A flexible pipeline combining clustering and correction tools for prokaryotic and eukaryotic metabarcoding

Short title:

A flexible metabarcoding pipeline based on read correction

Miriam I. Brandt[1], Blandine Trouche[2], Laure Quintric[3], Patrick Wincker[4,5], Julie Poulain[4,5], and Sophie Arnaud-Haond[1]

[1]MARBEC, Ifremer, Univ. Montpellier, IRD, CNRS, Sète, France

[2] Univ. Brest, CNRS, Ifremer, Laboratoire de Microbiologie des Environnements Extrêmes, Plouzané, France

[3]Ifremer, Cellule Bioinformatique, Brest, France

[4] Génomique Métabolique, Génoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France

[5] Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/ Tara

Corresponding author: sarnaud@ifremer.fr, miriam.isabelle.brandt@gmail.com,

ABSTRACT

Environmental metabarcoding is an increasingly popular tool for studying biodiversity in marine and terrestrial biomes. With sequencing costs decreasing, multiple-marker metabarcoding, spanning several branches of the tree of life, is becoming more accessible. 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 Illumina 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 v2, a network-based clustering algorithm, and to further curate the ASVs/OTUs based on sequence similarity and co-occurrence rates using a recently developed algorithm, LULU. Finally, flexible taxonomic assignment was implemented *via* Ribosomal Database Project (RDP) Bayesian classifier and BLAST. We validate this pipeline with ribosomal and mitochondrial markers using eukaryotic mock communities and 42 deep-sea sediment samples. The results show that ASVs, reflecting genetic diversity, may not be appropriate for alpha diversity estimation of organisms fitting the biological species concept. The results underline the advantages of clustering and LULU-curation for producing more reliable metazoan biodiversity inventories, and show that LULU is an effective tool for filtering metazoan molecular clusters, although the minimum identity threshold applied to co-occurring OTUs has to be increased for 18S. The comparison of BLAST and the RDP Classifier underlined the potential of the latter to deliver very good assignments, but highlighted the need for a concerted effort to build comprehensive, ecosystem-specific, databases adapted to the studied communities.

Key words: Biodiversity, bioinformatics, environmental DNA, metabarcoding, mock communities, eukaryotes (18S and COI), prokaryotes (16S)

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, non-invasive tools, allowing detection of a wide range of taxa in a rapid, cost-effective way using a variety of sample types (Valentini et al. 2009; Taberlet et al. 2012; Creer et al. 2016; Stat et al. 2017). In the last decade, these tools have been used to describe past and present biodiversity in terrestrial (Ji et al. 2013; Yoccoz et al. 2012; Yu et al. 2012; Slon et al. 2017; Pansu et al. 2015), freshwater (Valentini et al. 2016; Deiner et al. 2016; Bista et al. 2015; Dejean et al. 2011; Evans et al. 2016), and marine (Fonseca et al. 2010; Sinniger et al. 2016; Pawlowski et al. 2011; Massana et al. 2015; De Vargas et al. 2015; Salazar et al. 2016; Boussarie et al. 2018; Bik et al. 2012) 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 Fig.1 in Fonseca et al, 2018). Recent studies have addressed many issues regarding sampling methods (Dickie et al. 2018), contamination risks (Goldberg et al. 2016), DNA extraction protocols (Brannock and Halanych 2015; Deiner et al. 2015; Zinger et al. 2016), amplification biases and required PCR replication levels (Nichols et al. 2018; Alberdi et al. 2017; Ficetola et al. 2015). 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 (Coissac et al. 2012; Bokulich et al.

2013). 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 and Halanych 2015; Clare et al. 2016; Brown et al. 2015; Xiong and 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 and sequencing errors (and to some extent intra-individual variability linked to the existence of 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/haplotype at the same locus) that would best fit a “species level”, i.e. the Operational Taxonomic Units defined using a classical phenetic *proxy* (Sokal and Crovello 1970). Recent bioinformatic algorithms alleviate the influence of errors and intraspecific variability in metabarcoding datasets. 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 fine-scale 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 intra-individual variability (pseudogenes, heteroplasmy), based on their similarity and co-occurrence rate with more abundant clusters, allowing obtaining curated datasets 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). However, clustering sequences may still be necessary in metazoan datasets, 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 the latter are not necessarily the most abundant ones (Bazin et al. 2006). Such bias in biodiversity inventories based on ASVs is likely to be magnified in presence-absence metabarcoding datasets, commonly used for metazoan communities (Ji et al. 2013). Similarly, imposing a "universal" clustering threshold on metabarcoding datasets is also introducing bias, penalizing groups with lower interspecific divergence, and overestimating species diversity in groups with higher interspecific divergence. However, this can be alleviated with tools such as swarm v2, a single-linkage clustering algorithm (Mahe et al. 2015). Based on network theory, swarm v2 aggregates sequences iteratively and locally around seed sequences and determines coherent groups of sequences, independent of amplicon input order, allowing highly scalable and fine-scale clustering. Finally, it is widely recognized that homogeneous entities sharing a set of evolutionary and ecological properties, i.e. *species* (Mayr 1942; de Queiroz 2005), sometimes referred to "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. Maintaining ASV information for feeding databases and cross-comparing studies is not incompatible with their clustering into OTUs, and this choice depends on the purpose of the study, i.e. 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 evaluate 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 (Leray et al. 2013) and the V1-V2 region of 18S ribosomal RNA (rRNA) (Sinniger et al. 2016), 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 test the different bioinformatic tools on a deep-sea sediment dataset in order to select an optimal trade-off between inflating biodiversity estimates and loosing 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 natural samples (Parada et al. 2016).

Our objectives were to (1) discuss the use of ASV *vs* OTU-centred datasets depending on taxonomic compartment and study objectives, and (2) determine the most adequate swarm-clustering and LULU curation thresholds that avoid inflating biodiversity estimates while retaining rare biodiversity.

1 MATERIALS AND METHODS

1.1 Preparation of samples

Mock communities

Genomic-DNA mass-balanced metazoan mock communities (5 ng/µL) were prepared using standardized 10 ng/µL DNA extracts of ten 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 $MgCl_2$ 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, Waltham, MA, USA) and sent to Eurofins (Eurofins Scientific, Luxembourg) for Sanger sequencing. The barcode sequences obtained for all mock specimens were added to the databases used for taxonomic assignments of metabarcoding datasets, and were submitted on Genbank under accession numbers MN826120-MN826130 and MN844176-MN844185.

Environmental DNA

Sediment cores were collected from thirteen 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, 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. 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.

1.2 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.

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, Waltham, MA, USA) 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. PCR amplifications (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) 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.

Eukaryotic COI gene amplicon generation

Metazoan COI barcodes were generated using the mlCOIintF (5'-GGWACWGGWTGAACWGTWTAYCCYCC-3') and jgHCO2198 (5'-TAIACYTCIGGRTGICCRAARAAYCA-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, Kusatsu, Japan). Cycling conditions included a 10 min denaturation step followed by 16 cycles of 95 °C for 10 s, 30s at 62°C (−1°C per cycle), 68 °C for 60 s, followed by 15 cycles of 95 °C for 10 s, 30s at 46°C, 68 °C for 60 s and a final extension of 68 °C for 7 min.

Prokaryotic 16S rRNA gene amplicon generation

Prokaryotic barcodes were generated using 515F-Y (5'- GTGYCAGCMGCCGCGGTAA-3') and 926R (5'- CCGYCAATTYMTTTRAGTTT-3') 16S-V4V5 primers (Parada et al. 2016). Triplicate PCR mixtures were prepared as described above for 18S-V1V2, 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.

Amplicon library preparation

PCR triplicates were pooled and PCR products purified using 1X AMPure XP beads (Beckman Coulter, Brea, CA, USA) clean up. Aliquots of purified amplicons were 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 fluorimeter (Invitrogen, Carlsbad, CA, USA). 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, 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 purified using 1X AMPure XP beads.

Sequencing library quality control

Amplicon libraries were quantified by Quant-iT dsDNA HS assay kits using a Fluoroskan Ascent microplate fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and 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, USA).

Sequencing procedures

Library concentrations were normalized to 10 nM by addition of 10 mM Tris-Cl (pH 8.5) and applied to cluster generation according to the Illumina Cbot User Guide (Part # 15006165). 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 (8–9 pM instead of 12–14 pM for standard libraries) and PhiX DNA spike-in was increased (20% instead of 1%) in order to minimize the impacts on the run quality.

Libraries were sequenced on HiSeq2500 (System User Guide Part # 15035786) instruments (Illumina, San Diego, CA, USA) in a 250 bp paired-end mode.

1.3 Bioinformatic analyses

All bioinformatic analyses were performed using a Unix shell script on a home-based cluster (DATARMOR, Ifremer), available on Gitlab (https://gitlab.ifremer.fr/abyss-project/). The mock communities were analysed alongside the natural 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.

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 (non-paired reads). Optionally, sequence file names can also be renamed if necessary using a CSV correspondence file.

Read correction, amplicon cluster generation and taxonomic assignment

Pairs of Illumina reads were corrected with DADA2 v.1.10 (Callahan et al. 2016) following the online tutorial for paired-end HiSeq data (https://benjjneb.github.io/dada2/bigdata_paired.html). Reads were filtered and trimmed with the *filterAndTrim* function and all reads containing ambiguous bases removed. The parameters were set based on tutorial recommendations 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).

The error model was calculated for forward and reverse reads (R1F/R2R pairs and then R2F/R1R pairs) with *learnErrors* based on 100 million randomly chosen bases (default), and reads were dereplicated using *derepFastq*. After read correction with the *dada* function, forward and reverse reads were merged with a minimum overlap of 12 nucleotides, allowing no mismatches (default). The amplicons were then filtered by size. The size range was 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.

Chimeras were removed with *removeBimeraDenovo* and ASVs were taxonomically assigned via the RDP naïve Bayesian classifier method, the default assignment method implemented in DADA2. A second taxonomic assignment method was optionally implemented in the pipeline, allowing assigning ASVs using BLAST+ (Basic Local Alignment Search Tool v2.6.0) based on minimum similarity and minimum coverage (-perc_identity 70 and –qcov_hsp 80). An initial test implementing BLASTn+ to assign taxonomy only to the COI dataset using a 96% percent identity threshold led to the exclusion of the majority of the clusters. Given observed inter-specific 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. The Silva132 reference database was used for the 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/assignation/). Finally, to evaluate the effect of clustering, ASV tables produced by DADA2 were clustered with swarm v2 (Mahe et al. 2015) at *d=1,3,4,5 and 11* for 18S and 16S, and *d=1,5,6,7, and 13* for COI in FROGS (http://frogs.toulouse.inra.fr/) (Escudié et al. 2018). Resulting OTUs were taxonomically assigned via RDP and BLAST+ using the databases stated above.

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 taxonomically based on their RDP or BLAST taxonomy. For both assignment methods, unassigned clusters were removed. Non-target 18S and COI clusters (bacterial, non-metazoan) as well as all clusters with a terrestrial assignment (taxonomic groups known to be terrestrial-only, such as Insecta, Arachnida, Diplopoda, Amphibia, terrestrial mammals, Stylommatophora, Aves, Onychophora, Succineidae, Cyclophoridae, Diplommatinidae, Megalomastomatidae, Pupinidae, Veronicellidae) were removed. Samples were checked to ensure that a minimum of 10,000 metazoan reads were left after refining. Finally, as tag-switching is always 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 (Wangensteen and Turon 2016, script from 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 similarities and co-occurrence rate with more abundant ("parent") clusters. LULU was tested with a minimum relative co-occurrence of 0.90, 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 vast majority of prokaryotes usually show low levels (< 1% divergence) of intra genomic 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 d=1. Although LULU curation may still be useful to merge redundant phylotypes 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 an unrealistic calculation time, even on a powerful cluster (e.g. LULU curation was at 20-40% after 4 days of calculation on our cluster).

1.4 Statistical analyses

Sequence tables were analysed using R with the packages phyloseq v1.22.3 (McMurdie and 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 datasets were normalized by rarefaction to their common minimum sequencing depth, before analysis of mock communities and natural samples.

To evaluate the functionality of the pipeline with the mock communities, taxonomically assigned metazoan clusters were considered as derived from one of the ten species used for the mock communities when the assignment delivered the corresponding species, genus, family, or class. Clusters not fitting the expected taxa were labelled as ‘Others’. Apart from PCR errors, these non-target clusters may also originate from contamination by external DNA from associated microfauna, or gut content in the case of whole polyps used for cnidarians.

Alpha diversity detected using each pipeline in the natural samples was evaluated with the number of observed target-taxa in the rarefied datasets via analyses of variance (ANOVA) on generalized linear models based on quasipoisson distribution models. Homogeneity of multivariate dispersions were verified with the *betapart* package v.1.5.1 (Baselga and Orme 2012). Beta-diversity patterns were visualised via Principal Coordinates Analyses (PCoA), using Jaccard dissimilarities for metazoans and Bray-Curtis dissimilarities for prokaryotes. The effect of site and LULU curation on community composition was tested by means of PERMANOVA, using the function *adonis*2 (vegan), with the same dissimilarities as in PCoAs, and permuting 999 times.Finally, BLAST and RDP taxonomic assignments of the mock samples and the global dataset were compared at the most adequate pipeline settings for each locus. BLAST-refined (minimum identity at 70%) and RDP-refined (minimum phylum bootstrap at 80%) datasets were compared on ASV-level for prokaryotes, and OTU-level for metazoans (swarm *d=3*, LULU at 84% for COI and 90% 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.

2 RESULTS

2.1 Alpha diversity in mock communities

A number of 2 million (18S) and 1.5 million (COI) raw reads were obtained from the two mock communities (Table S4).After refining, these numbers were decreased to 1.3 million for 18S and 0.7 million for COI.

Seven out of ten mock species were recovered in the 18S dataset and all species were detected in the COI dataset (Table 1), even with minimum relative DNA abundance levels as low as 0.7% (Mock 5). Taxonomically unresolved species were correctly assigned up to their common family or class level. Dominant species generally produced more reads in both the clustered and non-clustered datasets (Table S6).

When ASVs were clustered with swarm v2, this generally led to a slight loss of taxonomic resolution: *Chorocaris* sp. was not detected in Mock 5 for 18S at $d > 1$, and the two bivalves *P. kilmeri* and *C. regab* were taxonomically misidentified for COI at $d \geq 1$.

Clustering sequences with swarm v2 reduced the number of clusters produced per species, but some species still produced multiple OTUs even at *d* values as high as *d=11* for 18S (*A. arbuscula*, *Munidopsis* sp., and *E. norvegica*) and *d=13* for COI *D. dianthus*, *A. muricola*, *Chorocaris* sp., and *Paralepetopsis* sp.). Curating with LULU allowed reducing the number of clusters produced per species to nearly one for both loci, but the best results were obtained in datasets clustered at $d > 1$ for 18S and $d \geq 1$ for COI. Moreover, LULU curation tended to decrease the number of non-target clusters (“Others”) (Table 1). In the clustered COI dataset, curating with LULU at 84% *minimum match* resulted in the most accurate detection of community composition, and this for all *d* values tested. However, curating with LULU the 18S data (ASVs or OTUs) led to the loss of one shrimp species (*Chorocaris* sp) when the *minimum match* parameter was at 90% and an additional species was lost (the limpet *Paralepetopsis* sp.) when this parameter was at 84%. LULU consistently merged the shrimp species *Chorocaris* sp with another shrimp species as the latter were always co-occurring in our mock samples.

Table 1. Number of ASVs/OTUs detected per species in the mock communities using different bioinformatic pipelines. White cells indicate an exact match with the number of OTUs expected, grey cells indicate a number of OTUs differing by ±3 from the number expected, and dark grey cells indicate a number of OTUs >3 from the one expected.

18S	**DADA2**	**DADA2 +LULU 90%**	**DADA2+ LULU 84%**		**DADA2+swarm d1/d3/d4/d5/d11**	**DADA2+swarm d1/d3/d4/d5/d11 + LULU 90%**	**DADA2+swarm d1/d3/d4/d5/d11 + LULU 84%**
Mock 3							
Alcyonacea;*A.arbuscula*	64	1	1	Alcyonacea;*A.arbuscula*	29/11/9/7/6	1/1/1/1/1	1/1/1/1/1
Caryophylliidae;*D.dianthus*	2	1	1	Caryophylliidae;*D.dianthus*	2/2/1/1/1	1/1/1/1/1	1/1/1/1/1
Alvinocaris muricola	2	1	1	*Alvinocaris muricola*	2/1/1/1/1	1/1/1/1/1	1/1/1/1/1
Chorocaris sp.	1	0	0	*Chorocaris* sp.	2/1/1/1/1	0/0/0/0/0	0/0/0/0/0
Munidopsis sp.	6	1	1	*Munidopsis* sp.	5/4/3/3/2	1/1/1/1/1	1/1/1/1/1
Gastropoda;*Paralepetopsis* sp.	1	1	0	Gastropoda;*Paralepetopsis* sp.	1/1/1/1/1	1/1/1/1/1	0/0/0/0/0
Vesicomyidae;*P. kilmeri/C. regab/V. gigas*	8	1	1	Bivalvia;*P. kilmeri/C. regab/V. gigas*	5/4/4/4/2	1/2/2/2/1	1/1/1/1/1
Polychaeta;*E.norvegica*	8	3	2	Polychaeta;*E.norvegica*	5/4/4/4/3	3/2/2/2/2	2/1/2/2/2
Others	3	3	2	Others	4/4/4/4/4	2/2/2/2/3	2/2/2/2/2
Mock 5							
Alcyonacea;*A.arbuscula*	54	1	1	Alcyonacea;*A.arbuscula*	28/11/9/7/6	1/1/1/1/1	1/1/1/1/1
Caryophylliidae;*D.dianthus*	1	1	1	Caryophylliidae;*D.dianthus*	1/1/1/1/1	1/1/1/1/1	1/1/1/1/1
Alvinocaris muricola	1	1	1	*Alvinocaris muricola*	1/1/1/1/1	1/1/1/1/1	1/1/1/1/1
Chorocaris sp.	1	0	0	*Chorocaris* sp.	1/0/0/0/0	0/0/0/0/0	0/0/0/0/0
Munidopsis sp.	4	1	1	*Munidopsis* sp.	4/3/3/3/2	1/1/1/1/1	1/1/1/1/1
Gastropoda;*Paralepetopsis* sp.	1	1	0	Gastropoda;*Paralepetopsis* sp.	1/1/1/1/1	1/1/1/1/1	0/0/0/0/0
Vesicomyidae;*P. kilmeri/C. regab/V. gigas*	5	1	1	Bivalvia;*P. kilmeri/C. regab/V. gigas*	5/3/3/3/2	1/1/1/1/1	1/1/1/1/1
Polychaeta;*E.norvegica*	11	3	2	Polychaeta;*E.norvegica*	5/4/4/4/3	3/2/2/2/1	2/1/2/2/2
Others	4	3	2	Others	3/4/4/4/2	4/2/2/2/1	4/2/2/2/3
COI	**DADA2**	**DADA2 +LULU 90%**	**DADA2+ LULU 84%**		**DADA2+swarm d1/d5/d6/d7/d13**	**DADA2+swarm d1/d5/d6/d7/d13 + LULU 90%**	**DADA2+swarm d1/d5/d6/d7/d13 + LULU 84%**
Mock 3							
Acanella arbuscula	1	1	1	*Acanella arbuscula*	1/1/1/1/1	1/1/1/1/1	1/1/1/1/1
Hexacorallia;*D.dianthus*	3	3	3	Hexacorallia;*D.dianthus*	3/4/4/4/3	3/3/3/3/3	3/3/3/3/3
Alvinocaris ;A. muricola	26	2	2	Alvinocaris;*A. muricola*	21/12/10/10/5	1/1/1/1/1	1/1/1/1/1
Chorocaris sp.	2	1	1	*Chorocaris* sp.	3/3/3/3/3	1/1/1/1/1	1/1/1/1/1
Munidopsis sp.	2	1	1	*Munidopsis* sp.	3/2/1/1/1	2/1/1/1/1	1/1/1/1/1
Gastropoda;*Paralepetopsis* sp.	8	2	3	Gastropoda;*Paralepetopsi* s sp.	3/3/3/3/2	2/2/2/2/2	2/2/2/2/2
Phreagena kilmeri	2	1	1	Bivalvia;*P. kilmeri*	2/3/3/3/3	2/2/2/2/2	2/2/2/2/2
Bivalvia;*C. regab*	2	1	1	Bivalvia;*C. regab*			
Vesicomya gigas	1	1	1	*Vesicomya gigas*	1/1/1/1/1	1/1/1/1/1	1/1/1/1/1
Polychaeta;*E.norvegica*	3	2	1	*Eunice norvegica*	2/1/1/1/1	2/1/1/1/1	1/1/1/1/1
Others	7	6	6	Others	3/3/3/3/4	4/5/5/5/5	5/5/5/5/5
Mock 5							
Acanella arbuscula	1	1	1	*Acanella arbuscula*	1/1/1/1/1	1/1/1/1/1	1/1/1/1/1
Hexacorallia;*D.dianthus*	3	3	3	Hexacorallia;*D.dianthus*	3/3/3/3/3	3/3/3/3/3	3/3/3/3/3
Alvinocaris ;A. muricola	26	2	2	Alvinocaris;*A. muricola*	21/12/10/10/5	1/1/1/1/1	1/1/1/1/1
Chorocaris sp.	1	1	1	*Chorocaris* sp.	2/2/2/2/2	1/1/1/1/1	1/1/1/1/1
Munidopsis sp.	2	1	1	*Munidopsis* sp.	2/2/1/1/1	1/1/1/1/1	1/1/1/1/1
Gastropoda;*Paralepetopsis* sp.	5	2	2	Gastropoda;*Paralepetopsis* sp.	3/2/2/2/2	2/2/2/2/2	2/2/2/2/2
Phreagena kilmeri	1	1	1	Bivalvia;*P. kilmeri*	2/2/2/2/2	2/2/2/2/2	2/2/2/2/2
Bivalvia;*C. regab*	2	1	1	Bivalvia;*C. regab*			
Vesicomya gigas	1	1	1	*Vesicomya gigas*	1/1/1/1/1	1/1/1/1/1	1/1/1/1/1
Polychaeta;*E.norvegica*	3	2	1	*Eunice norvegica*	2/2/2/2/2	1/1/1/1/1	1/1/1/1/1
Others	6	5	4	Others	2/2/2/2/2	1/2/2/2/2	1/1/1/1/1

.

2.2 Alpha-diversity patterns in natural samples

High-throughput sequencing results

A number of 44 million (18S), 33 million (COI) and 16 million (16S) reads were obtained from 42 sediment samples, 4 field controls, 2 extraction blanks, and 4-10 PCR blanks (Table S4). Two sediment samples failed amplification for the COI marker gene (PCT_FA_CT2_0_1 and CHR_CT1_0_1). For metazoans, less reads were retained after bioinformatic processing in negative controls (36% for 18S, 47% for COI) compared to true samples (~60% for 18S, ~70% for COI), while the opposite was observed for 16S (74% of reads retained in control samples against 53% in true samples). Negative control samples (field, extraction, and PCR controls) contained 2,186,230 (~8%) 18S reads, 1,015,700 (~4%) COI reads, and 2,618,729 (28%) 16S reads. These reads were mostly originating from the field controls for metazoans (48% for 18S, 55% for COI) and extractions controls for 16S (50%).

After blank correction, data refining, and abundance renormalization, rarefaction curves showed that a plateau was achieved for all samples in both clustered and non-clustered datasets, suggesting an overall sequencing depth adequate to capture the diversity present (Fig. S1). The final 18S datasets (with and without clustering at selected *d* values) contained 8.9-9.6 million marine metazoan reads in 42 sediment samples (Table S4), and comprised 57,661 ASVs and 19,504-44,948 OTUs (Table S6). The final COI datasets contained 4.5-6.9 million marine metazoan reads in 40 sediment samples, and comprised 78,785 ASVs and 44,684-64,669 OTUs. The 16S datasets contained from 6.6 to 6.7 million prokaryotic reads in 42 sediment samples, producing 56,577 ASVs and 41,746-14,631 OTUs.

Number of clusters among pipelines

The number of metazoan clusters detected in the deep-sea sediment samples varied significantly between bioinformatic pipelines chosen (, and also varied significantly among sites (Table 2). However, the pipeline effect was consistent across sites although mean cluster numbers detected per sample spanned a wide range in all loci (100-800 for 18S, 150-1,500 for COI datasets, and 1,500-5,000 for 16S, Fig. 1).

Expectedly, clustering significantly reduced the number of detected 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 (Fig. 1, Table 2). DADA2 detected on average 863 (SE=61) metazoan COI ASVs per sample, and clustering reduced this number to around 500, regardless the *d*-value. For ribosomal loci, clustering at *d*=3-5 reduced OTU numbers of around 25-30% compared to without clustering, while at *d=11*, cluster numbers were halved.

Table 2. Effect of pipeline and site on the number of metazoan and prokaryote clusters. 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*. Significance codes: ***: p<0.001; **: p<0.01; *: p<0.05.

LOCUS	**F-value**	***p-value***	**Significant pairwise comparisons**
COI			
Pipeline	123.13	p<0.001	Dada2 > DS***; DS(d1) > DS(d13)***;
Site	356.37	p<0.001	Dada2 > DL***; DS > DSL 84%***; D(S)L 90% > D(S)L 84%***
Pipeline x Site	0.16	p>0.05	DL > DSL***; DL 90% > DS***
18S V1-V2			
Pipeline	129.16	p<0.001	Dada2 > DS(d>1)***; DS(d1) > DS(d>1)***; DS(d11) < DS(d1-5)***;
Site	154.52	p<0.001	Dada2 > DL***; DS > DSL 84%***; D(S)L 90% > D(S)L 84%***;
Pipeline x Site	0.49	p>0.05	DL 84% < DS***
16S V4-V5			
Pipeline	179.19	p<0.001	Dada2 > DS***;
Site	18.46	p<0.001	DS(d1) > DS(d>1)***; DS(d11) < DS(d1-5)***
Pipeline x Site	0.06	p>0.05	

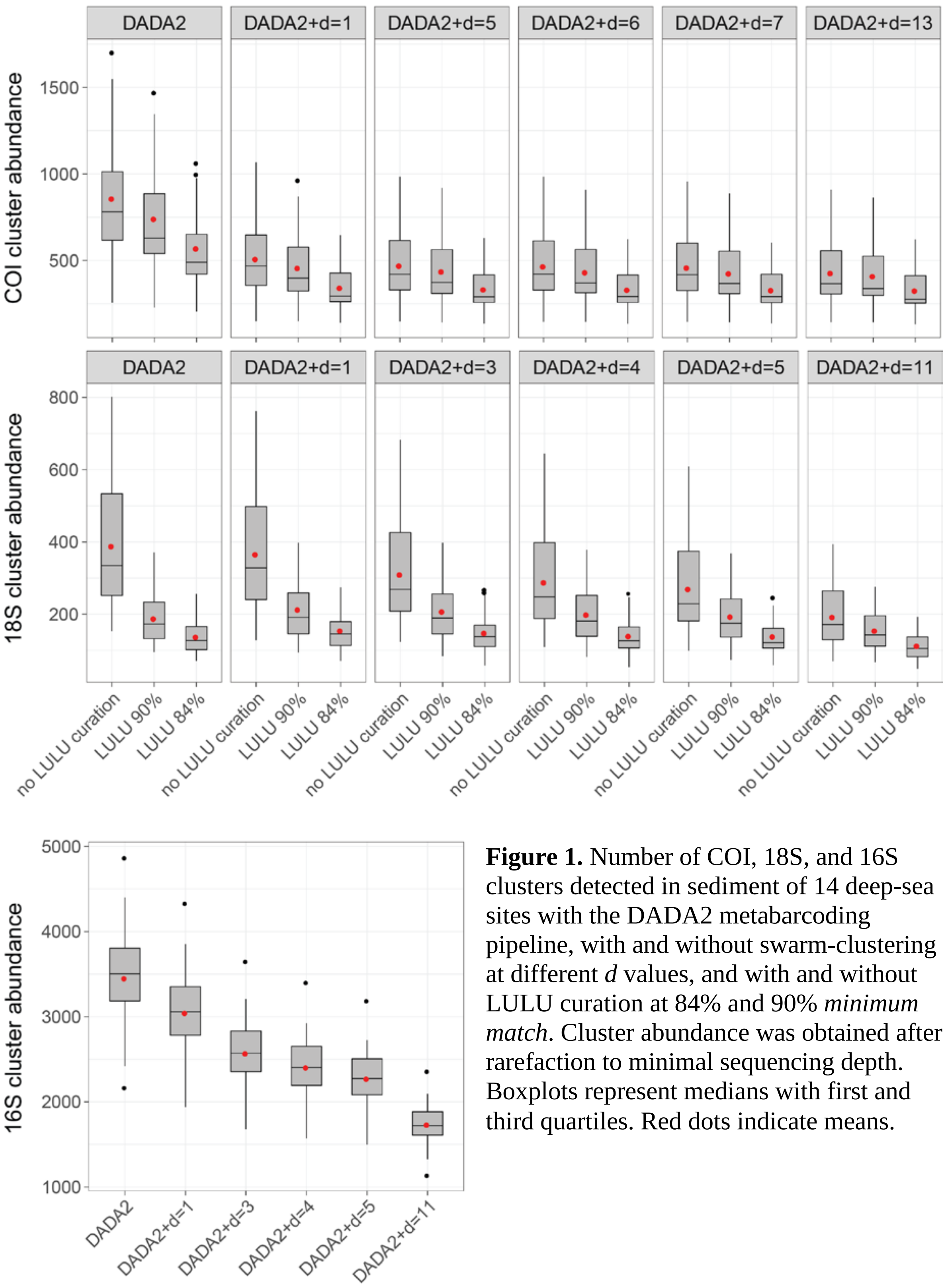

Figure 1. Number of COI, 18S, and 16S clusters detected in sediment of 14 deep-sea sites with the DADA2 metabarcoding pipeline, with and without swarm-clustering at different *d* values, and with and without LULU curation at 84% and 90% *minimum match*. Cluster abundance was obtained after rarefaction to minimal sequencing depth. Boxplots represent medians with first and third quartiles. Red dots indicate means.

LULU curation of metazoan ASVs significantly decreased the number of clusters detected at both tested *minimum match* values (Table 2). For OTU datasets, the decrease was significant only when the *minimum match* parameter was at 84%. The effect of LULU curation was stronger at a lower *minimum match* value for both loci, as LULU curation at 90% of ASVs or OTUs resulted in significantly more clusters than when the minimum match was at 84% (Table 2). The effect of LULU curation of was also more pronounced for the 18S locus: LULU decreased by 31-65% the number of 18S ASVs/OTUs, compared to 7-33% for COI. LULU curation of ASVs or OTUs resulted in comparable cluster numbers in the 18S datasets, regardless the *d*-value used for clustering. For example, at 84% *minimum match*, LULU curation produced on average 137 ± 7 and 140 ± 8 clusters per sample after application on ASVs and OTUs (*d=4*) respectively. At 90%, these numbers were at 189 ± 11 and 200 ± 12 (Fig. 1). This was not the case for COI, where LULU curation of ASVs resulted in significantly more clusters (574 ± 38 at 84% and 742 ± 53 at 90%) than LULU curation of OTUs (334 ± 21 and 433 ± 31 for *d=6*).

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 (Fig S2).

2.3 Patterns of beta-diversity between pipelines

Community differences were visualized using PCoA ordinations (Jaccard and Bray-Curtis dissimilarities for metazoans and prokaryotes respectively) in clustered and non-clustered datasets (Fig. 2, Fig. S3). Expectedly, PERMANOVAs confirmed that sites differed significantly in terms of community structure, accounting from 45% to 89% of variation in data. Evaluating the effect of

LULU curation (at 84% and 90%) for metazoans showed that LULU-curated data resolved similar ecological patterns than non-curated data, accounting from 0.5% (COI) to 1.3% (18S) of variation in data (Fig. 2).

Although ASV and OTU datasets detected similar levels of variation due to sites in PERMANOVAs, clustering levels affected the ecological patterns resolved by ordinations in rRNA loci (Fig 2). At low *d* values (*d=1-3*), ecological patterns were consistent to patterns observed in the ASV datasets, with samples segregating by site and depth. Increasing *d* values produced stronger segregation among sites, thus resulting in differentiation among ocean basins rather than depth. This change in resolution occurred with *d* values as low as *d=4* for 18S but was strongest at *d=11* for both rRNA loci (Fig. S3, Fig. 2).

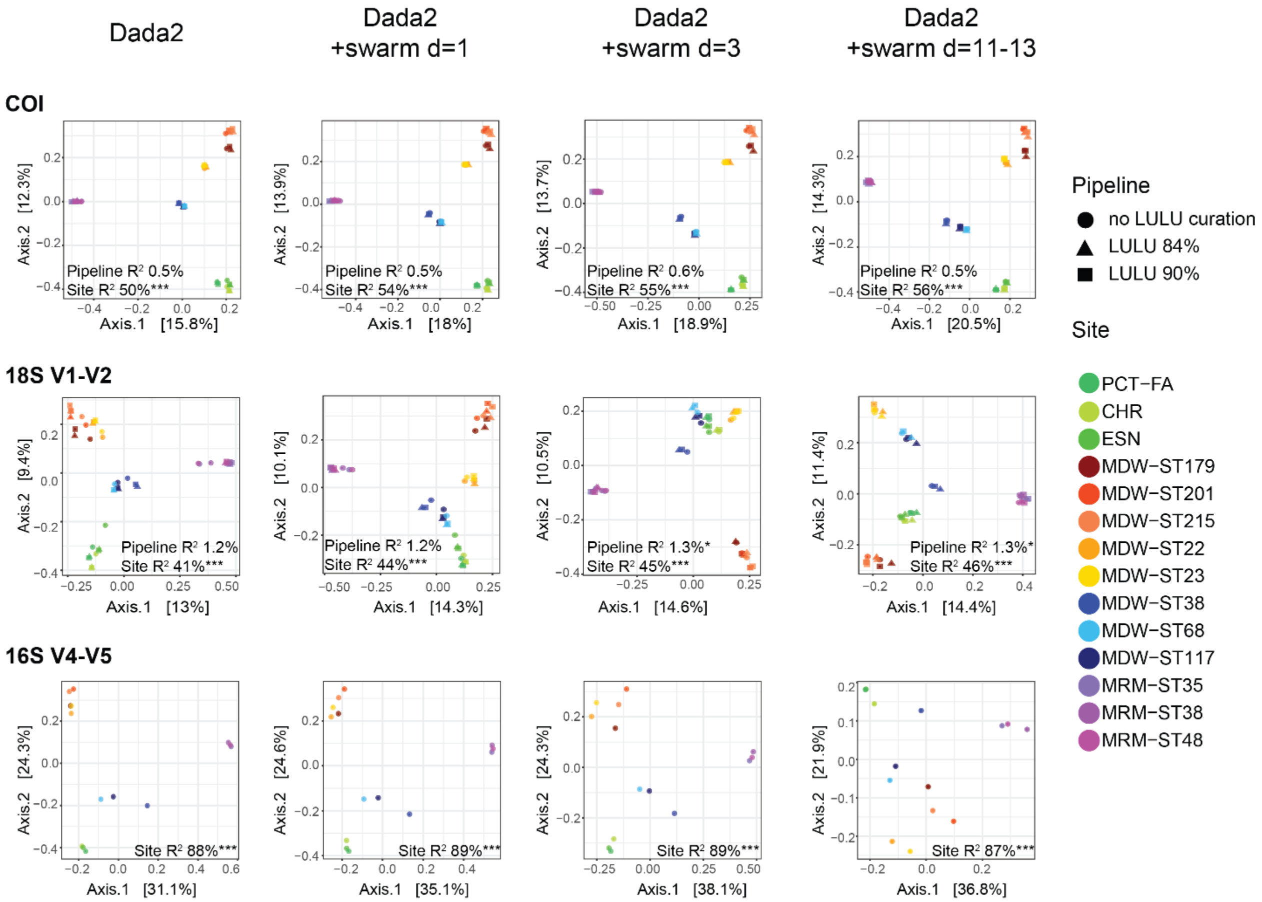

Figure 2. Beta-diversity patterns in ASV and OTU-centred datasets. PCoA ordinations showing community differentiation observed between sites and LULU *vs* not LULU curated samples, for the DADA2 metabarcoding pipeline with and without clustering. Metazoan datasets were clustered at *d=1-13* (COI) *d=1-11* (18S) and curated with LULU at two minimum match values. The prokaryote 16S dataset was clustered at *d=1-11*. R^2 values and associated p-values obtained in PERMANOVAs are shown in the ordination plots. Significance codes: ***: p<0.001; **: p<0.01; *: p<0.05. Colour codes: Green: Mediterranean < 1,000 m; Red-yellow: Mediterranean-Atlantic transition zone 300-1,000 m; Blue: North Atlantic < 1,000 m; Purple: Arctic < 1,000 m.

2.4 Taxonomic assignment quality

BLAST and RDP Bayesian Classifier assignments were compared in the mock communities and natural samples, on data clustered at *d=3* and curated with LULU at 84% for COI and 90% for 18S. For prokaryotes, assignment methods were compared on the ASV-level. BLAST and RDP assigned similar amounts of OTUs in the prokaryote dataset, but BLAST assigned 20-70% less OTUs in the metazoan datasets (Table S7). Assigning with BLAST at a minimum of 70% hit identity resulted in comparable results as described above. Eight of the ten species were recovered with COI and six species were recovered with 18S, while the vesicomyid bivalves were taxonomically unresolved with both loci (Fig. S4). Although most species produced one single OTU, between one and three species still resulted in 2-3 OTUs in each mock sample. Assigning the 18S dataset with RDP resulted in comparable taxonomic resolutions, although more species produced more than one OTU. Assigning the COI dataset 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, all 10 species were detected, and this at expected OTU abundances, once data was filtered for phylum bootstrap levels ≥ 80% (Fig S4). However, applying a phylum bootstrap minimum of 80% resulted in a strong decrease in the number of final target OTUs, particularly for COI where only 226 OTUs remained after filtering (Table S7). This reduced recovery with RDP after applying a minimum phylum bootstrap level was not observed in prokaryotes, where 51,000-55,000 ASVs were left after filtering with both assignment methods (Table S7).

BLAST hit identities of the overall datasets varied strongly depending on phyla and marker gene (Fig. 3). For 18S, most clusters had hit identities ≥ 90%. Poorly assigned clusters (hit identity < 90%) represented less than 20% of the dataset and were mostly assigned to

Nematoda, Cnidaria, Tardigrada, Porifera, and Xenacoelomorpha. For COI, nearly all clusters had similarities to sequences in databases lower than 90%. Overall, arthropods and echinoderms were detected at similar levels by both markers. The 18S barcode marker performed better in the detection of nematodes, annelids, platyhelminths, and xenacoelomorphs while COI mostly detected cnidarians, molluscs, and poriferans (Fig. 3), highlighting the complementarity of these two loci. BLAST hit identity was much higher for prokaryotes, with most clusters assigned with more than 90% similarity to sequences in databases. When datasets were filtered for RDP phylum bootstrap levels ≥ 80%, most assignments also had high genus bootstrap values for ribosomal loci. However, for COI, a considerable number of OTUs assigned to arthropods, cnidarians, molluscs, vertebrates, and poriferans still had genus bootstraps < 60%.

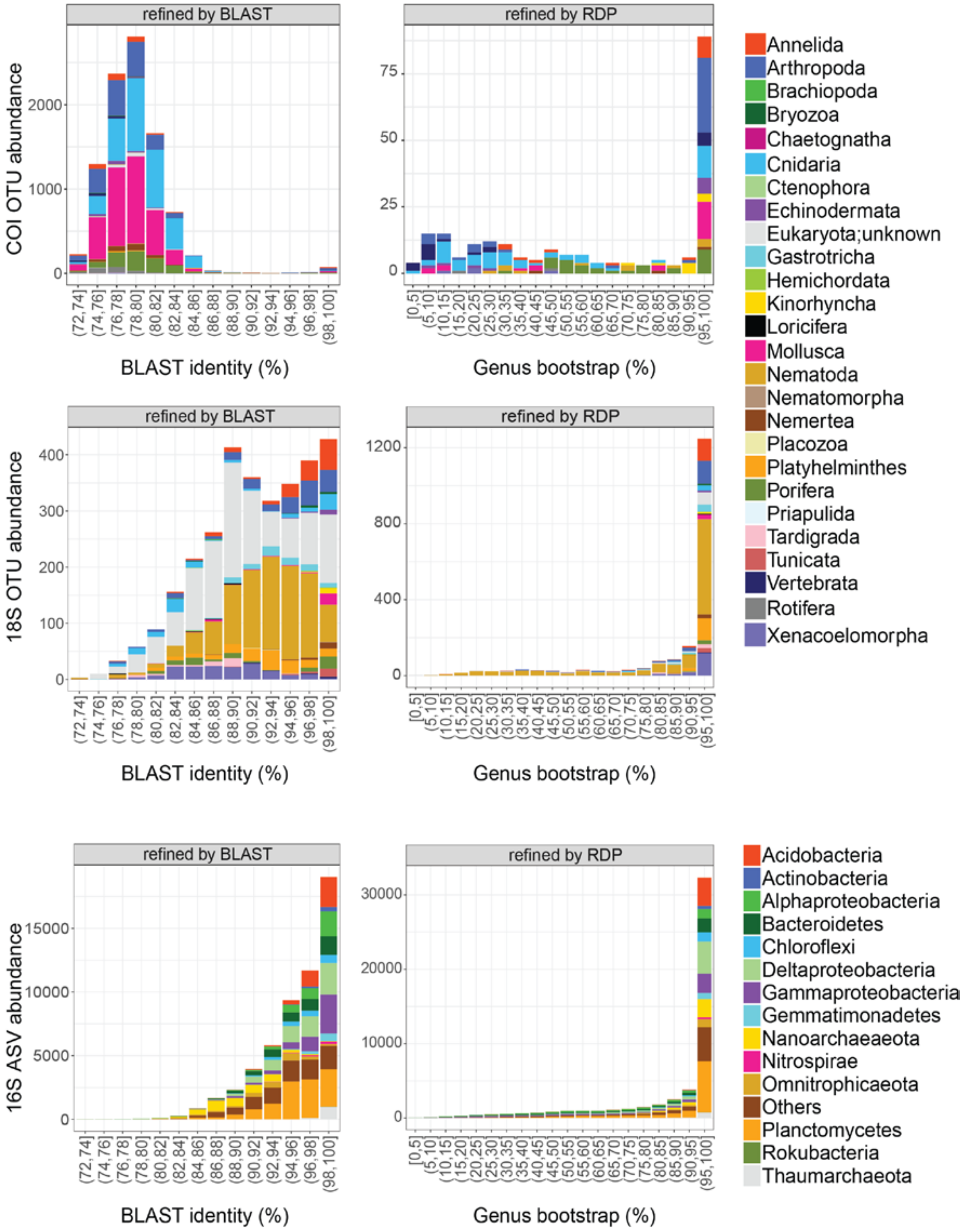

Figure 3. Taxonomic assignment quality of BLAST and RDP methods on metazoan and prokaryote metabarcoding datasets of 14 deep-sea sites. BLAST hit identity of all target clusters detected is given at hit identities > 70%. RDP-assigned data was filtered for phylum bootstraps ≥ 80%, and associated genus bootstraps are displayed. Taxonomic assignments were performed on the Silva132 database for 18S and 16S, and on the MIDORI-UNIQUE database, subsampled to marine taxa for COI.

3 DISCUSSION

3.1 ASVs and OTUs for genetic *vs* species diversity

The rise of HTS and the subsequent use of DNA metabarcoding have revolutionized microbiology by unlocking the access to uncultivable microorganisms, which represent by far the great majority of prokaryotes (Klappenbach et al. 2001). The development and improvement of molecular and bioinformatic methods to perform inventories were historically primarily developed for 16S rRNA barcode loci, before being transferred to the eukaryotic kingdom based on the use of barcode markers such as 28S and 18S rRNA, ITS, or mitochondrial markers such as COI (Valentini et al. 2009; Bellemain et al. 2010). Thus, most bioinformatics pipelines were initially developed accounting for intrinsic properties of prokaryotes and concepts inherent to microbiology (Caporaso et al. 2010; Schloss et al. 2009; Boyer et al. 2016), before being transferred to eukaryotes in general, or metazoans in particular. Such application transfers require adaptations to account for differences in both concepts and basic biological features. One example is the question of the relevance of using ASVs, 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 working on metazoans.

First, metazoans are well known to exhibit variable and sometimes very high intraspecific polymorphism in 18S-V1 and above all in COI. Second, the results on the mock samples showed that single individuals produced very different numbers of ASVs, indicating that ASV-centred datasets do not reflect actual species composition in metazoans. As this "demultiplication" will be highly variable across taxa (as seen in Fig. S2, and references such as Plouviez et al. 2009 and Teixeira et al. 2013), the taxonomic compositions of samples based on ASVs will reflect genetic rather than species diversity.

Clustering ASVs into OTUs and/or curating with LULU alleviated the numerical inflation, but some species still produced more than one OTU, even at *d*-values as high as *d=11-13*. While clustering and LULU curation improved numerical results in the mock communities, they were associated with a decrease in taxonomic resolution, especially for 18S where some closely related species were merged with increasing clustering/filtering thresholds (i.e. the vesicomyid bivalves, the gastropod, and the shrimp species; Table 1). When studying natural habitats, very likely to harbour closely related co-occurring species, both LULU curation and clustering are thus likely to lead to the loss of true species diversity, particularly for low-resolution markers such as 18S. Optimal results in the mock samples, i.e. delivering the best balance between the limitation of spurious clusters and the loss of true species diversity, were obtained with LULU curation at 90% for 18S and 84% for COI, highlighting the importance of adjusting bioinformatic correction tools to each barcode marker, a step for which mock communities are most adequate.

3.2 ASVs *vs* OTUs in natural communities: adapting pipeline parameters to marker properties

Life histories of organisms, together with intrinsic properties of marker genes, determine the level of intragenomic and intraspecific diversity. 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 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). Together with clustering, LULU curation at 84% proved effective in limiting the number of multiple clusters produced by single individuals, confirming its efficiency to correct for intragenomic diversity (Table 1).

The mock communities we used here did not contain several haplotypes of the same species (intraspecific variation), as is most often the case in environmental samples. This prevents us from generalizing the comparable results obtained after LULU curation of ASVs and OTUs, and the apparently limited effect of clustering in the mock samples to communities that are more complex. However, LULU curation of ASVs is not suited to account for natural haplotype diversity: not all haplotypes co-occur and when they do so, they may vary in proportion and dominance relationships, making clustering more suited to account for natural haplotypic diversity. Thus, clustering ASVs will still be necessary to produce inventories of metazoan communities that reflect species rather than gene diversity.

As expected, evaluation of clustering and LULU curation on natural samples showed distinct results for 18S and COI. Indeed, 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 expected for the slower evolving 18S gene (Carranza et al. 1996), than for COI which exhibits faster evolutionary rates (Machida and Knowlton 2012; Machida et al. 2012). This is reflected in the lower ASV (DADA2) to OTU (DADA2+swarm) ratios observed here for 18S (1.0-2.2.) compared to COI (2.0-2.7) data at clustering *d*-values comprised between one and seven (Table S6), underlining the different influence –and importance– of clustering on these loci, and the need for a versatile, marker by marker choice for clustering and curation parameters. When applying LULU to ASVs (DADA2) *versus* OTUs (DADA2+swarm) on 18S, similar cluster numbers were obtained (Fig. 1), suggesting a limited added effect of clustering for this marker once DADA2 and LULU are applied. This is in line with its slow evolutionary rate (Carranza et al. 1996) leading to a limited number of haplotypes per species compared to COI. In contrast, for COI, LULU curation of the ASV dataset led to nearly twice the number of clusters (574 ± 38 at 84% and 742 ± 53 at

90%) compared to LULU curation of OTUs (at *d=6*: 334 ± 21 for 84% and 433 ± 31 for 90%). This confirms the higher intraspecific diversity of COI, and the need to combine clustering with LULU curation to account for intraspecific diversity in natural samples, especially with highly polymorphic markers such as COI.

Finally, the reproductive mode and pace of selection in microbial populations may lead to locally lower levels of intraspecific variation than the one expected for metazoans. Prokaryotic alpha diversity was however also affected by the clustering of ASVs (Fig. 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, Fig. 1, decrease of ~25%) suggests the occurrence of very closely related 16S rRNA sequences, possibly belonging to the same ecotype/species. Such entities may still be important to delineate 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.

3.3 Influence on beta diversity

After focusing on alpha diversity estimates, i.e. on the numerical accuracy of inventories, the analysis of community structures showed that the LULU-curated datasets resolved similar ecological patterns as datasets not curated with LULU. However, clustering affected resolution of ecological patterns in ribosomal loci when *d* values were high, and this was not the case for COI, where similar patterns were resolved in all datasets (Fig. 2). This is in accordance with other studies reporting severe impacts of bioinformatic parameters on alpha diversity while comparable patterns of beta diversity are observed in ASV and OTU datasets, at least down to a minimum level of clustering stringency (Xiong and Zhan 2018; Bokulich et al. 2013).

Clustering and LULU curation mainly led to the decrease of the number of clusters assigned to particular taxa in both loci, such as annelids, arthropods, nematodes, or platyhelminthes for 18S, and chordates, cnidarians, echinoderms, or poriferans for COI (Fig. S2). The strong decrease in cluster numbers observed in these phyla suggests that the latter have greater intraspecific polymorphism, although the decrease could also be due to the merging of closely related species, as both markers have lower taxonomic resolution in particular taxa. This has been acknowledged for 18S in general, but in nematodes in particular (Derycke et al. 2010), and reported in cnidarians with COI (Hebert et al. 2003).

Overall, based on alpha and beta diversity results observed in mock communities and natural samples, applying LULU at 84% seems to efficiently curate metazoan COI datasets without significant loss of species, but clustering is required, at least at *d=1*, in order to address high intraspecific polymorphism. For 18S, LULU curation seems to require values above 84% (e.g. 90%) in order to avoid the loss of species, as seen in the mock communities. However, the low taxonomic resolution obtained with this marker suggests that clustering should be performed at low *d-values* (*d<4*) to address intraspecific polymorphism without affecting beta-diversity patterns. For prokaryotes, clustering 16S ASVs at *d=1* reduces the number of detected clusters by ~25%, which may help addressing intragenomic variation when needed.

3.4 Taxonomic resolution and assignment quality

The COI locus allowed the detection of all ten species present in the mock samples, compared to seven in the 18S dataset (Table 1). This locus 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 (Tang et al. 2012; Clarke et al. 2017; Andújar et al. 2018). Our results also support approaches combining nuclear and mitochondrial markers to achieve more comprehensive biodiversity inventories (Cowart 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 high complementarity of COI and 18S in terms of targeted taxa (highlighted in Fig. S2), also supports the approach taken by Stefanni et al. (2018), as subsampling each gene dataset for its “best targeted phyla” and subsequently combining both seems to be a very efficient way to produce comprehensive and non-redundant biodiversity inventories.

Finally, compared to BLAST assignments, similar taxonomic resolution was observed using the RDP Bayesian Classifier on the mock samples for 18S (Fig. S4) and for COI when using the MIDORI-UNIQUE marine-only database. Poor performance of RDP using the full MIDORI database is likely due to the size of the database, and to its low coverage of deep-sea species. 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 (Fig. 3). Using minimum similarities of 80% for COI and 86% for 18S as cut-off values for metazoans has been shown to improve the taxonomic quality of metazoan metabarcoding datasets (Stefanni et al. 2018). However, phylogenies of marine invertebrates have found high levels of species divergence (up to ~30%), even within genera (Zanol et al. 2010). Consequently, 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 (Shank et al. 1999; Herrera et al. 2015). At present, it thus seems difficult to work at taxonomic levels beyond phylum-level with deep-sea metabarcoding data when using large public databases. Small databases, taxonomically similar to the targeted communities, and with sequences of the same length as the amplified fragment of interest, are known to maximise accurate identification (Macheriotou et al. 2019). When using the reduced marine-only COI database, RDP provided the most accurate assignments in the mock samples when the phylum bootstrap level was ≥ 80 (Fig. S 4), although this filtering threshold drastically reduced the number of OTUs in the overall dataset (Table S7). The development of custom-built marine RDP training sets, without overrepresentation of terrestrial species, is therefore needed for this Bayesian assignment method to be effective on deep-sea datasets. With reduced and more specific databases, removing clusters with phylum bootstrap-level < 80 should be an efficient way to increase taxonomic quality of deep-sea metabarcoding datasets. At present, if 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, extracting the clusters belonging to the groups of interest, then re-assigning taxonomy to these target taxa using RDP and a smaller, taxon-specific database.

CONCLUSIONS AND PERSPECTIVES

Using mock communities and natural 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 datasets can be easily merged *a posteriori*, and clustered if necessary afterwards. Nevertheless, clustering ASVs into OTUs will be required to obtain accurate species-level inventories, at least for metazoan communities, with a more severe influence of clustering observed on alpha diversity estimates than beta-diversity patterns. Considering 16S polymorphism observed in prokaryotic species (Acinas et al. 2004) and the possible geographic segregation of their populations, clustering may also be required in prokaryotic datasets, for example in studies screening for species associations (i.e. symbiotic relationships) as symbionts may be prone to differential fixation through enhanced drift (Shapiro, Leducq, & Mallet, 2016).

Our results also demonstrated that LULU effectively curates metazoan biodiversity inventories obtained through metabarcoding. They also underline the need to adapt parameters for curation (e.g. LULU curation at 90% for 18S and 84% for COI) and clustering to each gene used and taxonomic compartment targeted, in order to identify an optimal balance between the correction for spurious clusters and the merging of closely related species.

Finally, our findings 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.

The pipeline is publicly available on Gitlab (https://gitlab.ifremer.fr/abyss-project/), and 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.

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 and Patrick Durand for bioinformatic support, Cathy Liautard-Haag for her help with lab management, and Babett Guenther for help in barcoding. We also wish to express our gratitude to the participants and mission chiefs of the Essnaut (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 an anonymous referee for their comments on a previous version of this manuscript.

REFERENCES

Acinas, Silvia G, Luisa A Marcelino, Vanja Klepac-Ceraj, and Martin F Polz. 2004. ‘Divergence and Redundancy of 16S RRNA Sequences in Genomes with Multiple Rrn Operons’. *Journal of Bacteriology* 186 (9): 2629–35. https://doi.org/10.1128/JB.186.9.2629-2635.2004.

Alberdi, Antton, Ostaizka Aizpurua, M. Thomas P. Gilbert, and Kristine Bohmann. 2017. ‘Scrutinizing Key Steps for Reliable Metabarcoding of Environmental Samples’. Edited by Andrew Mahon. *Methods in Ecology and Evolution*, 2017. https://doi.org/10.1111/2041-210X.12849.

Andújar, Carmelo, Paula Arribas, Douglas W. Yu, Alfried P. Vogler, and Brent C. Emerson. 2018. ‘Why the COI Barcode Should Be the Community DNA Metabarcode for the Metazoa’. *Molecular Ecology* 27 (20): 3968–75. https://doi.org/10.1111/mec.14844.

Baselga, Andrés, and C. David L. Orme. 2012. ‘Betapart : An R Package for the Study of Beta Diversity’. *Methods in Ecology and Evolution* 3 (5): 808–12. https://doi.org/10.1111/j.2041-210X.2012.00224.x.

Bazin, Eric, Sylvain Glémin, and Nicolas Galtier. 2006. ‘Population Size Does Not Influence

Mitochondrial Genetic Diversity in Animals'. *Science* 312 (5773): 570–72. https://doi.org/10.1126/science.1122033.

Bellemain, Eva, Tor Carlsen, Christian Brochmann, Eric Coissac, Pierre Taberlet, and Håvard Kauserud. 2010. 'ITS as an Environmental DNA Barcode for Fungi: An in Silico Approach Reveals Potential PCR Biases'. *BMC Microbiology* 10 (July): 189. https://doi.org/10.1186/1471-2180-10-189.

Bensasson, Douda, De Xing Zhang, Daniel L. Hartl, and Godfrey M. Hewitt. 2001. 'Mitochondrial Pseudogenes: Evolution's Misplaced Witnesses'. *Trends in Ecology and Evolution*. https://doi.org/10.1016/S0169-5347(01)02151-6.

Bhadury, P, M C Austen, D T Bilton, P J D Lambshead, A D Rogers, and G R Smerdon. 2006. 'Molecular Detection of Marine Nematodes from Environmental Samples: Overcoming Eukaryotic Interference'. *Aquatic Microbial Ecology* 44 (1): 97–103. https://doi.org/Doi 10.3354/Ame044097.

Bik, Holly M., Way Sung, Paul De Ley, James G Baldwin, Jyotsna Sharma, Axayácatl Rocha-Olivares, and W Kelley Thomas. 2012. 'Metagenetic Community Analysis of Microbial Eukaryotes Illuminates Biogeographic Patterns in Deep-Sea and Shallow Water Sediments.' *Molecular Ecology* 21 (5): 1048–59. https://doi.org/10.1111/j.1365-294X.2011.05297.x.

Bista, Iliana, G Carvalho, K Walsh, M Christmas, Mehrdad Hajibabaei, P Kille, D Lallias, and Simon Creer. 2015. 'Monitoring Lake Ecosystem Health Using Metabarcoding of Environmental DNA: Temporal Persistence and Ecological Relevance'. *Genome* 58 (5): 197.

Bokulich, Nicholas A, Sathish Subramanian, Jeremiah J Faith, Dirk Gevers, Jeffrey I Gordon, Rob Knight, David A Mills, and J Gregory Caporaso. 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, Germain, Judith Bakker, Owen S. Wangensteen, Stefano Mariani, Lucas Bonnin, Jean Baptiste Juhel, Jeremy J. Kiszka, et al. 2018. ‘Environmental DNA Illuminates the Dark Diversity of Sharks’. *Science Advances* 4 (5): eaap9661. https://doi.org/10.1126/sciadv.aap9661.

Boyer, F, C Mercier, A Bonin, Y Le Bras, Pierre Taberlet, and Eric Coissac. 2016. ‘OBITOOLS: A UNIX-Inspired Software Package for DNA Metabarcoding’. *Molecular Ecology Resources* 16 (1): 176–82. https://doi.org/10.1111/1755-0998.12428.

Brannock, P M, and K M Halanych. 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, F J J Chain, T J Crease, H J MacIsaac, and M E Cristescu. 2015. ‘Divergence Thresholds and Divergent Biodiversity Estimates: Can Metabarcoding Reliably Describe Zooplankton Communities?’ *Ecology and Evolution* 5 (11): 2234–51. https://doi.org/10.1002/ece3.1485.

Bucklin, Ann, Dirk Steinke, and Leocadio Blanco-Bercial. 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, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. ‘DADA2: High-Resolution Sample Inference from Illumina Amplicon Data’. *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Callahan, Benjamin J., Paul J McMurdie, and Susan P Holmes. 2017. ‘Exact Sequence Variants Should Replace Operational Taxonomic Units in Marker-Gene Data Analysis’. *ISME*

Journal 11 (12): 2639–43. https://doi.org/10.1038/ismej.2017.119.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. ‘QIIME Allows Analysis of High-Throughput Community Sequencing Data’. *Nature Methods* 7 (5): 335–36. https://doi.org/10.1038/nmeth.f.303.

Carranza, Salvador, Gonzalo Giribet, Carles Ribera, Jaume Baguñà, and Marta Riutort. 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): 824–32. https://doi.org/10.1093/oxfordjournals.molbev.a025643.

Carugati, Laura, Cinzia Corinaldesi, Antonio Dell’Anno, and Roberto Danovaro. 2015. ‘Metagenetic Tools for the Census of Marine Meiofaunal Biodiversity: An Overview’. *Marine Genomics* 24 (December): 11–20. https://doi.org/10.1016/j.margen.2015.04.010.

Clare, Elizabeth L., Frédéric J.J. Chain, Joanne E. Littlefair, and Melania E. Cristescu. 2016. ‘The Effects of Parameter Choice on Defining Molecular Operational Taxonomic Units and Resulting Ecological Analyses of Metabarcoding Data’. Edited by Kristy Deiner. *Genome* 59 (11): 981–90. https://doi.org/10.1139/gen-2015-0184.

Clarke, Laurence J., Jason M. Beard, Kerrie M. Swadling, and Bruce E. Deagle. 2017. ‘Effect of Marker Choice and Thermal Cycling Protocol on Zooplankton DNA Metabarcoding Studies’. *Ecology and Evolution* 7 (3): 873–83. https://doi.org/10.1002/ece3.2667.

Cohan, Frederick M. 2001. ‘Bacterial Species and Speciation’. Edited by M. Kane. *Systematic Biology* 50 (4): 513–24. https://doi.org/10.1080/10635150118398.

Coissac, Eric, Tiayyba Riaz, and Nicolas Puillandre. 2012. ‘Bioinformatic Challenges for DNA Metabarcoding of Plants and Animals’. *Molecular Ecology* 21 (8): 1834–47. https://doi.org/10.1111/j.1365-294X.2012.05550.x.

Cowart, Dominique A, Miguel Pinheiro, Olivier Mouchel, Marion Maguer, Jacques Grall, Jacques Miné, and Sophie Arnaud-Haond. 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, Simon, Kristy Deiner, Serita Frey, Dorota Porazinska, Pierre Taberlet, W. Kelley Thomas, Caitlin Potter, and Holly M. Bik. 2016. 'The Ecologist's Field Guide to Sequence-Based Identification of Biodiversity'. Edited by Robert Freckleton. *Methods in Ecology and Evolution* 7 (9): 1008–18. https://doi.org/10.1111/2041-210X.12574.

Davis, Nicole M., Diana M. Proctor, Susan P. Holmes, David A. Relman, and Benjamin J. Callahan. 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.

Deiner, Kristy, Emanuel A. Fronhofer, Elvira Mächler, Jean Claude Walser, and Florian Altermatt. 2016. 'Environmental DNA Reveals That Rivers Are Conveyer Belts of Biodiversity Information'. *Nature Communications* 7 (1): 12544. https://doi.org/10.1038/ncomms12544.

Deiner, Kristy, Jean-Claude C Walser, E Machler, Florian Altermatt, Elvira Mächler, and Florian Altermatt. 2015. 'Choice of Capture and Extraction Methods Affect Detection of Freshwater Biodiversity from Environmental DNA'. *Biological Conservation* 183 (March): 53–63. https://doi.org/10.1016/j.biocon.2014.11.018.

Dejean, Tony, Alice Valentini, Antoine Duparc, Stephanie Pellier-Cuit, Francois Pompanon, Pierre Taberlet, and Claude Miaud. 2011. 'Persistence of Environmental DNA in Freshwater Ecosystems'. *PLoS One* 6 (8). https://doi.org/10.1371/journal.pone.0023398.

Derycke, Sofie, Jan Vanaverbeke, Annelien Rigaux, Thierry Backeljau, and Tom Moens. 2010.

‘Exploring the Use of Cytochrome Oxidase c Subunit 1 (COI) for DNA Barcoding of Free-Living Marine Nematodes’. Edited by Peter Roopnarine. *PLoS ONE* 5 (10): e13716. https://doi.org/10.1371/journal.pone.0013716.

Dickie, Ian A., Stephane Boyer, Hannah L. Buckley, Richard P. Duncan, Paul P. Gardner, Ian D. Hogg, Robert J. Holdaway, et al. 2018. ‘Towards Robust and Repeatable Sampling Methods in EDNA-Based Studies’. *Molecular Ecology Resources*. Wiley/Blackwell (10.1111). https://doi.org/10.1111/1755-0998.12907.

Drummond, A J, R D Newcomb, T R Buckley, D Xie, A Dopheide, B C M Potter, J Heled, et al. 2015. ‘Evaluating a Multigene Environmental DNA Approach for Biodiversity Assessment’. *Gigascience* 4. https://doi.org/ARTN 4610.1186/s13742-015-0086-1.

Eren, A Murat, Joseph H Vineis, Hilary G Morrison, and Mitchell L Sogin. 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.

Escudié, Frédéric, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, et al. 2018. ‘FROGS: Find, Rapidly, OTUs with Galaxy Solution’. Edited by Bonnie Berger. *Bioinformatics* 34 (8): 1287–94. https://doi.org/10.1093/bioinformatics/btx791.

Evans, N T, B P Olds, M A Renshaw, C R Turner, Y Y Li, C L Jerde, A R Mahon, M E Pfrender, G A Lamberti, and D M Lodge. 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, Gentile Francesco, Johan Pansu, Aurélie Bonin, Eric Coissac, Charline Giguet-Covex, Marta De Barba, Ludovic Gielly, et al. 2015. ‘Replication Levels, False Presences and the Estimation of the Presence/Absence from EDNA Metabarcoding Data’. *Molecular Ecology*

Resources 15 (3): 543–56. https://doi.org/10.1111/1755-0998.12338.

Fonseca, Vera G. 2018. ‘Pitfalls in Relative Abundance Estimation Using Edna Metabarcoding’. *Molecular Ecology Resources* 18 (5): 923–26. https://doi.org/10.1111/1755-0998.12902.

Fonseca, Vera G., Gary R Carvalho, Way Sung, Harriet F Johnson, Deborah M Power, Simon P Neill, Margaret Packer, et al. 2010. ‘Second-Generation Environmental Sequencing Unmasks Marine Metazoan Biodiversity’. *Nature Communications* 1. https://doi.org/9810.1038/ncomms1095.

Frøslev, Tobias Guldberg, Rasmus Kjøller, Hans Henrik Bruun, Rasmus Ejrnæs, Ane Kirstine Brunbjerg, Carlotta Pietroni, and Anders Johannes Hansen. 2017. ‘Algorithm for Post-Clustering Curation of DNA Amplicon Data Yields Reliable Biodiversity Estimates’. *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-01312-x.

Gevers, Dirk, Frederick M. Cohan, Jeffrey G. Lawrence, Brian G. Spratt, Tom Coenye, Edward J. Feil, Erko Stackebrandt, et al. 2005. ‘Re-Evaluating Prokaryotic Species’. *Nature Reviews Microbiology* 3 (9): 733–39. https://doi.org/10.1038/nrmicro1236.

Goldberg, Caren S., Cameron R. Turner, Kristy Deiner, Katy E. Klymus, Philip Francis Thomsen, Melanie A. Murphy, Stephen F. Spear, et al. 2016. ‘Critical Considerations for the Application of Environmental DNA Methods to Detect Aquatic Species’. Edited by M. Gilbert. *Methods in Ecology and Evolution* 7 (11): 1299–1307. https://doi.org/10.1111/2041-210X.12595.

Hashimoto, Joel G, Bradley S Stevenson, and Thomas M Schmidt. 2003. ‘Rates and Consequences of Recombination between RRNA Operons’. *Journal of Bacteriology* 185 (3): 966–72. https://doi.org/10.1128/JB.185.3.966-972.2003.

Hebert, Paul D.N., Sujeevan Ratnasingham, and Jeremy R. de Waard. 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): S96-9. https://doi.org/10.1098/rsbl.2003.0025.

Herrera, Santiago, Hiromi Watanabe, and Timothy M. Shank. 2015. ‘Evolutionary and Biogeographical Patterns of Barnacles from Deep-Sea Hydrothermal Vents’. *Molecular Ecology* 24 (3): 673–89. https://doi.org/10.1111/mec.13054.

Ji, Yinqiu, Louise Ashton, Scott M Pedley, David P Edwards, Yong Tang, Akihiro Nakamura, Roger Kitching, et al. 2013. ‘Reliable, Verifiable and Efficient Monitoring of Biodiversity via Metabarcoding’. *Ecology Letters* 16 (10): 1245–57. https://doi.org/10.1111/ele.12162.

Klappenbach, J A, Paul R. Saxman, Cole James R., and Thomas M. Schmidt. 2001. ‘Rrndb: The Ribosomal RNA Operon Copy Number Database’. *Nucleic Acids Research* 29 (1): 181–84. https://doi.org/10.1093/nar/29.1.181.

Leray, Matthieu, J Y Yang, C P Meyer, S C Mills, N Agudelo, V Ranwez, J T Boehm, and Ryuji J. Machida. 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. https://doi.org/10.1186/1742-9994-10-34.

Macheriotou, Lara, Katja Guilini, Tania Nara Bezerra, Bjorn Tytgat, Dinh Tu Nguyen, Thi Xuan Phuong Nguyen, Febe Noppe, et al. 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, Ryuji J., and Nancy Knowlton. 2012. ‘PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences’. Edited by Jack Anthony Gilbert. *PLoS ONE* 7 (9): e46180. https://doi.org/10.1371/journal.pone.0046180.

Machida, Ryuji J., Matthew Kweskin, and Nancy Knowlton. 2012. ‘PCR Primers for Metazoan

Mitochondrial 12S Ribosomal DNA Sequences’. *PLoS ONE* 7 (4). https://doi.org/10.1371/journal.pone.0035887.

Machida, Ryuji J., Matthieu Leray, Shian Lei Ho, and Nancy Knowlton. 2017. ‘Data Descriptor: Metazoan Mitochondrial Gene Sequence Reference Datasets for Taxonomic Assignment of Environmental Samples’. *Scientific Data* 4. https://doi.org/10.1038/sdata.2017.27.

Mahe, F, Torbjørn Rognes, C Quince, Colomban De Vargas, and M Dunthorn. 2015. ‘Swarm v2: Highly-Scalable and High-Resolution Amplicon Clustering’. *PeerJ* 3. https://doi.org/Artn E142010.7717/Peerj.1420.

Massana, Ramón Ramon, Angélique Gobet, Stéphane Audic, David Bass, Lucie Bittner, Christophe Boutte, Aurélie Chambouvet, et al. 2015. ‘Marine Protist Diversity in European Coastal Waters and Sediments as Revealed by High-Throughput Sequencing’. *Environmental Microbiology* 17 (10): 4035–49. https://doi.org/10.1111/1462-2920.12955.

Mayr, Ernst. 1942. *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*. New York, NY: Columbia University Press. http://www.hup.harvard.edu/catalog.php?isbn=9780674862500.

McMurdie, Paul J., and Susan Holmes. 2013. ‘Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data’. Edited by Michael Watson. *PLoS ONE* 8 (4): e61217. https://doi.org/10.1371/journal.pone.0061217.

Minoche, André E, Juliane C Dohm, and Heinz Himmelbauer. 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, Jacob T., Gavin M. Douglas, André M. Comeau, and Morgan G.I. Langille. 2018. ‘Denoising the Denoisers: An Independent Evaluation of Microbiome Sequence Error-Correction Approaches’. *PeerJ* 6: e5364. https://doi.org/10.7717/peerj.5364.

Nichols, Ruth V., Christopher Vollmers, Lee A. Newsom, Yue Wang, Peter D. Heintzman, McKenna Leighton, Richard E. Green, and Beth Shapiro. 2018. 'Minimizing Polymerase Biases in Metabarcoding'. *Molecular Ecology Resources* 18 (5): 927–39. https://doi.org/10.1111/1755-0998.12895.

Oksanen, Jari, Michael Blanchet, Guillaume F. Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, R. Peter Minchin, R.B. O'Hara, et al. 2018. 'Vegan: Community Ecology Package'. https://cran.r-project.org/package=vegan.

Pansu, Johan, Charline Giguet-Covex, Francesco Ficetola, Ludovic Gielly, Frederic Boyer, Eric Coissac, Isabelle Domaizon, Lucie Zinger, Jerome Poulenard, and Fabien Arnaud. 2015. 'Environmental DNA Metabarcoding to Investigate Historic Changes in Biodiversity'. *Genome* 58 (5): 264.

Parada, A E, D M Needham, and J A Fuhrman. 2016. 'Every Base Matters: Assessing Small Subunit RRNA Primers for Marine Microbiomes with Mock Communities, Time Series and Global Field Samples'. *Environ Microbiol* 18 (5): 1403–14. https://doi.org/10.1111/1462-2920.13023.

Pawlowski, Jan W., Richard Christen, Beatrice Lecroq, Dipankar Bachar, Hamid Reza Shahbazkia, Linda Amaral-Zettler, and Laure Guillou. 2011. 'Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing'. *PLoS One* 6 (4). https://doi.org/e1816910.1371/journal.pone.0018169.

Pei, Anna Y., William E Oberdorf, Carlos W Nossa, Ankush Agarwal, Pooja Chokshi, Erika A Gerz, Zhida Jin, et al. 2010. 'Diversity of 16S RRNA Genes within Individual Prokaryotic Genomes'. *Applied and Environmental Microbiology* 76 (12): 3886–97. https://doi.org/10.1128/AEM.02953-09.

Phillips, Jarrett D., Daniel J. Gillis, and Robert H. Hanner. 2019. 'Incomplete Estimates of

Genetic Diversity within Species: Implications for DNA Barcoding'. *Ecology and Evolution*. John Wiley & Sons, Ltd. https://doi.org/10.1002/ece3.4757.

Plouviez, S., T. M. Shank, B. Faure, C. Daguin-Thiebaut, F. Viard, F. H. Lallier, and D. Jollivet. 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): 3903–17. https://doi.org/10.1111/j.1365-294X.2009.04325.x.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2012. 'The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools'. *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.

Queiroz, Kevin de. 2005. 'Ernst Mayr and the Modern Concept of Species'. *Proceedings of the National Academy of Sciences* 102 (Supplement 1): 6600–6607. https://doi.org/10.1073/pnas.0502030102.

R Core Team. 2018. 'R: A Language and Environment for Statistical Computing.' R Foundation for Statistical Computing, Vienna, Austria.

Salazar, Guillem, Francisco M Cornejo-Castillo, Veronica Benitez-Barrios, Eugenio Fraile-Nuez, X Anton Alvarez-Salgado, Carlos M Duarte, Josep M Gasol, and Silvia G Acinas. 2016. 'Global Diversity and Biogeography of Deep-Sea Pelagic Prokaryotes'. *Isme Journal* 10 (3): 596–608. https://doi.org/10.1038/ismej.2015.137.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. 'Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities'. *Applied and Environmental Microbiology* 75 (23): 7537–41. https://doi.org/10.1128/AEM.01541-09.

Schnell, Ida Bærholm, Kristine Bohmann, and M. Thomas P. Gilbert. 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, Timothy M., Michael B. Black, Kenneth M. Halanych, Richard A. Lutz, and Robert C. Vrijenhoek. 1999. 'Miocene Radiation of Deep-Sea Hydrothermal Vent Shrimp (Caridea: Bresiliidae): Evidence from Mitochondrial Cytochrome Oxidase Subunit I'. *Molecular Phylogenetics and Evolution* 13 (2): 244–54. https://doi.org/10.1006/mpev.1999.0642.

Shapiro, B. Jesse, Jean Baptiste Leducq, and James Mallet. 2016. 'What Is Speciation?' Edited by Ivan Matic. *PLoS Genetics* 12 (3): e1005860. https://doi.org/10.1371/journal.pgen.1005860.

Sinniger, Frederic, Jan W. Pawlowski, Saki Harii, Andrew J. Gooday, Hiroyuki Yamamoto, Pierre Chevaldonné, Tomas Cedhagen, Gary Carvalho, and Simon Creer. 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, Viviane, Charlotte Hopfe, Clemens L Weiß, Fabrizio Mafessoni, Marco De La Rasilla, Carles Lalueza-Fox, Antonio Rosas, et al. 2017. 'Neandertal and Denisovan DNA from Pleistocene Sediments'. *Science* 356 (6338): 605–8. https://doi.org/10.1126/science.aam9695.

Sokal, Robert R., and Theodore J. Crovello. 1970. 'The Biological Species Concept : A Critical Evaluation'. *The American Naturalist* 104 (936): 127–53.

Song, Hojun, Jennifer E Buhay, Michael F Whiting, and Keith A Crandall. 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): 13486–91. https://doi.org/10.1073/pnas.0803076105.

Stat, Michael, Megan J. Huggett, Rachele Bernasconi, Joseph D. Dibattista, Tina E. Berry, Stephen J. Newman, Euan S. Harvey, and Michael Bunce. 2017. 'Ecosystem Biomonitoring with EDNA: Metabarcoding across the Tree of Life in a Tropical Marine Environment'. *Scientific Reports* 7. https://doi.org/10.1038/s41598-017-12501-5.

Stefanni, Sergio, David Stanković, Diego Borme, Alessandra de Olazabal, Tea Juretić, Alberto Pallavicini, and Valentina Tirelli. 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, Pierre, Eric Coissac, Mehrdad Hajibabaei, and Loren H. Rieseberg. 2012. 'Environmental DNA'. *Molecular Ecology* 21 (8): 1789–93. https://doi.org/10.1111/j.1365-294X.2012.05542.x.

Tang, Cuong Q., Francesca Leasi, Ulrike Obertegger, Alexander Kieneke, Timothy G Barraclough, and Diego Fontaneto. 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): 16208–12. https://doi.org/10.1073/pnas.1209160109.

Teixeira, Sara, Karine Olu, Carole Decker, Regina L Cunha, Sandra Fuchs, Stéphane Hourdez, Ester A. Serrão, and Sophie Arnaud-Haond. 2013. 'High Connectivity across the Fragmented Chemosynthetic Ecosystems of the Deep Atlantic Equatorial Belt: Efficient Dispersal Mechanisms or Questionable Endemism?' *Molecular Ecology* 22 (18): 4663–80. https://doi.org/10.1111/mec.12419.

Valentini, Alice, François Pompanon, and Pierre Taberlet. 2009. 'DNA Barcoding for

Ecologists'. *Trends in Ecology and Evolution*. Elsevier Current Trends. https://doi.org/10.1016/j.tree.2008.09.011.

Valentini, Alice, Pierre Taberlet, Claude Miaud, Raphaël Raphael Civade, Jelger Herder, Philip Francis Thomsen, Eva Bellemain, et al. 2016. 'Next-Generation Monitoring of Aquatic Biodiversity Using Environmental DNA Metabarcoding'. *Molecular Ecology* 25 (4): 929–42. https://doi.org/10.1111/mec.13428.

Vargas, Colomban De, Stéphane Audic, Nicolas Henry, Johan Decelle, Frédéric Mahé, Ramiro Logares, Enrique Lara, et al. 2015. 'Eukaryotic Plankton Diversity in the Sunlit Ocean'. *Science* 348 (6237). https://doi.org/10.1126/science.1261605.

Wangensteen, Owen S., and Xavier Turon. 2016. 'Metabarcoding Techniques for Assessing Biodiversity of Marine Animal Forests'. In *Marine Animal Forests*, edited by S. Rossi, L. Bramanti, A. Gori, and C. Orejas Saco del Valle, 1–29. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17001-5_53-1.

Xiong, Wei, and Aibin Zhan. 2018. 'Testing Clustering Strategies for Metabarcoding-Based Investigation of Community–Environment Interactions'. *Molecular Ecology Resources* 18 (6): 1326–38. https://doi.org/10.1111/1755-0998.12922.

Yoccoz, N G, K A Brathen, L Gielly, J Haile, M E Edwards, T Goslar, H von Stedingk, et al. 2012. 'DNA from Soil Mirrors Plant Taxonomic and Growth Form Diversity'. *Molecular Ecology* 21 (15): 3647–55. https://doi.org/10.1111/j.1365-294X.2012.05545.x.

Yu, Douglas W, Yinqiu Ji, Brent C Emerson, Xiaoyang Wang, Chengxi Ye, Chunyan Yang, and Zhaoli Ding. 2012. 'Biodiversity Soup: Metabarcoding of Arthropods for Rapid Biodiversity Assessment and Biomonitoring'. *Methods in Ecology and Evolution* 3 (4): 613–23. https://doi.org/10.1111/j.2041-210X.2012.00198.x.

Zanol, Joana, Kenneth M. Halanych, Torsten H. Struck, and Kristian Fauchald. 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): 660–76. https://doi.org/10.1016/j.ympev.2009.12.024.

Zhan, Aibin, Sarah A. Bailey, Daniel D. Heath, and Hugh J. Macisaac. 2014. 'Performance Comparison of Genetic Markers for High-Throughput Sequencing-Based Biodiversity Assessment in Complex Communities'. *Molecular Ecology Resources* 14 (5): 1049–59. https://doi.org/10.1111/1755-0998.12254.

Zinger, Lucie, Jérôme Chave, Eric Coissac, Amaia Iribar, Eliane Louisanna, Sophie Manzi, Vincent Schilling, Heidy Schimann, Guilhem Sommeria-Klein, and Pierre Taberlet. 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.soilbio.2016.01.008.

DATA ACCESSIBILITY

The data for this work can be accessed in the European Nucleotide Archive (ENA) database (Study accession number will be given upon manuscript acceptance). The data set, including sequences, databases, as well as raw and refined ASV/OTU tables, has been deposited on ftp://ftp.ifremer.fr/ifremer/dataref/bioinfo/merlin/abyss/BioinformaticPipelineComparisons/. Bioinformatic scripts, config files, and R scripts are available on Gitlab (https://gitlab.ifremer.fr/abyss-project/).

AUTHOR CONTRIBUTIONS

MIB and SAH designed the study, MIB and JP carried out the laboratory and molecular work; MIB and BT performed the bioinformatic and statistical analyses. LQ assisted in the bioinformatic development and participated in the study design. MIB and SAH wrote the manuscript. All authors contributed to the final manuscript.