biosphere: Challenges, Innovations, and Opportunities. *Front. Earth Sci.* 7, 225.
- Castelle, C.J., and Banfield, J.F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* 172, 1181–1197.
- Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., et al. (2015). Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Curr. Biol.* 25, 690–701.
- Castelle, C.J., Méheust, R., Jaffe, A.L., Seitz, K., Gong, X., Baker, B.J., and Banfield, J.F. (2021). Protein Family Content Uncovers Lineage Relationships and Bacterial Pathway Maintenance Mechanisms in DPANN Archaea. *Front. Microbiol.* 12, 660052.

REFERENCES

- Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database (Oxford University Press).
- Cho, J.-C., and Tiedje, J.M. (2000). Biogeography and Degree of Endemicity of Fluorescent *Pseudomonas* Strains in Soil. *Appl. Environ. Microbiol.* 66, 5448–5456.
- Comolli, L.R., Baker, B.J., Downing, K.H., Siegerist, C.E., and Banfield, J.F. (2009). Three-dimensional analysis of the structure and ecology of a novel, ultra-small archaeon. *ISME J.* 3, 159–167.
- Corliss, J.B., Dymond, J., Gordon, L.I., Edmond, J.M., Herzen, R.P. von, Ballard, R.D., Green, K., Williams, D., Bainbridge, A., Crane, K., et al. (1979). Submarine Thermal Springs on the Galápagos Rift. *Science* 203, 1073–1083.
- Costello, M.J., and Breyer, S. (2017). Ocean Depths: The Mesopelagic and Implications for Global Warming. *Curr. Biol.* 27, R36–R38.
- Costello, M.J., Tsai, P., Wong, P.S., Cheung, A.K.L., Basher, Z., and Chaudhary, C. (2017). Marine biogeographic realms and species endemism. *Nat. Commun.* 8.
- Crick, F.H.C. (1958). On protein synthesis. *Symp Soc Exp Biol* 12.
- Crits-Christoph, A., Olm, M.R., Diamond, S., Bouma-Gregson, K., and Banfield, J.F. (2020). Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME J.* 14, 1834–1846.
- Csardi, M.G. (2013). Package 'igraph.' Last Accessed 3, 2013.
- Cui, G., Li, J., Gao, Z., and Wang, Y. (2019). Spatial variations of microbial communities in abyssal and hadal sediments across the Challenger Deep. *PeerJ* 7, e6961.
- Cuvelier, D., Legendre, P., Laës-Huon, A., Sarradin, P.-M., and Sarrazin, J. (2017). Biological and environmental rhythms in (dark) deep-sea hydrothermal ecosystems. *Biogeosciences* 14, 2955–2977.
- Czech, L., Barbera, P., and Stamatakis, A. (2020). Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* 36, 3263–3265.
- Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, giab008.
- Danovaro, R., Della Croce, N., Dell'Anno, A., and Pusceddu, A. (2003). A depocenter of organic matter at 7800m depth in the SE Pacific Ocean. *Deep Sea Res. Part Oceanogr. Res. Pap.* 50, 1411–1420.
- Danovaro, R., Corinaldesi, C., Rastelli, E., and Dell'Anno, A. (2015). Towards a better quantitative assessment of the relevance of deep-sea viruses, Bacteria and Archaea in the functioning of the ocean seafloor. *Aquat. Microb. Ecol.* 75, 81–90.
- Danovaro, R., Molari, M., Corinaldesi, C., and Dell'Anno, A. (2016). Macroecological drivers of archaea and bacteria in benthic deep-sea ecosystems. *Sci. Adv.* 2, e1500961.
- Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., Bik, H.M., and Eisen, J.A. (2014). PhyloSift:

REFERENCES

phylogenetic analysis of genomes and metagenomes. *PeerJ* 2, e243.

Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A., and Callahan, B.J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6.

De Anda, V., Chen, L.-X., Dombrowski, N., Hua, Z.-S., Jiang, H.-C., Banfield, J.F., Li, W.-J., and Baker, B.J. (2021). Brockarchaeota, a novel archaeal phylum with unique and versatile carbon cycling pathways. *Nat. Commun.* 12, 2404.

De Schepper, S., Ray, J.L., Skaar, K.S., Sadatzki, H., Ijaz, U.Z., Stein, R., and Larsen, A. (2019). The potential of sedimentary ancient DNA for reconstructing past sea ice evolution. *ISME J.* 13, 2566–2577.

Delmas, E., Besson, M., Brice, M.-H., Burkle, L.A., Dalla Riva, G.V., Fortin, M.-J., Gravel, D., Guimarães, P.R., Hembry, D.H., Newman, E.A., et al. (2019). Analysing ecological networks of species interactions: Analyzing ecological networks. *Biol. Rev.* 94, 16–36.

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

Delmont, T.O., Kiefl, E., Kilinc, O., Esen, O.C., Uysal, I., Rappé, M.S., Giovannoni, S., and Eren, A.M. (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *ELife* 8, e46497.

DeLong, E.F. (1992). Archaea in coastal marine environments. *Proc. Natl. Acad. Sci.* 89, 5685–5689.

Deming, J. (1985). Bacterial growth in deep-sea sediment trap and boxcore samples. *Mar. Ecol. Prog. Ser.* 25, 305–312.

Devol, A.H., Anderson, J.J., Kuivila, K., and Murray, J.W. (1984). A model for coupled sulfate reduction and methane oxidation in the sediments of Saanich Inlet. *Geochim. Cosmochim. Acta* 48, 993–1004.

D'Hondt, S., Jørgensen, B.B., Miller, D.J., Batzke, A., Blake, R., Cragg, B.A., Cypionka, H., Dickens, G.R., Ferdelman, T., Hinrichs, K.-U., et al. (2004). Distributions of Microbial Activities in Deep Subseafloor Sediments. *Sci. New Ser.* 306, 2216–2221.

D'Hondt, S., Spivack, A.J., Pockalny, R., Ferdelman, T.G., Fischer, J.P., Kallmeyer, J., Abrams, L.J., Smith, D.C., Graham, D., Hasiuk, F., et al. (2009). Subseafloor sedimentary life in the South Pacific Gyre. *Proc. Natl. Acad. Sci.* 106, 11651–11656.

van Dijk, E.L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Exp. Cell Res.* 322, 12–20.

Divins, D. (2003). Total sediment thickness of the world's oceans and marginal seas. NOAA Natl. Geophys. Data Cent.

Dombrowski, N., Lee, J.-H., Williams, T.A., Offre, P., and Spang, A. (2019). Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* 366.

Dombrowski, N., Williams, T.A., Sun, J., Woodcroft, B.J., Lee, J.-H., Minh, B.Q., Rinke, C., and Spang, A. (2020). Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* 11, 3939.

REFERENCES

- Durbin, A.M., and Teske, A. (2010). Sediment-associated microdiversity within the Marine Group I Crenarchaeota: Sediment Marine Group I Archaea. *Environ. Microbiol. Rep.* 2, 693–703.
- Durbin, A.M., and Teske, A. (2011). Microbial diversity and stratification of South Pacific abyssal marine sediments. *Environ. Microbiol.* 13, 3219–3234.
- Edgar, R.C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. 6.
- Elkins, J.G., Podar, M., Graham, D.E., Makarova, K.S., Wolf, Y., Randau, L., and Hedlund, B.P. (2008). A korarchaeal genome reveals insights into the evolution of the Archaea. *PNAS* 105, 8102-8107.
- Eloe-Fadrosh, E.A., Ivanova, N.N., Woyke, T., and Kyrpides, N.C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.* 1, 15032.
- Emerson, S., Jahnke, R., Bender, M., Froelich, P., Klinkhammer, G., Bowser, C., and Setlock, G. (1980). Early diagenesis in sediments from the eastern equatorial Pacific, I. Pore water nutrient and carbonate results. *Earth Planet. Sci. Lett.* 49, 57–80.
- Eren, a. M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G., and Sogin, M.L. (2013a). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* n/a-n/a.
- Eren, A.M., Vineis, J.H., Morrison, H.G., and Sogin, M.L. (2013b). A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLOS ONE* 8, e66643.
- Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2014). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.*
- Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3, e1319.
- Eren, A.M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S.E., Schechter, M.S., Fink, I., Pan, J.N., Yousef, M., Fogarty, E.C., et al. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* 6, 3–6.
- Espinoza, J.L., Shah, N., Singh, S., Nelson, K.E., and Dupont, C.L. (2020). Applications of weighted association networks applied to compositional data in biology. *Environ. Microbiol.* 22, 3020–3038.
- Fang, J., Zhang, L., and Bazylinski, D.A. (2010). Deep-sea piezosphere and piezophiles: geomicrobiology and biogeochemistry. *Trends Microbiol.* 18, 413–422.
- Farag, I.F., Biddle, J.F., Zhao, R., Martino, A.J., House, C.H., and León-Zayas, R.I. (2020). Metabolic potentials of archaeal lineages resolved from metagenomes of deep Costa Rica sediments. *ISME J.* 14, 1345–1358.
- Farag, I.F., Zhao, R., and Biddle, J.F. (2021). “ *Sifarchaeota* ,” a Novel Asgard Phylum from Costa Rican Sediment Capable of Polysaccharide Degradation and Anaerobic Methylophony. *Appl. Environ. Microbiol.* 87.

REFERENCES

- Fierer, N., and Jackson, R.B. (2006). The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* 103, 626–631.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Bult, C.J., Tomb, J.-F., Sutton, G., Fields, C., Liu, L., Spriggs, T., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269.
- Fossing, H., Gallardo, V.A., Jorgensen, B.B., Hüttel, M., Nielsen, L.P., Schulz, H., Canfield, D.E., Forster, S., Glud, R.N., Gundersen, J.K., et al. (1995). Concentration and transport of nitrate by the mat-forming sulphur bacterium *Thioploca*. *Nature* 374.
- Fox, G.E., Pechman, K.R., and Woese, C.R. (1977). Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Prokaryotic Systematics. *Int. J. Syst. Bacteriol.* 27, 44–57.
- Francis, C.A., Roberts, K.J., Beman, J.M., Santoro, A.E., and Oakley, B.B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. 6.
- François, D.X., Godfroy, A., Mathien, C., Aubé, J., Cathalot, C., Lesongeur, F., L'Haridon, S., Philippon, X., and Roussel, E.G. (2021). *Persephonella atlantica* sp. nov.: How to adapt to physico-chemical gradients in high temperature hydrothermal habitats. *Syst. Appl. Microbiol.* 44, 10.
- Friedline, C.J., Franklin, R.B., McCallister, S.L., and Rivera, M.C. (2012). Bacterial assemblages of the eastern Atlantic Ocean reveal both vertical and latitudinal biogeographic signatures. *Biogeosciences* 9, 2177–2193.
- Froelich, P.N., Klinkhammer, G.P., Bender, M.L., Luedtke, N.A., Heath, G.R., Cullen, D., Dauphin, P., and Blaynehartman, D.Hammond. (1979). Early oxidation of organic matter in pelagic sediments of the eastern equatorial Atlantic: suhoxic diagenesis. 16.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152.
- Fuhrman, J.A., and Davis, A.A. (1997). Widespread Archaea and novel Bacteria from the deep sea as shown by 16S rRNA gene sequences. *Mar. Ecol. Prog. Ser.* 150, 275–285.
- Fuhrman, J.A., McCallum, K., and Davis, A.A. (1992). Novel major archaeobacterial group from marine plankton. *Nature* 356, 148.
- Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P.J., Soen, Y., and Shental, N. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* 6, 17.
- Fullerton, H., and Moyer, C.L. (2016). Comparative Single-Cell Genomics of Chloroflexi from the Okinawa Trough Deep-Subsurface Biosphere. *Appl. Environ. Microbiol.* 82, 3000–3008.
- Galperin, M.Y., Wolf, Y.I., Makarova, K.S., Vera Alvarez, R., Landsman, D., and Koonin, E.V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* 49, D274–D281.
- García-López, R., Cornejo-Granados, F., Lopez-Zavala, A.A., Sánchez-López, F., Cota-Huizar, A., Sotelo-Mundo, R.R., Guerrero, A., Mendoza-Vargas, A., Gómez-Gil, B., and Ochoa-Leyva, A. (2020). Doing More with Less: A Comparison of 16S Hypervariable Regions

REFERENCES

- in Search of Defining the Shrimp Microbiota. *Microorganisms* 8, 134.
- del Giorgio, P.A., and Duarte, C.M. (2002). Respiration in the open ocean. *Nature* 420, 379–384.
- Giovannelli, D., Molari, M., d’Errico, G., Baldrighi, E., Pala, C., and Manini, E. (2013). Large-Scale Distribution and Activity of Prokaryotes in Deep-Sea Surface Sediments of the Mediterranean Sea and the Adjacent Atlantic Ocean. *PLoS ONE* 8, e72996.
- Glud, R.N. (2008). Oxygen dynamics of marine sediments. *Mar. Biol. Res.* 4, 243–289.
- Glud, R.N., Berg, P., Thamdrup, B., Larsen, M., Stewart, H.A., Jamieson, A.J., Glud, A., Oguri, K., Sanei, H., Rowden, A.A., et al. (2021). Hadal trenches are dynamic hotspots for early diagenesis in the deep sea. *Commun. Earth Environ.* 2, 21.
- Golyshina, O.V., Toshchakov, S.V., Makarova, K.S., Gavrillov, S.N., Korzhenkov, A.A., Cono, V.L., Arcadi, E., Nechitaylo, T.Y., Ferrer, M., Kublanov, I.V., et al. (2017). ‘ARMAN’ archaea depend on association with euryarchaeal host in culture and in situ. *Nat. Commun.* 8, 1–12.
- Gooday, A.J. (1994). The Biology of Deep-Sea Foraminifera: A Review of Some Advances and Their Applications in Paleoceanography. *PALAIOS* 9, 14.
- Graham, E.D., Heidelberg, J.F., and Tully, B.J. (2017). BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* 5, e3035.
- Gralka, M., Szabo, R., Stocker, R., and Cordero, O.X. (2020). Trophic Interactions and the Drivers of Microbial Community Assembly. *Curr. Biol.* 30, R1176–R1188.
- Green, J., and Bohannan, B.J.M. (2006). Spatial scaling of microbial biodiversity. *Trends Ecol. Evol.* 21, 501–507.
- Green, J.L., Bohannan, B.J.M., and Whitaker, R.J. (2008). Microbial Biogeography: From Taxonomy to Traits. *Science* 320, 1039–1043.
- Gruber-Vodicka, H.R., Seah, B.K.B., and Pruesse, E. (2020). phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. *MSystems* 5.
- Guy, L., and Ettema, T.J.G. (2011). The archaeal ‘TACK’ superphylum and the origin of eukaryotes. *Trends Microbiol.* 19, 580–587.
- Han, R., Zhang, X., Liu, J., Long, Q., Chen, L., Liu, D., and Zhu, D. (2017). Microbial community structure and diversity within hypersaline Keke Salt Lake environments. *Can. J. Microbiol.* 63, 895–908.
- Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C., and Martiny, J.B.H. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat. Rev. Microbiol.* 10, 497–506.
- Hedge, J., and Wilson, D.J. (2016). Practical Approaches for Detecting Selection in Microbial Genomes. *PLOS Comput. Biol.* 12, e1004739.
- Hedges, J.I., Clark, W.A., and Come, G.L. (1988). Fluxes and reactivities of organic matter in a coastal marine bay: Fluxes and reactions. *Limnol. Oceanogr.* 33, 1137–1152.
- Hijmans, R.J. (2019). *geosphere: Spherical Trigonometry*.

REFERENCES

- Hinrichs, K.-U., and Boetius, A. (2002). The Anaerobic Oxidation of Methane: New Insights in Microbial Ecology and Biogeochemistry. In *Ocean Margin Systems*, G. Wefer, D. Billett, D. Hebbeln, B.B. Jørgensen, M. Schlüter, and T.C.E. van Weering, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 457–477.
- Hinrichs, K.-U., Hayes, J.M., Bach, W., Spivack, A.J., Hmelo, L.R., Holm, N.G., Johnson, C.G., and Sylva, S.P. (2006). Biological formation of ethane and propane in the deep marine subsurface. *Proc. Natl. Acad. Sci.* *103*, 14684–14689.
- Hiraoka, S., Hirai, M., Matsui, Y., Makabe, A., Minegishi, H., Tsuda, M., Juliarni, Rastelli, E., Danovaro, R., Corinaldesi, C., et al. (2020). Microbial community and geochemical analyses of trans-trench sediments for understanding the roles of hadal environments. *ISME J.* *14*, 740–756.
- Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* *35*, 518–522.
- Hooper, A.B., Vannelli, T., Bergmann, D.J., and Arciero, D.M. (1997). Enzymology of the oxidation of ammonia to nitrite by bacteria. *9*.
- Horner-Devine, M.C., Lage, M., Hughes, J.B., and Bohannon, B.J.M. (2004). A taxa–area relationship for bacteria. *Nature* *432*, 750–753.
- Hoshino, T., Doi, H., Uramoto, G.-I., Wörmer, L., Adhikari, R.R., Xiao, N., Morono, Y., D’Hondt, S., Hinrichs, K.-U., and Inagaki, F. (2020). Global diversity of microbial communities in marine sediment. *Proc. Natl. Acad. Sci.* 201919139.
- Hu, H., Natarajan, V.P., and Wang, F. (2021). Towards enriching and isolation of uncultivated archaea from marine sediments using a refined combination of conventional microbial cultivation methods. *Mar. Life Sci. Technol.* *3*, 231–242.
- Huber, H., Hohn, M.J., Rachel, R., Fuchs, T., Wimmer, V.C., and Stetter, K.O. (2002a). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* *417*, 63–67.
- Huber, J.A., Butterfield, D.A., and Baross, J.A. (2002b). Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge seafloor habitat. *Appl. Environ. Microbiol.* *68*, 1585–1594.
- Huber, J.A., Johnson, H.P., Butterfield, D.A., and Baross, J.A. (2006). Microbial life in ridge flank crustal fluids. *Environ. Microbiol.* *8*, 88–99.
- Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hemsdorf, A.W., Amano, Y., Ise, K., et al. (2016a). A new view of the tree of life. *Nat. Microbiol.* *1*, 1–6.
- Hug, L.A., Thomas, B.C., Sharon, I., Brown, C.T., Sharma, R., Hettich, R.L., Wilkins, M.J., Williams, K.H., Singh, A., and Banfield, J.F. (2016b). Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages: N- and C-cycling organisms in the subsurface. *Environ. Microbiol.* *18*, 159–173.
- Hugenholtz, P., Chuvpochina, M., Oren, A., Parks, D.H., and Soo, R.M. (2021). Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J.*

REFERENCES

- Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119.
- Ichino, M.C., Clark, M.R., Drazen, J.C., Jamieson, A., Jones, D.O.B., Martin, A.P., Rowden, A.A., Shank, T.M., Yancey, P.H., and Ruhl, H.A. (2015). The distribution of benthic biomass in hadal trenches: A modelling approach to investigate the effect of vertical and lateral organic matter transport to the seafloor. *Deep Sea Res. Part Oceanogr. Res. Pap.* 100, 21–33.
- Imachi, H., Nobu, M.K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., et al. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* 577, 519–525.
- International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Iversen, N., and Jørgensen, B.B. (1985). Anaerobic methane oxidation rates at the sulfate-methane transition in marine sediments from Kattegat and Skagerrak (Denmark): Anaerobic methane oxidation. *Limnol. Oceanogr.* 30, 944–955.
- Jacob, M., Soltwedel, T., Boetius, A., and Ramette, A. (2013). Biogeography of Deep-Sea Benthic Bacteria at Regional Scale (LTER HAUSGARTEN, Fram Strait, Arctic). *PLoS ONE* 8, e72779.
- Jahnke, R.A. (1996). The global ocean flux of particulate organic carbon: Areal distribution and magnitude. *Glob. Biogeochem. Cycles* 10, 71–88.
- Jamieson, A. (2015). *The Hadal Zone: Life in the Deepest Oceans* (Cambridge: Cambridge University Press).
- Jamieson, A.J., Fujii, T., Mayor, D.J., Solan, M., and Priede, I.G. (2010). Hadal trenches: the ecology of the deepest places on Earth. *Trends Ecol. Evol.* 25, 190–197.
- Jamieson, A.J., Fang, J., and Cui, W. (2018). Exploring the Hadal Zone: Recent Advances in Hadal Science and Technology. *Deep Sea Res. Part II Top. Stud. Oceanogr.* 155, 1–3.
- Jeong, J., Yun, K., Mun, S., Chung, W.-H., Choi, S.-Y., Nam, Y., Lim, M.Y., Hong, C.P., Park, C., Ahn, Y.J., et al. (2021). The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. *Sci. Rep.* 11, 1727.
- Jochum, L.M., Chen, X., Lever, M.A., Loy, A., Jørgensen, B.B., Schramm, A., and Kjeldsen, K.U. (2017). Depth Distribution and Assembly of Sulfate-Reducing Microbial Communities in Marine Sediments of Aarhus Bay. *Appl. Environ. Microbiol.* 83.
- Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11, 431.
- Jørgensen, B.B., and Boetius, A. (2007). Feast and famine — microbial life in the deep-sea bed. *Nat. Rev. Microbiol.* 5, 770–781.
- Jørgensen, B.B., Andrén, T., and Marshall, I.P.G. (2020). Sub-seafloor biogeochemical processes and microbial life in the Baltic Sea. *Environ. Microbiol.*
- Jørgensen, S.L., Thorseth, I.H., Pedersen, R.B., Baumberger, T., and Schleper, C. (2013). Quantitative and phylogenetic study of the Deep Sea Archaeal Group in sediments of the

REFERENCES

- Arctic mid-ocean spreading ridge. *Front. Microbiol.* **4**.
- Jurasinsk, G., and Retzer, V. (2012). simba: A Collection of functions for similarity analysis of vegetation data.
- Kahle, D., and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *R J.* **5**, 144.
- Kallmeyer, J., Pockalny, R., Adhikari, R.R., Smith, D.C., and D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Natl. Acad. Sci.* **109**, 16213–16216.
- Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30.
- Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165.
- Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359.
- Kang, Y.J., Cheng, J., Mei, L.J., Hu, J., Piao, Z., and Yin, S.X. (2010). Multiple copies of 16S rRNA gene affect the restriction patterns and DGGE profile revealed by analysis of genome database. *Microbiology* **79**, 655–662.
- Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzon, F., Claverie, J.-M., et al. (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biol.* **9**, e1001177.
- Karst, S.M., Ziels, R.M., Kirkegaard, R.H., Sørensen, E.A., McDonald, D., Zhu, Q., Knight, R., and Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169.
- Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066.
- Kebschull, J.M., and Zador, A.M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* gkv717.
- Kerou, M., Ponce-Toledo, R.I., Zhao, R., Abby, S.S., Hirai, M., Nomaki, H., Takaki, Y., Nunoura, T., Jørgensen, S.L., and Schleper, C. (2021). Genomes of Thaumarchaeota from deep sea sediments reveal specific adaptations of three independently evolved lineages. *ISME J.* 1–17.
- Kim, M., Oh, H.-S., Park, S.-C., and Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64**, 346–351.
- Kioka, A., Schwestermann, T., Moernaut, J., Ikehara, K., Kanamatsu, T., McHugh, C.M., dos Santos Ferreira, C., Wiemer, G., Haghpor, N., Kopf, A.J., et al. (2019). Megathrust earthquake drives drastic organic carbon supply to the hadal trench. *Sci. Rep.* **9**, 1553.
- Kirkpatrick, J.B., Walsh, E.A., and D'Hondt, S. (2019). Microbial Selection and Survival in Subseafloor Sediment. *Front. Microbiol.* **10**.
- Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O.

REFERENCES

- (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* *41*, e1–e1.
- Knittel, K., Lösekann, T., Boetius, A., Kort, R., and Amann, R. (2005). Diversity and Distribution of Methanotrophic Archaea at Cold Seeps. *Appl. Environ. Microbiol.* *71*, 467–479.
- Könneke, M., Bernhard, A.E., de la Torre, J.R., Walker, C.B., Waterbury, J.B., and Stahl, D.A. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* *437*, 543–546.
- Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* *28*, 3211–3217.
- Kormas, K.A., Tamaki, H., Hanada, S., and Kamagata, Y. (2009). Apparent richness and community composition of Bacteria and Archaea in geothermal springs. *Aquat Microb Ecol* *10*.
- Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* *28*, 2520–2522.
- Krause, S., Bremges, A., Münch, P.C., McHardy, A.C., and Gescher, J. (2017). Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. *Sci. Rep.* *7*, 1–13.
- Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., and Bonneau, R.A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* *11*, e1004226.
- Kvenvolden, K.A. (1993). Gas hydrates-geological perspective and global change. *Rev. Geophys.* *31*, 173–187.
- LaBrie, R., Bélanger, S., Benner, R., and Maranger, R. (2020). Spatial abundance distribution of prokaryotes is associated with dissolved organic matter composition and ecosystem function. *Limnol. Oceanogr.*
- Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci.* *82*, 6955–6959.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* *10*, R25.
- Lapidus, A.L., and Korobeynikov, A.I. (2021). Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. *Front. Microbiol.* *12*, 613791.
- Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* *25*.
- Lecours, V., Devillers, R., Schneider, D., Lucieer, V., Brown, C., and Edinger, E. (2015). Spatial scale and geographic context in benthic habitat mapping: review and future directions. *Mar. Ecol. Prog. Ser.* *535*, 259–284.
- Lee, M.D. (2019). GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* *35*, 4162–4164.
- Lejzerowicz, F., Esling, P., Majewski, W., Szczuciński, W., Decelle, J., Obadia, C., Arbizu,

REFERENCES

- P.M., and Pawlowski, J. (2013). Ancient DNA complements microfossil record in deep-sea subsurface sediments. *Biol. Lett.* 9, 20130283.
- Lever, M.A., Heuer, V.B., Morono, Y., Masui, N., Schmidt, F., Alperin, M.J., Inagaki, F., Hinrichs, K.-U., and Teske, A. (2010). Acetogenesis in Deep Subseafloor Sediments of The Juan de Fuca Ridge Flank: A Synthesis of Geochemical, Thermodynamic, and Gene-based Evidence. *Geomicrobiol. J.* 27, 183–211.
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676.
- Li, M., Mi, T., He, H., Chen, Y., Zhen, Y., and Yu, Z. (2021). Active bacterial and archaeal communities in coastal sediments: Biogeography pattern, assembly process and co-occurrence relationship. *Sci. Total Environ.* 750, 142252.
- Lipp, J.S., Morono, Y., Inagaki, F., and Hinrichs, K.-U. (2008). Significant contribution of Archaea to extant biomass in marine subsurface sediments. *Nature* 454, 991–994.
- Liu, J., Zhu, S., Liu, X., Yao, P., Ge, T., and Zhang, X.-H. (2020). Spatiotemporal dynamics of the archaeal community in coastal sediments: assembly process and co-occurrence relationship. *ISME J.* 14, 1463–1478.
- Liu, X., Li, M., Castelle, C.J., Probst, A.J., Zhou, Z., Pan, J., Liu, Y., Banfield, J.F., and Gu, J.-D. (2018). Insights into the ecology, evolution, and metabolism of the widespread Woese archaeal lineages. *Microbiome* 6, 102.
- Liu, X., Wang, Y., and Gu, J.-D. (2021). Ecological distribution and potential roles of Woese archaeota in anaerobic biogeochemical cycling unveiled by genomic analysis. *Comput. Struct. Biotechnol. J.* 19, 794–800.
- Lloyd, K.G., Bird, J.T., Buongiorno, J., Deas, E., Kevorkian, R., Noordhoek, T., Rosalsky, J., and Roy, T. (2020). Evidence for a growth zone for deep subsurface microbial clades in near-surface anoxic sediments (*Microbiology*).
- Locey, K.J., Muscarella, M.E., Larsen, M.L., Bray, S.R., Jones, S.E., and Lennon, J.T. (2020). Dormancy dampens the microbial distance–decay relationship. *Philos. Trans. R. Soc. B Biol. Sci.* 11.
- Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F.M., Ferrera, I., Sarmiento, H., Hingamp, P., Ogata, H., de Vargas, C., Lima-Mendez, G., et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities: Using *mi tag* s to explore microbial communities. *Environ. Microbiol.* 16, 2659–2671.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.
- Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, null, Buchner, A., Lai, T., Steppi, S., Jobb, G., et al. (2004). ARB: a software environment for sequence data. *Nucleic Acids Res.* 32, 1363–1371.
- Lugli, G.A., Milani, C., Duranti, S., Alessandri, G., Turroni, F., Mancabelli, L., Tatoni, D., Ossiprandi, M.C., van Sinderen, D., and Ventura, M. (2019). Isolation of novel gut bifidobacteria using a combination of metagenomic and cultivation approaches. *Genome Biol.*

REFERENCES

20, 96.

MacArthur, R.H., and Wilson, E.O. (1967). *The Theory of Island Biogeography*.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3, e1420.

Mallawaarachchi, V., Wickramarachchi, A., and Lin, Y. (2020). GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* 36, 3307–3313.

Mantyla, A.W., and Reid, J.L. (1983). Abyssal characteristics of the World Ocean waters. *Deep Sea Res. Part Oceanogr. Res. Pap.* 30, 805–833.

Marshall, I.P.G., Ren, G., Jaussi, M., Lomstein, B.Aa., Jørgensen, B.B., Røy, H., and Kjeldsen, K.U. (2019). Environmental filtering determines family-level structure of sulfate-reducing microbial communities in subsurface marine sediments. *ISME J.* 13, 1920–1932.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *17*, 3.

Martín-Cuadrado, A.-B., López-García, P., Alba, J.-C., Moreira, D., Monticelli, L., Strittmatter, A., Gottschalk, G., and Rodríguez-Valera, F. (2007). Metagenomics of the Deep Mediterranean, a Warm Bathypelagic Habitat. *PLoS ONE* 2, e914.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-Devine, M.C., Kane, M., Krumins, J.A., Kuske, C.R., et al. (2006). Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* 4, 102–112.

Martiny, J.B.H., Eisen, J.A., Penn, K., Allison, S.D., and Horner-Devine, M.C. (2011). Drivers of bacterial diversity depend on spatial scale. *Proc. Natl. Acad. Sci.* 108, 7850–7854.

Mat, A.M., Sarrazin, J., Markov, G.V., Apremont, V., Dubreuil, C., Eché, C., Fabioux, C., Klopp, C., Sarradin, P.-M., Tanguy, A., et al. (2020). Biological rhythms in the deep-sea hydrothermal mussel *Bathymodiolus azoricus*. *Nat. Commun.* 11, 3454.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618.

McMurdie, P.J., and Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* 8, e61217.

McNichol, J., Berube, P.M., Biller, S.J., and Fuhrman, J.A. (2021). Evaluating and Improving Small Subunit rRNA PCR Primer Coverage for Bacteria, Archaea, and Eukaryotes Using Metagenomes from Global Ocean Surveys. *mSystems* 6, 13.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37, 1530–1534.

Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011). Evaluation of genomic high-

REFERENCES

throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12**, R112.

Mittelbach, G.G., and Schemske, D.W. (2015). Ecological and evolutionary perspectives on community assembly. *Trends Ecol. Evol.* **30**, 241–247.

Molari, M., and Manini, E. (2012). Reliability of CARD-FISH Procedure for Enumeration of Archaea in Deep-Sea Surficial Sediments. *Curr. Microbiol.* **64**, 242–250.

Molari, M., Manini, E., and Dell'Anno, A. (2013). Dark inorganic carbon fixation sustains the functioning of benthic deep-sea ecosystems. *Glob. Biogeochem. Cycles* **27**, 212–221.

Mosier, A.C., Allen, E.E., Kim, M., Ferriera, S., and Francis, C.A. (2012). Genome Sequence of “Candidatus Nitrosopumilus salaria” BD31, an Ammonia-Oxidizing Archaeon from the San Francisco Bay Estuary. *J. Bacteriol.* **194**, 2121–2122.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273.

Murray, A.E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kämpfer, P., Konstantinidis, K.T., Lane, C.E., Papke, R.T., Parks, D.H., et al. (2020). Roadmap for naming uncultivated Archaea and Bacteria. *Nat. Microbiol.* **5**, 987–994.

Nealson, K.H. (1997). SEDIMENT BACTERIA: Who's There, What Are They Doing, and What's New? *Annu. Rev. Earth Planet. Sci.* **25**, 403–434.

Nekola, J.C., and White, P.S. (1999). The distance decay of similarity in biogeography and ecology. *J. Biogeogr.* **26**, 867–878.

Nemergut, D.R., Schmidt, S.K., Fukami, T., O'Neill, S.P., Bilinski, T.M., Stanish, L.F., Knelman, J.E., Darcy, J.L., Lynch, R.C., Wickey, P., et al. (2013). Patterns and Processes of Microbial Community Assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–356.

Newberry, C.J., Webster, G., Cragg, B.A., Parkes, R.J., Weightman, A.J., and Fry, J.C. (2004). Diversity of prokaryotes and methanogenesis in deep subsurface sediments from the Nankai Trough, Ocean Drilling Program Leg 190. *Environ. Microbiol.* **6**, 274–287.

Niggemann, J., Ferdelman, T.G., Lomstein, B.A., Kallmeyer, J., and Schubert, C.J. (2007). How depositional conditions control input, composition, and degradation of organic matter in sediments from the Chilean coastal upwelling region. *Geochim. Cosmochim. Acta* **71**, 1513–1527.

Nøhr Glud, R., Gundersen, J.K., Barker Jørgensen, B., Revsbech, N.P., and Schulz, H.D. (1994). Diffusive and total oxygen uptake of deep-sea sediments in the eastern South Atlantic Ocean: in situ and laboratory measurements. *Deep Sea Res. Part Oceanogr. Res. Pap.* **41**, 1767–1788.

Nunoura, T., Hirai, M., Yoshida-Takashima, Y., Nishizawa, M., Kawagucci, S., Yokokawa, T., Miyazaki, J., Koide, O., Makita, H., Takaki, Y., et al. (2016). Distribution and Niche Separation of Planktonic Microbial Communities in the Water Columns from the Surface to the Hadal Waters of the Japan Trench under the Eutrophic Ocean. *Front. Microbiol.* **7**.

Nunoura, T., Nishizawa, M., Hirai, M., Shimamura, S., Harnvoravongchai, P., Koide, O., Morono, Y., Fukui, T., Inagaki, F., Miyazaki, J., et al. (2018). Microbial Diversity in Sediments

REFERENCES

- from the Bottom of the Challenger Deep, the Mariana Trench. *Microbes Environ.* 33, 186–194.
- Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834.
- Oda, Y., Star, B., Huisman, L.A., Gottschal, J.C., and Forney, L.J. (2003). Biogeography of the Purple Nonsulfur Bacterium *Rhodospseudomonas palustris*. *Appl. Environ. Microbiol.* 69, 5186–5191.
- O'Donnell, J.L., Kelly, R.P., Lowell, N.C., and Port, J.A. (2016). Indexed PCR Primers Induce Template-Specific Bias in Large-Scale DNA Sequencing Studies. *PLOS ONE* 11.
- Offre, P., Spang, A., and Schleper, C. (2013). Archaea in Biogeochemical Cycles. *Annu. Rev. Microbiol.* 67, 437–457.
- Oguri, K., Kawamura, K., Sakaguchi, A., Toyofuku, T., Kasaya, T., Murayama, M., Fujikura, K., Glud, R.N., and Kitazato, H. (2013). Hadal disturbance in the Japan Trench induced by the 2011 Tohoku–Oki Earthquake. *Sci. Rep.* 3, 1915.
- Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. (2015). *vegan: Community Ecology Package*.
- Olsen, G.J., and Woese, C.R. (1993). Ribosomal RNA: a key to phylogeny. *FASEB J.* 7, 113–123.
- Olson, P., Reynolds, E., Hinnov, L., and Goswami, A. (2016). Variation of ocean sediment thickness with crustal age: OCEAN SEDIMENT THICKNESS. *Geochem. Geophys. Geosystems* 17, 1349–1369.
- Orcutt, B.N., Sylvan, J.B., Knab, N.J., and Edwards, K.J. (2011). Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. *Microbiol. Mol. Biol. Rev. MMBR* 75, 361–422.
- Orejas, C., Addamo, A., Alvarez, M., Aparicio, A., Alcoverro, D., Arnaud-Haond, S., Bilan, M., Boavida, J., Cainzos, V., Calderon, R., et al. (2017). Cruise Summary Report - Medwaves Survey (Mediterranean Out Flow Water And Vulnerable Ecosystems) (Zenodo).
- Ortiz-Alvarez, R., and Casamayor, E.O. (2016). High occurrence of *Pacearchaeota* and *Woesearchaeota* (Archaea superphylum DPANN) in the surface waters of oligotrophic high-altitude lakes: Archaeal occurrence in high-altitude lakes. *Environ. Microbiol. Rep.* 8, 210–217.
- Overmann, J., Huang, S., Nübel, U., Hahnke, R.L., and Tindall, B.J. (2019). Relevance of phenotypic information for the taxonomy of not-yet-cultured microorganisms. *Syst. Appl. Microbiol.* 42, 22–29.
- Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2015). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples: Primers for marine microbiome studies. *Environ. Microbiol.* 18, 1403–1414.
- Park, S.-J., Kim, J.-G., Jung, M.-Y., Kim, S.-J., Cha, I.-T., Kwon, K., Lee, J.-H., and Rhee, S.-K. (2012a). Draft Genome Sequence of an Ammonia-Oxidizing Archaeon, “Candidatus Nitrosopumilus koreensis” AR1, from Marine Sediment. *J. Bacteriol.* 194, 6940–6941.

REFERENCES

- Park, S.-J., Kim, J.-G., Jung, M.-Y., Kim, S.-J., Cha, I.-T., Ghai, R., Martín-Cuadrado, A.-B., Rodríguez-Valera, F., and Rhee, S.-K. (2012b). Draft Genome Sequence of an Ammonia-Oxidizing Archaeon, “Candidatus Nitrosopumilus sediminis” AR2, from Svalbard in the Arctic Circle. *J. Bacteriol.* *194*, 6948–6949.
- Parker, C.T., Tindall, B.J., and Garrity, G.M. (2015). International Code of Nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.*
- Parkes, R.J., Cragg, B.A., Bale, S.J., Getliff, J.M., Goodman, K., Rochelle, P.A., Fry, J.C., Weightman, A.J., and Harvey, S.M. (1994). Deep bacterial biosphere in Pacific Ocean sediments. *Nature* *371*, 410–413.
- Parkes, R.J., Webster, G., Cragg, B.A., Weightman, A.J., Newberry, C.J., Ferdelman, T.G., Kallmeyer, J., Jørgensen, B.B., Aiello, I.W., and Fry, J.C. (2005). Deep sub-seafloor prokaryotes stimulated at interfaces over geological time. *Nature* *436*, 390–394.
- Parkes, R.J., Cragg, B., Roussel, E., Webster, G., Weightman, A., and Sass, H. (2014). A review of prokaryotic populations and processes in sub-seafloor sediments, including biosphere:geosphere interactions. *Mar. Geol.* *352*, 409–425.
- Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* *25*, 1043–1055.
- Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* *2*, 1533–1542.
- Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* *14*.
- Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* *38*, 1079–1086.
- Paulson, J.N., Pop, M., and Bravo, H.C. (2013). metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. *Bioconductor Package 1*, 191.
- Pei, A.Y., Oberdorf, W.E., Nossa, C.W., Agarwal, A., Chokshi, P., Gerz, E.A., Jin, Z., Lee, P., Yang, L., Poles, M., et al. (2010). Diversity of 16S rRNA Genes within Individual Prokaryotic Genomes. *Appl. Environ. Microbiol.* *76*, 3886–3897.
- Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* *28*, 1420–1428.
- Peoples, L.M., Donaldson, S., Osuntokun, O., Xia, Q., Nelson, A., Blanton, J., Allen, E.E., Church, M.J., and Bartlett, D.H. (2018). Vertically distinct microbial communities in the Mariana and Kermadec trenches. *PLOS ONE* *13*, e0195102.
- Peoples, L.M., Grammatopoulou, E., Pombrol, M., Xu, X., Osuntokun, O., Blanton, J., Allen, E.E., Nunnally, C.C., Drazen, J.C., Mayor, D.J., et al. (2019). Microbial Community Diversity Within Sediments from Two Geographically Separated Hadal Trenches. *Front. Microbiol.* *10*.

REFERENCES

- Pereira, O., Hochart, C., Auguet, J.C., Debroas, D., and Galand, P.E. (2019). Genomic ecology of Marine Group II, the most common marine planktonic Archaea across the surface ocean. *MicrobiologyOpen* 8.
- Petitjean, C., Deschamps, P., López-García, P., and Moreira, D. (2015). Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. *Genome Biol. Evol.* 7, 191–204.
- Petro, C., Starnawski, P., Schramm, A., and Kjeldsen, K. (2017). Microbial community assembly in marine sediments. *Aquat. Microb. Ecol.* 79, 177–195.
- Petro, C., Zäncker, B., Starnawski, P., Jochum, L.M., Ferdelman, T.G., Jørgensen, B.B., Røy, H., Kjeldsen, K.U., and Schramm, A. (2019). Marine Deep Biosphere Microbial Communities Assemble in Near-Surface Sediments in Aarhus Bay. *Front. Microbiol.* 10.
- Pierre, C. (1999). The oxygen and carbon isotope distribution in the Mediterranean water masses. *Mar. Geol.* 153, 41–55.
- Pischedda, L., Poggiale, J.C., Cuny, P., and Gilbert, F. (2008). Imaging Oxygen Distribution in Marine Sediments. The Importance of Bioturbation and Sediment Heterogeneity. *Acta Biotheor.* 56, 123–135.
- Podar, M., Makarova, K.S., Graham, D.E., Wolf, Y.I., Koonin, E.V., and Reysenbach, A.-L. (2013). Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct* 8, 9.
- Polz, M.F., and Cavanaugh, C.M. (1998). Bias in Template-to-Product Ratios in Multitemplate PCR. *APPL Env. MICROBIOL* 64, 7.
- Probst, A.J., Holman, H.-Y.N., DeSantis, T.Z., Andersen, G.L., Birarda, G., Bechtel, H.A., Piceno, Y.M., Sonnleitner, M., Venkateswaran, K., and Moissl-Eichinger, C. (2013). Tackling the minority: sulfate-reducing bacteria in an archaea-dominated subsurface biofilm. *ISME J.* 7, 635–651.
- Probst, A.J., Ladd, B., Jarett, J.K., Geller-McGrath, D.E., Sieber, C.M.K., Emerson, J.B., Anantharaman, K., Thomas, B.C., Malmstrom, R.R., Stieglmeier, M., et al. (2018). Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nat. Microbiol.* 3, 328–336.
- Props, R., Monsieurs, P., Vandamme, P., Leys, N., Deneff, V.J., and Boon, N. (2019). Gene Expansion and Positive Selection as Bacterial Adaptations to Oligotrophic Conditions. *MSphere* 4, e00011-19, /msphere/4/1/mSphere011-19.atom.
- Pruesse, E., Peplies, J., and Glöckner, F.O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590-596.
- Raes, J., Letunic, I., Yamada, T., Jensen, L.J., and Bork, P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* 7, 473.

REFERENCES

- Ramirez-Llodra, E. (2020). Deep-Sea Ecosystems: Biodiversity and Anthropogenic Impacts. *Law Seabed* 36–60.
- Reysenbach, A.-L., Liu, Y., Banta, A.B., Beveridge, T.J., Kirshtein, J.D., Schouten, S., Tivey, M.K., Damm, K.L.V., and Voytek, M.A. (2006). A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents. *442*, 4.
- Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437.
- Rinke, C., Chuvochina, M., Mussig, A.J., Chaumeil, P.-A., Davin, A.A., Waite, D.W., Whitman, W.B., Parks, D.H., and Hugenholtz, P. (2021). A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat. Microbiol.* 6, 946–959.
- Rommevaux, C., Henri, P., Degboe, J., Chavagnac, V., Lesongeur, F., Godfroy, A., Boulart, C., Destrigneville, C., and Castillo, A. (2019). Prokaryote Communities at Active Chimney and *In Situ* Colonization Devices After a Magmatic Degassing Event (37°N MAR, EMSO-Azores Deep-Sea Observatory). *Geochem. Geophys. Geosystems* 20, 3065–3089.
- Roy, H., Kallmeyer, J., Adhikari, R.R., Pockalny, R., Jorgensen, B.B., and D’Hondt, S. (2012). Aerobic Microbial Respiration in 86-Million-Year-Old Deep-Sea Red Clay. *Science* 336, 922–925.
- Ryan, W.B.F., Carbotte, S.M., Coplan, J.O., O’Hara, S., Melkonian, A., Arko, R., Weissel, R.A., Ferrini, V., Goodwillie, A., Nitsche, F., et al. (2009). Global Multi-Resolution Topography synthesis: GLOBAL MULTI-RESOLUTION TOPOGRAPHY SYNTHESIS. *Geochem. Geophys. Geosystems* 10, n/a-n/a.
- Salazar, G., and Sunagawa, S. (2017). Marine microbial diversity. *Curr. Biol.* 27, R489–R494.
- Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., and Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12, 87.
- Sanford, R.A., Lloyd, K.G., Konstantinidis, K.T., and Löffler, F.E. (2021). Microbial Taxonomy Run Amok. *Trends Microbiol.* 29, 394–404.
- Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* 74, 5463–5467.
- Schauberger, C., Middelboe, M., Larsen, M., Peoples, L.M., Bartlett, D.H., Kirpekar, F., Rowden, A.A., Wenzhöfer, F., Thamdrup, B., and Glud, R.N. (2021a). Spatial variability of prokaryotic and viral abundances in the Kermadec and Atacama Trench regions. *Limnol. Oceanogr.* Ino.11711.
- Schauberger, C., Glud, R.N., Hausmann, B., Trouche, B., Maignien, L., Poulain, J., Wincker, P., Arnaud-Haond, S., Wenzhöfer, F., and Thamdrup, B. (2021b). Microbial community structure in hadal sediments: high similarity along trench axes and strong changes along redox gradients. *ISME J. Multidiscip. J. Microb. Ecol.* 14.
- Schiffries, C.M., Mangum, A.J., Mays, J., Hoon-Starr, M., and Hazen, R. (2019). The Deep Carbon Observatory: An Interdisciplinary Quest to Study Carbon in Earth. In AGU Fall Meeting Abstracts, pp. DI43A-0029.

REFERENCES

- Schippers, A., Neretin, L.N., Kallmeyer, J., Ferdelman, T.G., Cragg, B.A., John Parkes, R., and Jørgensen, B.B. (2005). Prokaryotic cells of the deep sub-seafloor biosphere identified as living bacteria. *Nature* 433, 861–864.
- Schleper, C. (2010). Ammonia oxidation: different niches for bacteria and archaea? *ISME J.* 4, 1092–1094.
- Schratzberger, M., and Ingels, J. (2018). Meiofauna matters: The roles of meiofauna in benthic ecosystems. *J. Exp. Mar. Biol. Ecol.* 502, 12–25.
- Schrenk, M.O., Huber, J.A., and Edwards, K.J. (2010). Microbial Provinces in the Subseafloor. *Annu. Rev. Mar. Sci.* 2, 279–304.
- Schwank, K., Bornemann, T.L.V., Dombrowski, N., Spang, A., Banfield, J.F., and Probst, A.J. (2019). An archaeal symbiont-host association from the deep terrestrial subsurface. *ISME J.* 13, 2135–2139.
- Scoma, A. (2020). Updated definitions on piezophily as suggested by hydrostatic pressure dependence on temperature (*Microbiology*).
- Seiter, K., Hensen, C., and Zabel, M. (2005). Benthic carbon mineralization on a global scale. *Glob. Biogeochem. Cycles* 19.
- Shafiee, R.T., Diver, P.J., Snow, J.T., Zhang, Q., and Rickaby, R.E.M. (2021). Marine ammonia-oxidising archaea and bacteria occupy distinct iron and copper niches. *ISME Commun.* 1, 1.
- Shaiber, A., Willis, A.D., Delmont, T.O., Roux, S., Chen, L.-X., Schmid, A.C., Yousef, M., Watson, A.R., Lolans, K., Esen, Ö.C., et al. (2020). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.* 21, 292.
- Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3, 836–843.
- Sintes, E., Bergauer, K., De Corte, D., Yokokawa, T., and Herndl, G.J. (2013). Archaeal *amo* A gene diversity points to distinct biogeography of ammonia-oxidizing *Crenarchaeota* in the ocean. *Environ. Microbiol.* 15, 1647–1658.
- Soininen, J., McDonald, R., and Hillebrand, H. (2007). The distance decay of similarity in ecological communities. *Ecography* 30, 3–12.
- Soltwedel, T., Bauerfeind, E., Bergmann, M., Bracher, A., Budaeva, N., Busch, K., Cherkasheva, A., Fahl, K., Grzelak, K., Hasemann, C., et al. (2016). Natural variability or anthropogenically-induced variation? Insights from 15 years of multidisciplinary observations at the arctic marine LTER site HAUSGARTEN. *Ecol. Indic.* 65, 89–102.
- Sørensen, K.B., and Teske, A. (2006). Stratified Communities of Active Archaea in Deep Marine Subsurface Sediments. *APPL Env. MICROBIOL* 72, 8.
- Spang, A. (2019). Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* 14.
- Spang, A., Caceres, E.F., and Ettema, T.J.G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* 357.

REFERENCES

- St. John, E., Liu, Y., Podar, M., Stott, M.B., Meneghin, J., Chen, Z., Lagutin, K., Mitchell, K., and Reysenbach, A.-L. (2019). A new symbiotic nanoarchaeote (*Candidatus Nanoclepta minutus*) and its host (*Zestosphaera tikiterensis* gen. nov., sp. nov.) from a New Zealand hot spring. *Syst. Appl. Microbiol.* **42**, 94–106.
- Stackebrandt, E., and Goebel, B.M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int. J. Syst. Evol. Microbiol.* **44**, 846–849.
- Stahl, D.A., and de la Torre, J.R. (2012). Physiology and Diversity of Ammonia-Oxidizing Archaea. *Annu. Rev. Microbiol.* **66**, 83–101.
- Staley, J.T., and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* **39**, 321–346.
- Starnawski, P., Bataillon, T., Ettema, T.J.G., Jochum, L.M., Schreiber, L., Chen, X., Lever, M.A., Polz, M.F., Jørgensen, B.B., Schramm, A., et al. (2017). Microbial community assembly and evolution in subseafloor sediment. *Proc. Natl. Acad. Sci.* **114**, 2940–2945.
- Stegen, J.C., Lin, X., Fredrickson, J.K., Chen, X., Kennedy, D.W., Murray, C.J., Rockhold, M.L., and Konopka, A. (2013). Quantifying community assembly processes and identifying features that impose them. *ISME J.* **7**, 2069–2079.
- Stieglmeier, M., Klingl, A., Alves, R.J.E., Rittmann, S.K.-M.R., Melcher, M., Leisch, N., and Schleper, C. (2014). *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum Thaumarchaeota. *Int. J. Syst. Evol. Microbiol.* **64**, 2738–2752.
- Sun, Y., Liu, Y., Pan, J., Wang, F., and Li, M. (2019). Perspectives on Cultivation Strategies of Archaea. *Microb. Ecol.* **15**.
- Sutcliffe, I.C., Dijkshoorn, L., Whitman, W.B., and Executive Board, on behalf of the I. (2020). Minutes of the International Committee on Systematics of Prokaryotes online discussion on the proposed use of gene sequences as type for naming of prokaryotes, and outcome of vote. *Int. J. Syst. Evol. Microbiol.* **70**, 4416–4417.
- Suzuki, M.T., and Giovannoni, S.J. (1996). Bias Caused by Template Annealing in the Amplification of Mixtures of 16S rRNA Genes by PCR. *APPL ENV. MICROBIOL.* **62**, 6.
- Takai, K., and Horikoshi, K. (1999). Genetic Diversity of Archaea in Deep-Sea Hydrothermal Vent Environments. *Genetics* **152**, 1285–1297.
- Tara Oceans Coordinators, Sunagawa, S., Acinas, S.G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., et al. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445.
- Tatusov, R.L. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36.
- Teske, A.P. (2005). The deep subsurface biosphere is alive and well. *Trends Microbiol.* **13**, 402–404.
- Teske, A., Durbin, A., Ziervogel, K., Cox, C., and Arnosti, C. (2011). Microbial Community Composition and Function in Permanently Cold Seawater and Sediments from an Arctic Fjord

REFERENCES

of Svalbard. *Appl. Environ. Microbiol.* **77**, 2008–2018.

Teske, A., Wegener, G., Chanton, J.P., White, D., MacGregor, B., Hoer, D., de Beer, D., Zhuang, G., Saxton, M.A., Joye, S.B., et al. (2021). Microbial Communities Under Distinct Thermal and Geochemical Regimes in Axial and Off-Axis Sediments of Guaymas Basin. *Front. Microbiol.* **12**, 633649.

Thamdrup, B. (2012). New Pathways and Processes in the Global Nitrogen Cycle. *Annu. Rev. Ecol. Evol. Syst.* **43**, 407–428.

The Genome Standards Consortium, Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731.

Tobias-Hünefeldt, S.P., Wenley, J., Baltar, F., and Morales, S.E. (2020). Ecological drivers switch from bottom-up to top-down during model microbial community successions. *ISME J.*

Topçuoğlu, B.D., Stewart, L.C., Morrison, H.G., Butterfield, D.A., Huber, J.A., and Holden, J.F. (2016). Hydrogen Limitation and Syntrophic Growth among Natural Assemblages of Thermophilic Methanogens at Deep-sea Hydrothermal Vents. *Front. Microbiol.* **7**.

Torti, A., Lever, M.A., and Jørgensen, B.B. (2015). Origin, dynamics, and implications of extracellular DNA pools in marine sediments. *Mar. Genomics* **24**, 185–196.

Torti, A., Jørgensen, B.B., and Lever, M.A. (2018). Preservation of microbial DNA in marine sediments: insights from extracellular DNA pools. *Environ. Microbiol.* **20**, 4526–4542.

Trembath-Reichert, E., Butterfield, D.A., and Huber, J.A. (2019). Active subseafloor microbial communities from Mariana back-arc venting fluids share metabolic strategies across different thermal niches and taxa. *ISME J.* **13**, 2264–2279.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J.I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* **449**, 804–810.

Turnewitsch, R., Falahat, S., Stehlikova, J., Oguri, K., Glud, R.N., Middelboe, M., Kitazato, H., Wenzhöfer, F., Ando, K., Fujio, S., et al. (2014). Recent sediment dynamics in hadal trenches: Evidence for the influence of higher-frequency (tidal, near-inertial) fluid dynamics. *Deep Sea Res. Part Oceanogr. Res. Pap.* **90**, 125–138.

Vellend, M. (2010). Conceptual Synthesis in Community Ecology. *Q. Rev. Biol.* **85**, 183–206.

Vellend, M., Srivastava, D.S., Anderson, K.M., Brown, C.D., Jankowski, J.E., Kleynhans, E.J., Kraft, N.J.B., Letaw, A.D., Macdonald, A.A.M., Maclean, J.E., et al. (2014). Assessing the relative importance of neutral stochasticity in ecological communities. *Oikos* **123**, 1420–1430.

Vetriani, C., Jannasch, H.W., MacGregor, B.J., Stahl, D.A., and Reysenbach, A.-L. (1999). Population Structure and Phylogenetic Characterization of Marine Benthic Archaea in Deep-Sea Sediments. *Appl. Environ. Microbiol.* **65**, 4375–4384.

Vuillemin, A., Wankel, S.D., Coskun, Ö.K., Magritsch, T., Vargas, S., Estes, E.R., Spivack, A.J., Smith, D.C., Pockalny, R., Murray, R.W., et al. (2019). Archaea dominate oxic subseafloor communities over multimillion-year time scales. *Sci. Adv.* **5**, eaaw4108.

REFERENCES

- Vuillemin, A., Vargas, S., Coskun, Ö.K., Pockalny, R., Murray, R.W., Smith, D.C., D'Hondt, S., and Orsi, W.D. (2020). Atribacteria Reproducing over Millions of Years in the Atlantic Abyssal Subseafloor. *MBio* 11.
- Walker, C.B., de la Torre, J.R., Klotz, M.G., Urakawa, H., Pinel, N., Arp, D.J., Brochier-Armanet, C., Chain, P.S.G., Chan, P.P., Gollabgir, A., et al. (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc. Natl. Acad. Sci.* 107, 8818–8823.
- Walsh, E.A., Kirkpatrick, J.B., Rutherford, S.D., Smith, D.C., Sogin, M., and D'Hondt, S. (2016). Bacterial diversity and community composition from seasurface to subseafloor. *ISME J.* 10, 979–989.
- Wang, J., Kan, J., Zhang, X., Xia, Z., Zhang, X., Qian, G., Miao, Y., Leng, X., and Sun, J. (2017). Archaea Dominate the Ammonia-Oxidizing Community in Deep-Sea Sediments of the Eastern Indian Ocean—from the Equator to the Bay of Bengal. *Front. Microbiol.* 8.
- Wang, L.-G., Lam, T.T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C.W., Jones, B.R., Bradley, T., et al. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* 37, 599–603.
- Wang, P., Xiao, X., and Wang, F. (2005). Phylogenetic analysis of Archaea in the deep-sea sediments of west Pacific Warm Pool. *Extremophiles* 9, 209–217.
- Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267.
- Wang, Y., Huang, J.-M., Cui, G.-J., Nunoura, T., Takaki, Y., Li, W.-L., Li, J., Gao, Z.-M., Takai, K., Zhang, A.-Q., et al. (2018). Genomics insights into ecotype formation of ammonia-oxidizing archaea in the deep ocean. *Environ. Microbiol.* 14.
- Wang, Z., Wang, Y., Fuhrman, J.A., Sun, F., and Zhu, S. (2019). Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. *Brief. Bioinform.* 14.
- Watling, L., Guinotte, J., Clark, M.R., and Smith, C.R. (2013). A proposed biogeography of the deep ocean floor. *Prog. Oceanogr.* 111, 91–112.
- Watson, J.D., and Crick, F.H.C. (1953). A structure for Deoxyribose Nucleic Acid. *Nature* 171, 737–738.
- Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., et al. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int. J. Syst. Bacteriol.* 37, 463–464.
- Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K.F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, 100.
- Wellsbury, P., Goodman, K., Barth, T., Cragg, B.A., Barnes, S.P., and Parkes, R.J. (1997). Deep marine biosphere fuelled by increasing organic matter availability during burial and heating. *Nature* 388, 573–576.

REFERENCES

- Wellsbury, P., Mather, I., and Parkes, R.J. (2002). Geomicrobiology of deep, low organic carbon sediments in the Woodlark Basin, Pacific Ocean. *FEMS Microbiol. Ecol.* 42, 59–70.
- Wenzhöfer, F., Oguri, K., Middelboe, M., Turnewitsch, R., Toyofuku, T., Kitazato, H., and Glud, R.N. (2016). Benthic carbon mineralization in hadal trenches: Assessment by in situ O₂ microprofile measurements. *Deep Sea Res. Part Oceanogr. Res. Pap.* 116, 276–286.
- Whitaker, R.J. (2003). Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea. *Science* 301, 976–978.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York).
- Willis, C., Desai, D., and LaRoche, J. (2019). Influence of 16S rRNA variable region on perceived diversity of marine microbial communities of the Northern North Atlantic. *FEMS Microbiol. Lett.* 366, fnz152.
- Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* 74, 5088–5090.
- Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* 87, 4576–4579.
- Wörmer, L., Hoshino, T., Bowles, M.W., Viehweger, B., Adhikari, R.R., Xiao, N., Uramoto, G., Könneke, M., Lazar, C.S., Morono, Y., et al. (2019). Microbial dormancy in the marine subsurface: Global endospore abundance and response to burial. *Sci. Adv.* 5, eaav1024.
- Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607.
- Wurch, L., Giannone, R.J., Belisle, B.S., Swift, C., Utturkar, S., Hettich, R.L., Reysenbach, A.-L., and Podar, M. (2016). Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* 7, 1–10.
- Xu, D., Liu, S., Chen, Q., and Ni, J. (2017). Microbial community compositions in different functional zones of Carrousel oxidation ditch system for domestic wastewater treatment. *AMB Express* 7, 40.
- Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2016). *ggtree*: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*
- Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541, 353–358.
- Zeng, X., Birrien, J.-L., Fouquet, Y., Cherkashov, G., Jebbar, M., Querellou, J., Oger, P., Cambon-Bonavita, M.-A., Xiao, X., and Prieur, D. (2009). *Pyrococcus* CH1, an obligate piezophilic hyperthermophile: extending the upper pressure-temperature limits for life. *ISME J.* 3, 873–876.
- Zhang, S., Song, W., Wemheuer, B., Reveillaud, J., Webster, N., and Thomas, T. (2019). Comparative Genomics Reveals Ecological and Evolutionary Insights into Sponge-Associated *Thaumarchaeota*. *MSystems* 4, e00288-19, /msystems/4/4/msys.00288-19.atom.

REFERENCES

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.* 7, 203–214.

Zhao, R., Dahle, H., Ramírez, G.A., and Jørgensen, S.L. (2020). Indigenous Ammonia-Oxidizing Archaea in Oxidic Subseafloor Oceanic Crust. *MSystems* 5.

Zhong, H., Lehtovirta-Morley, L., Liu, J., Zheng, Y., Lin, H., Song, D., Todd, J.D., Tian, J., and Zhang, X.-H. (2020). Novel insights into the Thaumarchaeota in the deepest oceans: their metabolism and potential adaptation mechanisms. *Microbiome* 8, 78.

Zhou, J., and Ning, D. (2017). Stochastic Community Assembly: Does It Matter in Microbial Ecology? *Microbiol. Mol. Biol. Rev.* 81.

Zinger, L., Amaral-Zettler, L.A., Fuhrman, J.A., Horner-Devine, M.C., Huse, S.M., Welch, D.B.M., Martiny, J.B.H., Sogin, M., Boetius, A., and Ramette, A. (2011). Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems. *PLoS ONE* 6, e24570.

Zinger, L., Boetius, A., and Ramette, A. (2014). Bacterial taxa-area and distance-decay relationships in marine environments. *Mol. Ecol.* 23, 954–964.

SUPPLEMENTARY MATERIAL

Supplementary material for Chapter 1

1. Amplicon libraries preparation and sequencing

1.1. 16S-V4V5 rRNA gene amplicon generation

Prokaryotic barcodes were generated using the 515F-Y/926R (Parada et al., 2015) and 517F/958R (Topçuoğlu et al., 2016) primers, and the *Phusion* High Fidelity PCR Master Mix with GC buffer (ThermoFisher Scientific, Waltham, MA, USA). PCR mixtures (25 µL final volume) contained 2.5 ng or less of DNA template with 0.4 µM concentration of each primer, 3% of DMSO, and 1X *Phusion* Master Mix. PCR amplifications (98°C for 30 s; 25 cycles of 10 s at 98°C, 30 s at 53°C, 30 s at 72°C; and 72°C for 10 min) of all samples were carried out in triplicate in order to smooth the intra-sample variance while obtaining sufficient amounts of amplicons for Illumina sequencing.

PCR triplicates were pooled and cleaned up using 1X AMPure XP beads (Beckman Coulter, Brea, CA, USA). Aliquots of purified amplicons were then run on an Agilent Bioanalyzer using the DNA High Sensitivity LabChip kit (Agilent Technologies, Santa Clara, CA, USA) to check their lengths and quantified with a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA).

1.2. Amplicon library preparation

One hundred ng of amplicons were directly end-repaired, A-tailed at the 3' end, and ligated to Illumina compatible adaptors using the NEBNext DNA Modules Products (New England Biolabs, MA, USA) and NextFlex DNA barcodes (Bioo Scientific Corporation) with a liquid handler. This was done on a Biomek FX Laboratory Automation Workstation (Beckmann Coulter Genomics), able to perform up to 96 reactions in parallel. After two consecutive 1x AMPure XP clean ups, the ligated products were amplified using Kapa Hifi HotStart NGS

SUPPLEMENTARY MATERIAL

library Amplification kit (Kapa Biosystems, Wilmington, MA), followed by 1x AMPure XP purification.

1.3. Sequencing library quality control

Libraries were quantified by Quant-iT dsDNA HS assay kits using a Fluoroskan Ascent microplate fluorometer (ThermoFisher Scientific, Waltham, MA, USA) and then by qPCR with the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems, Wilmington, MA, USA) on a MxPro instrument (Agilent Technologies, Santa Clara, CA, USA). Library profiles were assessed using a high-throughput microfluidic capillary electrophoresis system (LabChip GX, Perkin Elmer, Waltham, MA, USA).

1.4. Amplicon sequencing procedures

Metabarcoding libraries were characterized by low diversity sequences at the beginning of the reads, partly due to the presence of the primer sequence used to amplify tags. Low-diversity libraries can interfere in correct cluster identification, resulting in a drastic loss of data output. Therefore, loading concentrations of the metabarcoding libraries were normalized to 8–9 pM (instead of 12–14 pM for standard libraries) and contained a 20% PhiX DNA spike-in (instead of 1%) in order to minimize the impacts on the run quality. Libraries were sequenced on HiSeq2500 instruments (Illumina, San Diego, CA, USA) in a 250 bp paired-end mode.

2. Metagenomic libraries preparation and sequencing

2.1. Sequencing library preparation

According to the relatively low DNA quantities extracted, 10 ng or less of genomic DNA were sonicated and the NEBNext Ultra II DNA Library prep kit for Illumina was manually applied. Fragments were end-repaired, 3'-adenylated and NEXTflex DNA barcoded adaptors were added by using NEBNext Ultra II DNA Library prep kit for Illumina (New England Biolabs, MA,

SUPPLEMENTARY MATERIAL

USA). After two consecutive 1x AMPure clean ups, the ligated products were PCR-amplified with NEBNext® Ultra II Q5 Master Mix included in the kit, followed by 0.8x AMPure XP purification.

2.2. Sequencing library quality control

All libraries were quantified first by Quant-it dsDNA HS using a Fluoroskan Ascent instrument (ThermoFisher Scientific, Waltham, MA, USA) then 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). Library profiles were assessed using a high throughput microfluidic capillary electrophoresis system (LabChip GX, Perkin Elmer, Waltham, MA).

2.3. Sequencing procedures

Library concentrations were normalized to 10 nM by addition of Tris-Cl 10 mM (pH 8.5) and applied to cluster generation according to the Illumina Cbot User Guide (Part # 15006165). Sequencing of libraries was performed according to the Novaseq 6000 System User Guide (Part # 20023471) in a paired-end mode using a read length of 150 bp.

Supplementary material for Chapter 2

1. PCR amplification

Prokaryotic barcodes were generated using the 515F-Y (5'- GTGYCAGCMGCCGCGGTAA-3') and 926R (5'- CCGYCAATTYMTTTRAGTTT-3') primers (Parada et al., 2015), and the Phusion High Fidelity PCR Master Mix with GC buffer (ThermoFisher Scientific, Waltham, MA, USA). PCR mixtures (25 µL final volume) contained 2.5 ng or less of DNA template with 0.4 µM concentration of each primer, 3% of DMSO, and 1X Phusion Master Mix. PCR amplifications (98°C for 30 s; 25 cycles of 10 s at 98°C, 30 s at 53°C, 30 s at 72°C; and 72°C for 10 min) of all samples were carried out in triplicate in order to smooth the intra-sample variance while obtaining sufficient amounts of amplicons for Illumina sequencing.

PCR triplicates were pooled and cleaned up using 1X AMPure XP beads (Beckman Coulter, Brea, CA, USA). Aliquots of purified amplicons were then run on an Agilent Bioanalyzer using the DNA High Sensitivity LabChip kit (Agilent Technologies, Santa Clara, CA, USA) to check their lengths and quantified with a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA).

2. Sequencing

2.1. Amplicon library preparation

One hundred ng of amplicons were directly end-repaired, A-tailed at the 3' end, and ligated to Illumina compatible adaptors using the NEBNext DNA Modules Products (New England Biolabs, MA, USA) and NextFlex DNA barcodes (Bioo Scientific Corporation) with a liquid handler. This was done on a Biomek FX Laboratory Automation Workstation (Beckmann Coulter Genomics), able to perform up to 96 reactions in parallel. After two consecutive 1x AMPure XP clean ups, the ligated product were amplified using Kapa Hifi HotStart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA), followed by 1x AMPure XP purification.

SUPPLEMENTARY MATERIAL

2.2. Sequencing library quality control

Libraries were quantified by Quant-iT dsDNA HS assay kits using a Fluoroskan Ascent microplate fluorometer (ThermoFisher Scientific, Waltham, MA, USA) and then by qPCR with the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems, Wilmington, MA, USA) on a MxPro instrument (Agilent Technologies, Santa Clara, CA, USA). Library profiles were assessed using a high-throughput microfluidic capillary electrophoresis system (LabChip GX, Perkin Elmer, Waltham, MA, USA).

2.3. Sequencing procedures

Metabarcoding libraries were characterized by low diversity sequences at the beginning of the reads, partly due to the presence of the primer sequence used to amplify tags. Low-diversity libraries can interfere in correct cluster identification, resulting in a drastic loss of data output. Therefore, loading concentrations of the metabarcoding libraries were normalized to 8–9 pM (instead of 12–14 pM for standard libraries) and contained a 20% PhiX DNA spike-in (instead of 1%) in order to minimize the impacts on the run quality. Libraries were sequenced on HiSeq2500 instruments (Illumina, San Diego, CA, USA) in a 250 bp paired-end mode.

3. Sediment characterization

3.1. Granulometric distribution

To determine granulometry of the sediments, the samples were processed using a Malvern Mastersizer 3000 and Hydro LV (Malvern Panalytical Ltd, Malvern, UK). For each sample, a spatula tip of matter was added to water and submitted to ultrasound treatment for 30 seconds at 100% of their power to break aggregates. The sample was then agitated for 30 seconds without ultrasounds in order to stabilize it. Following this, at least 4 granulometric measurements were performed in water, under a 2000 rpm agitation, with the following

SUPPLEMENTARY MATERIAL

parameters: Mie theory with a refractive index of 1.52 and absorption of 0.1, and an obscuration rate between 0.5 and 15%. The result kept was the most repeatable value obtained. Between each sample, the machine underwent an automatic cleaning process.

3.2. Humidity level and loss on ignition at 550°C

Approximately 2g of sediments were placed into dry and clean crucibles that had been previously weighed. The filled crucibles were then weighed before and after being placed overnight in a 100°C oven. Finally they were placed in a 550°C oven for 4 hours and weighed once more. Based on the weights measured, humidity level and loss on ignition values were computed.

Supplementary material for Chapter 4

Table S3: Metagenome information: sequence number and co-assembly groups.

Genoscope code	Exploitable sequence number	Sample name	Coassembly Group
CCT_AADQOSDA_4_H5YGCDSXX.12BA193	120425533	HADES_Kermadec_ST6_CT7_S_0-1	HKT
CCT_AAHXOSDA_4_H5YGCDSXX.12BA218	144357422	HADES_Kermadec_ST6_CT7_S_1-3	HKT
CCT_AASOOSDA_1_H7VT7DSXX.12BA198	146627799	HADES_Kermadec_ST6_CT7_S_10-15	HAK
CCT_AAVNOSDA_2_H7VT7DSXX.12BA223	186248267	HADES_Kermadec_ST6_CT7_S_15-30	HAK
CCT_AALSOSDA_1_H7VT7DSXX.12BA243	126433878	HADES_Kermadec_ST6_CT7_S_3-5	HKT
CCT_AAPGOSDA_1_H7VT7DSXX.12BA268	140784408	HADES_Kermadec_ST6_CT7_S_5-10	HKT
CCT_AADROSDA_4_H5YGCDSXX.12BA205	197116916	HADES_Kermadec_ST7_CT5_S_0-1	AK7
CCT_AAHYOSDA_4_H5YGCDSXX.12BA230	152551333	HADES_Kermadec_ST7_CT5_S_1-3	AK7
CCT_AASPOSDA_1_H7VT7DSXX.12BA210	205401756	HADES_Kermadec_ST7_CT5_S_10-15	AK7
CCT_AAVOOSDA_2_H7VT7DSXX.12BA235	139257884	HADES_Kermadec_ST7_CT5_S_15-30	AK7
CCT_AALTOSDA_1_H7VT7DSXX.12BA255	132489557	HADES_Kermadec_ST7_CT5_S_3-5	AK7
CCT_AAPHOSDA_1_H7VT7DSXX.12BA280	142603331	HADES_Kermadec_ST7_CT5_S_5-10	AK7
CCT_AAGLOSDA_4_H5YGCDSXX.12BA217	140624702	So261_Site 10_CT1_0_1	HAS
CCT_AAKCOSDA_4_H5YGCDSXX.12BA242	140447751	So261_Site 10_CT1_1_3	HAS
CCT_AAUPOSDA_2_H7VT7DSXX.12BA222	185040843	So261_Site 10_CT1_10_15	MD
CCT_AAWPOSDA_2_H7VT7DSXX.12BA247	157159132	So261_Site 10_CT1_15_30	MD
CCT_AANYOSDA_1_H7VT7DSXX.12BA267	131243022	So261_Site 10_CT1_3_5	HAK
CCT_AARLOSDA_1_H7VT7DSXX.12BA197	150800789	So261_Site 10_CT1_5_10	HAK
CCT_AAGROSDA_4_H5YGCDSXX.12BA229	116086397	So261_Site 3_CT1_0_1	HAS
CCT_AAKIOSDA_4_H5YGCDSXX.12BA254	164417092	So261_Site 3_CT1_1_3	H3T
CCT_AAUVOSDA_2_H7VT7DSXX.12BA234	210676544	So261_Site 3_CT1_10_15	H3D
CCT_AAWVOSDA_2_H7VT7DSXX.12BA259	210253317	So261_Site 3_CT1_15_30	H3D
CCT_AAOEOSDA_1_H7VT7DSXX.12BA279	172364443	So261_Site 3_CT1_3_5	H3T
CCT_AARROSDA_1_H7VT7DSXX.12BA209	134804636	So261_Site 3_CT1_5_10	H3D
CCT_AAGSOSDA_4_H5YGCDSXX.12BA241	121290647	So261_Site 3_CT2_0_1	HAS
CCT_AAKJOSDA_4_H5YGCDSXX.12BA266	163859413	So261_Site 3_CT2_1_3	H3T
CCT_AAUWOSDA_2_H7VT7DSXX.12BA246	209171901	So261_Site 3_CT2_10_15	H3D
CCT_AAWWOSDA_2_H7VT7DSXX.12BA271	190688026	So261_Site 3_CT2_15_30	H3D
CCT_AAOFOSDA_1_H7VT7DSXX.12BA196	141998402	So261_Site 3_CT2_3_5	H3T
CCT_AARSOSDA_1_H7VT7DSXX.12BA221	140640390	So261_Site 3_CT2_5_10	H3D
CCT_AAGTOSDA_4_H5YGCDSXX.12BA253	179320302	So261_Site 3_CT3_0_1	HAS
CCT_AAKKOSDA_4_H5YGCDSXX.12BA278	124729554	So261_Site 3_CT3_1_3	H3T
CCT_AAUXOSDA_2_H7VT7DSXX.12BA258	225070322	So261_Site 3_CT3_10_15	H3D
CCT_AAWXOSDA_2_H7VT7DSXX.12BA283	143495570	So261_Site 3_CT3_15_30	H3D
CCT_AAOGOSDA_1_H7VT7DSXX.12BA208	140501796	So261_Site 3_CT3_3_5	H3T

SUPPLEMENTARY MATERIAL

CCT_AARTOSDA_1_H7VT7DSXX.12BA233	128696046	So261_Site 3_CT3_5_10	H3D
CCT_AAHDOSDA_4_H5YGCDSSXX.12BA265	135359984	So261_Site 7_CT1_0_1	A7S
CCT_AAKUOSDA_4_H5YGCDSSXX.12BA195	131504517	So261_Site 7_CT1_1_3	A7S
CCT_AAVHOSDA_2_H7VT7DSXX.12BA270	198248813	So261_Site 7_CT1_10_15	A7D
		So261_Site 7_CT1_15_30	library failed
CCT_AAOQOSDA_1_H7VT7DSXX.12BA220	150571592	So261_Site 7_CT1_3_5	A7D
CCT_AASDOSDA_2_H7VT7DSXX.12BA245	203963265	So261_Site 7_CT1_5_10	A7D
CCT_AAHEOSDA_4_H5YGCDSSXX.12BA277	127667716	So261_Site 7_CT2_0_1	A7S
CCT_AAKVOSDA_4_H5YGCDSSXX.12BA207	122196082	So261_Site 7_CT2_1_3	A7S
		So261_Site 7_CT2_10_15	library failed
		So261_Site 7_CT2_15_30	library failed
CCT_AAOROSDA_1_H7VT7DSXX.12BA232	143613871	So261_Site 7_CT2_3_5	A7D
CCT_AASEOSDA_2_H7VT7DSXX.12BA257	219040563	So261_Site 7_CT2_5_10	A7D
CCT_AAHFOSDA_4_H5YGCDSSXX.12BA194	259670853	So261_Site 7_CT3_0_1	A7S
CCT_AAKWOSDA_4_H5YGCDSSXX.12BA219	149104779	So261_Site 7_CT3_1_3	A7S
CCT_AAVJOSDA_2_H7VT7DSXX.12BA199	173121549	So261_Site 7_CT3_10_15	A7D
		So261_Site 7_CT3_15_30	library failed
CCT_AAOSOSDA_1_H7VT7DSXX.12BA244	147362285	So261_Site 7_CT3_3_5	A7D
CCT_AASFOSDA_1_H7VT7DSXX.12BA269	141603124	So261_Site 7_CT3_5_10	A7D
CCT_AAHGOSDA_4_H5YGCDSSXX.12BA206	131406290	So261_Site 9_CT1_0_1	A9S
CCT_AAKXOSDA_4_H5YGCDSSXX.12BA231	153632844	So261_Site 9_CT1_1_3	A9S
CCT_AAVKOSDA_2_H7VT7DSXX.12BA211	192971942	So261_Site 9_CT1_10_15	MD
CCT_AAXKOSDA_2_H7VT7DSXX.12BA236	175179454	So261_Site 9_CT1_15_30	MD
CCT_AAOTOSDA_1_H7VT7DSXX.12BA256	130767932	So261_Site 9_CT1_3_5	A9S
CCT_AASGOSDA_1_H7VT7DSXX.12BA281	207971201	So261_Site 9_CT1_5_10	MD

SUPPLEMENTARY MATERIAL

Table S4: MAG information : length, GC content, completion, redundancy, taxonomy

Bins	Total length	Number contigs	N50	GC content	% comp	% red	Domain	Phylum	Class	Order	Family	Genus	amoA gene
AK7_Bin_00007	1197703	125	13187	33,94	93,42	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
H3D_Bin_00022	1333722	57	39922	32,11	89,47	0	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales			no
H3T_Bin_00024	1311402	127	17006	33,4	88,16	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		yes
HKT_Bin_00022	1048028	106	14070	34,58	86,84	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		yes
HKT_Bin_00027	1070626	162	8072	33,15	85,53	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A9S_Bin_00005	1138102	129	11586	34,7	84,21	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
HAS_Bin_00039	1114140	241	4799	34,42	84,21	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		yes
A7D_Bin_00024	960883	50	25507	33,89	78,95	0	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		yes
A7S_Bin_00015	915437	191	5013	31,63	78,95	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
AK7_Bin_00037	1004026	140	8972	33,09	75	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
H3T_Bin_00065	943479	229	4256	33,29	73,68	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7D_Bin_00052	907596	109	9868	34,03	71,05	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
H3D_Bin_00109	1396043	255	6088	43,48	71,05	0	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales			no
A7S_Bin_00027	715505	170	4031	31,59	67,11	0	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
HAK_Bin_00079	638897	174	3621	32,36	64,47	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7D_Bin_00065	723063	129	6182	33,52	63,16	3,95	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7S_Bin_00045	907537	218	4056	43,83	63,16	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales			no
A9S_Bin_00032	852889	178	5036	34,28	63,16	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
AK7_Bin_00057	823154	218	3756	34,74	63,16	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7D_Bin_00064	707323	135	5678	44,17	61,84	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales			no
A7D_Bin_00075	653153	21	43385	35,7	60,53	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		yes
A9S_Bin_00033	638039	188	3274	33,9	60,53	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
HKT_Bin_00058	730438	124	6520	32,47	60,53	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7D_Bin_00083	2483871	486	5156	33,51	57,89	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
AK7_Bin_00065	773638	166	4992	34,47	57,89	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
AK7_Bin_00081	861529	219	3889	33,97	56,58	7,89	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7S_Bin_00056	637572	158	3823	32,71	55,26	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no

SUPPLEMENTARY MATERIAL

HAS_Bin_00099	592642	76	9266	31,93	55,26	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
HKT_Bin_00075	739489	159	5185	32,51	55,26	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
HKT_Bin_00076	758361	186	4233	31,95	53,95	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7D_Bin_00092	722766	147	5084	33,94	52,63	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7S_Bin_00081	1126597	313	3443	34,14	52,63	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
H3T_Bin_00167	792049	215	3619	32,58	52,63	7,89	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	yes
HAS_Bin_00095	617621	128	4992	32,03	52,63	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	yes
MD_Bin_00050	659104	156	4140	34,03	52,63	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A9S_Bin_00040	632316	160	3961	43,85	51,32	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales			no
H3D_Bin_00215	596262	167	3562	32,66	51,32	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	yes
HAK_Bin_00119	618817	157	4193	32,41	51,32	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7D_Bin_00094	575220	27	24228	46,38	50	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales			no
A7S_Bin_00098	704346	171	4081	33,86	47,37	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A9S_Bin_00058	817430	215	3605	33,13	47,37	10,53	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
AK7_Bin_00110	766999	226	3263	32,95	47,37	10,53	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
AK7_Bin_00091	455968	91	5495	33,08	44,74	0	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7S_Bin_00119	809768	235	3278	33,53	39,47	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7S_Bin_00118	376681	111	3365	34,33	38,16	1,32	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		yes
MD_Bin_00109	405265	125	3253	33,22	36,84	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7D_Bin_00161	764758	211	3632	33,1	35,53	6,58	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7S_Bin_00141	349449	100	3573	34,25	35,53	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
AK7_Bin_00136	775050	206	3652	33,55	35,53	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
AK7_Bin_00137	421231	128	3072	33,24	35,53	5,26	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	no
A7D_Bin_00152	382324	82	5310	34,37	32,89	2,63	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7D_Bin_00162	351894	81	4490	34,28	32,89	3,95	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae		no
A7S_Bin_00182	1146398	309	3443	32,42	27,63	9,21	Archaea	Crenarchaeota	Nitrososphaeria	Nitrososphaerales	Nitrosopumilaceae	Nitrosopumilus	yes

Titre : Exploration des Bactéries et des Archées des environnements marins profonds, de la structure des communautés à la génomique comparative

Mots clés : écologie microbienne, biogéographie, sédiments benthiques, fosses hadales, métagénomique

Résumé : Les sédiments marins recouvrent environ 65% de la surface terrestre et les microorganismes qui les peuplent jouent un rôle essentiel dans les cycles biogéochimiques marins. Situés à l'interface entre les communautés pélagiques et de subsurface, les Bactéries et Archées benthiques déterminent la partition entre enfouissement de la matière organique et nutriments relargués dans la colonne d'eau. Comprenant une vaste diversité de microorganismes et des adaptations fonctionnelles spécifiques, elles sont encore peu décrites. Dans le cadre du projet « Pourquoi pas les abysses ? », cette thèse s'est intéressée à la structure et la diversité fonctionnelle des communautés microbiennes benthiques des grands fonds.

Dans ce but, nous avons mis en place des méthodes standardisées d'échantillonnage, d'extraction d'ADN et d'analyse bioinformatique. A l'aide de données de métabarcoding 16S, nous avons étudié la biogéographie des communautés de la transition entre Méditerranée et Atlantique, et observé une

importante influence de la limitation de dispersion et de la dérive écologique, de façon longitudinale et verticale.

Dans les sédiments de surface de deux fosses hadales du Pacifique Sud, la distribution des classes d'Archées dominantes, Nitrososphaeria et Nanoarchaea est influencée par la profondeur et l'horizon sédimentaire, avec plusieurs partenaires putatifs pour la lignée présumée symbiotique des Woeseearchaeales. A l'aide de données métagénomiques, nous avons reconstruit 90 MAGs d'Archées des mêmes sédiments, dont 53 affiliés aux Nitrososphaeria dont la variabilité génomique semble liée à la niche écologique.

Dans l'ensemble, les résultats obtenus posent de solides bases pour la caractérisation de la diversité fonctionnelle et des adaptations spécifiques des communautés microbiennes des sédiments marins profonds.

Title : Exploring the deep ocean seafloor Bacteria and Archaea, from microbial community structure to comparative genomics

Keywords : microbial ecology, biogeography, benthic sediments, hadal trenches, metagenomics

Abstract : Marine sediments cover around 65% of the Earth's surface and microorganisms inhabiting these environments play an essential role in marine biogeochemical cycles. Located at the interface between pelagic and subsurface communities, benthic Bacteria and Archaea are responsible for the early diagenesis of sinking organic matter and determine the partitioning between buried organic matter and nutrients released in the water column. Likely encompassing a broad range of unique biodiversity and functional adaptations, they are still sparsely described. As part of the "Pourquoi pas les Abysses ?" project, this thesis endeavored to shed light on the structure and functional diversity of deep-sea benthic microbial communities.

To this end, we implemented large-scale standardized methods of sampling, DNA extraction, and bioinformatic processing to recover environmental DNA data. Firstly, using 16S rRNA amplicon sequencing, we investigated the biogeographic patterns at the transition between

Mediterranean Sea and Atlantic Ocean, and observed important influences of dispersal limitation and drift, both longitudinally and vertically.

In the surface sediments of two South Pacific hadal trenches, the dominant archaeal classes, Nitrososphaeria and Nanoarchaea, were partitioned following depth zones and sediment horizon, with a variety of putative partners for the presumed symbiotic Woeseearchaeales lineage. Using metagenomic sequencing and analysis on the same hadal samples, 90 archaeal MAGs were reconstructed from the corresponding hadal metagenomes, including 53 affiliated with Nitrososphaeria. We studied their clade-level distribution and showed differing patterns of genomic variability between ecological niches. Overall, these results lay the foundations for continued investigation of the functional diversity and adaptation of deep-sea sediment microbial communities.