Molecular Ecology Resources
August 2021, Volume 21, Issue 6, Pages 1904-1921
https://doi.org/10.1111/1755-0998.13398
https://archimer.ifremer.fr/doc/00688/80057/

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

Brandt Miriam [1, *], Trouche Blandine [2], Quintric Laure [3], Günther Babett [6], Wincker Patrick [4, 5], Poulain Julie [4, 5], Arnaud-haond Sophie [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, Genoscope, 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, France
[6] MARBEC, Ifremer Univ. Montpellier IRD CNRS Sète, France

* Corresponding authors : Miriam Brandt, email address : miriam.isabelle.brandt@gmail.com ; Sophie Arnaud-Haond, email address : sarnaud@ifremer.fr

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 validate 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 datasets. 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, Multi-marker metabarcoding, swarm

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; 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 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 for improved detection probability (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, 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 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 (*minimum match*) and co-occurrence rate (*minimum relative cooccurence*) with more abundant clusters, allowing the acquisition of 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). Yet, 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 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 datasets 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 (Escudié 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 fine-scale clustering. Finally, it is widely recognized that homogeneous entities sharing a set of evolutionary and ecological properties, i.e. *species* (Mayr 1942; Queiroz, de 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 (Parada et al. 2016) on these environmental samples.

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

Two 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 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, 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. Four (16S), eight (18S), and ten (COI) control PCRs were performed alongside sample PCRs, depending on the amount of trials needed to achieve successful amplification.

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. 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 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 reactions 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 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

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, 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 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 (Escudié 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.

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 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 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 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. 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. 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/assignation). 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 (Clare et al. 2016; Atienza et al. 2020; Turon et al. 2020; Djurhuus et al. 2017; Cordier et al. 2019; Sawaya et al. 2019; Wood et al. 2019; Laroche et al. 2018; Andújar et al. 2018a). 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. Non-target clusters (i.e. either non-metazoan or non-bacterial) 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, Diplommatinidae, 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 (Wangensteen and 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 dataset (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 100/1000 times less abundant than a potential parent OTU.

The vast majority of prokaryotes usually show low levels (< 1%) 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 low *d*-values. 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).

In order to have reliable BLAST phylum assignments for pipeline comparison, final datasets 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).

1.4 Statistical analyses

Data was 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 (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 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'. These non-target 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 datasets 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 and Orme 2012). The effect of site and LULU curation on community composition was tested by PERMANOVA, using the function *adonis*2 (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 datasets 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 dataset, 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. ≥ 80% / 86% similarity, ≥ 80% phylum-level bootstraps).

2 RESULTS

2.1 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 non-target 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 ten mock species were detected in the COI dataset (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 non-clustered datasets, 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 datasets 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 dataset, 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 1000 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 dataset at both *minimum match* levels and *Paralepetopsis* sp. with *minimum match* at 90% (Table 1). With *minimum ratio* at 1000, both species were retained at both *minimum match* levels but more OTUs were retained for another species (*Munidopsis* sp., Table S6). As LULU curation with higher *minimum ratio* levels resulted in more accurate species compositions in the mock samples with 18S, we only present LULU curation with *minimum ratio* = 100 for the environmental samples.

2.2 Alpha-diversity patterns in environmental samples

High-throughput sequencing results

A total 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 (16S), 8 (18S), and 10 (COI) PCR blanks (Table S4). The final datasets 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 dataset produced 53,815 target ASVs and 12,800 – 38,972 OTUs after swarm clustering (d = 1-11).

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, Fig. 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 (Fig. 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 (Fig. 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 (Fig. 1, Table 2). Similarly, LULU curation of ASVs resulted in significantly more clusters than LULU curation of OTUs produced with any d-value (Fig. 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 (Fig S1).

2.3 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 non-curated data, accounting for < 1% of variation in data (Fig. 2).

Although ASV and OTU datasets detected similar amounts of variation due to sites in PERMANOVAs, clustering levels affected the ecological patterns resolved by ordinations in rRNA loci (Fig 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 (Fig. 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 (Fig. 2, Fig. S2).

2.4 Taxonomic assignment quality

Assigning with BLAST resulted in mock community assignments comparable to described above. With COI, eight of the ten 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 (Fig S3). With 18S, seven species were recovered (4 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 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, RDP results were comparable to BLAST, with seven species correctly assigned at genus-level. Assigning the 18S dataset 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 dataset, but BLAST assigned 20% (18S) and 70% (COI) less OTUs at phylum-level than RDP in the metazoan datasets, even at minimum hit identity of 70% (Table S8). BLAST hit identities of the overall datasets varied strongly depending on phyla and marker gene (Fig. 3). For 18S, 90% of metazoan OTUs had assignment identities ≥ 86%, corresponding roughly to accurate phylum-level (Stefanni et al. 2018; Edgar 2017). Only 34% had reliable genus-level assignments, i.e. 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% (Fig. 3).

3 DISCUSSION

3.1 ASVs *vs* 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 and 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 datasets 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 (Fig. 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*, i.e. "*a lineage or group of 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 e.g., the vesicomyid bivalves only being identified up to class-level in clustered datasets 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 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 ~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 (Fig. 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 fine-scale sequence variants. This is in accordance with other studies reporting differences in beta diversity patterns in ASV *vs* OTU datasets for ribosomal loci, when large divergence thresholds were used for clustering (Xiong and Zhan 2018; Bokulich et al. 2013). 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 (Jost 2008; Beaumont and Nichols 1996) 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).

3.2 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 non-ideal design of the mock samples. Indeed, LULU should be applied on datasets 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 1000, the loss of that species at 84% may also just reflect the non-ideal 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 datasets. 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, Fig. 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 datasets, 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, likely 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 datasets led to significantly more OTUs 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.

3.3 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. 2018b). 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 18S barcode marker performed better in the detection of nematodes, annelids, platyhelminths, and xenacoelomorphs while COI mostly detected cnidarians, molluscs, and poriferans (Fig. 3, Fig. 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 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 assignments were observed using the RDP Bayesian Classifier on the mock samples for 18S and for COI when using the MIDORI-UNIQUE marine-only database (Fig. S3). 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. Indeed, small databases, 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 (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 (Fig. 3). Minimum similarities of 80% for COI and 86% for 18S as cut-off values for metazoans have been used to improve the taxonomic quality of metazoan metabarcoding datasets (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 (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. When using the reduced marine-only COI database, RDP provided the most accurate assignments in the mock samples (Fig. S3). However, filtering to accurate phylum-level bootstraps (≥ 80) drastically reduced the number of OTUs in the overall dataset (1% of OTUs left, Table S8). 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-level bootstraps < 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 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.

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 datasets 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 (incl. 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 datasets, 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 datasets, 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.

DATA ACCESSIBILITY

The raw data of this work can be accessed in the European Nucleotide Archive (ENA) database (project: PRJEB33873), please refer to the supplementary metadata excel sheet for ENA file names. The dataset, 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/).

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).

CONFLICT OF INTEREST DISCLOSURE

The authors declare that they have no financial conflict of interest with the content of this article.

REFERENCES

Acinas, S.G., L.A. Marcelino, V. Klepac-Ceraj, and M.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, A., O. Aizpurua, M.T.P. Gilbert, and K. Bohmann. 2017. "Scrutinizing Key Steps for Reliable Metabarcoding of Environmental Samples." Edited by Mahon, A. *Methods in Ecology and Evolution*, 2017. https://doi.org/10.1111/2041-210X.12849.

Andújar, C., P. Arribas, C. Gray, C. Bruce, G. Woodward, D.W. Yu, and A.P. Vogler. 2018a. "Metabarcoding of Freshwater Invertebrates to Detect the Effects of a Pesticide Spill." *Molecular Ecology* 27 (1): 146–66. https://doi.org/10.1111/mec.14410.

Andújar, C., P. Arribas, D.W. Yu, A.P. Vogler, and B.C. Emerson. 2018b. "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.

Atienza, S., M. Guardiola, K. Præbel, A. Antich, X. Turon, and O.S. Wangensteen. 2020. "DNA Metabarcoding of Deep-Sea Sediment Communities Using COI: Community Assessment, Spatio-Temporal Patterns and Comparison with 18S RDNA." *Diversity* 12 (4). https://doi.org/10.3390/D12040123.

Baselga, A., and C.D.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, E., S. Glémin, and N. Galtier. 2006. "Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals." *Science* 312 (5773): 570–72. https://doi.org/10.1126/science.1122033.

Beaumont, M.A., and R.A. Nichols. 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): 1619–26. https://doi.org/10.1098/rspb.1996.0237.

Bensasson, D., D.X. Zhang, D.L. Hartl, and G.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, H.M., W. Sung, P. De Ley, J.G. Baldwin, J. Sharma, A. Rocha-Olivares, and W.K. 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, I., G. Carvalho, K. Walsh, M. Christmas, M. Hajibabaei, P. Kille, D. Lallias, and S. Creer. 2015. "Monitoring Lake Ecosystem Health Using Metabarcoding of Environmental DNA: Temporal Persistence and Ecological Relevance." *Genome* 58 (5): 197.

Bokulich, N.A., S. Subramanian, J.J. Faith, D. Gevers, J.I. Gordon, R. Knight, D.A. Mills, and J.G. 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, G., J. Bakker, O.S. Wangensteen, S. Mariani, L. Bonnin, J.B. Juhel, J.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.

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, A., D. Steinke, and L. 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, B.J., P.J. McMurdie, and S.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.

Callahan, B.J., P.J. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, and S.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.

Carranza, S., G. Giribet, C. Ribera, J. Baguñà, and M. 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, L., C. Corinaldesi, A. Dell'Anno, and R. Danovaro. 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., F.J.J. Chain, J.E. Littlefair, and M.E. Cristescu. 2016. "The Effects of Parameter Choice on Defining Molecular Operational Taxonomic Units and Resulting Ecological Analyses of Metabarcoding Data." Edited by Deiner, K. *Genome* 59 (11): 981–90. https://doi.org/10.1139/gen-2015-0184.

Clarke, L.J., J.M. Beard, K.M. Swadling, and B.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, F.M. 2001. "Bacterial Species and Speciation." Edited by Kane, M. *Systematic Biology* 50 (4): 513–24. https://doi.org/10.1080/10635150118398.

Coissac, E., T. Riaz, and N. 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.

Cordier, T., F. Frontalini, K. Cermakova, L. Apothéloz-Perret-Gentil, M. Treglia, E. Scantamburlo, V.

Bonamin, and J.W. Pawlowski. 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., M. Pinheiro, O. Mouchel, M. Maguer, J. Grall, J. Miné, and S. 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, S., K. Deiner, S. Frey, D. Porazinska, P. Taberlet, W.K. Thomas, C. Potter, and H.M. Bik. 2016. "The Ecologist's Field Guide to Sequence-Based Identification of Biodiversity." Edited by Freckleton, R. *Methods in Ecology and Evolution* 7 (9): 1008–18. https://doi.org/10.1111/2041-210X.12574.

Davis, N.M., D.M. Proctor, S.P. Holmes, D.A. Relman, and B.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, K., E.A. Fronhofer, E. Mächler, J.C. Walser, and F. 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, K., J.-C.C. Walser, E. Machler, F. Altermatt, E. Mächler, and F. Altermatt. 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., A. Valentini, A. Duparc, S. Pellier-Cuit, F. Pompanon, P. Taberlet, and C. Miaud. 2011. "Persistence of Environmental DNA in Freshwater Ecosystems." *PLoS One* 6 (8). https://doi.org/10.1371/journal.pone.0023398.

Dickie, I.A., S. Boyer, H.L. Buckley, R.P. Duncan, P.P. Gardner, I.D. Hogg, R.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.

Djurhuus, A., J. Port, C.J. Closek, K.M. Yamahara, O.C. Romero-Maraccini, K.R. Walz, D.B. Goldsmith, et al. 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., 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.

Dufresne, Y., F. Lejzerowicz, L.A. Perret-Gentil, J.W. Pawlowski, and T. Cordier. 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 [Preprint]*, 124156. https://doi.org/10.1101/124156.

Eren, A.M., J.H. Vineis, H.G. Morrison, and M.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, F., L. Auer, M. Bernard, M. Mariadassou, L. Cauquil, K. Vidal, S. Maman, et al. 2018. "FROGS: Find, Rapidly, OTUs with Galaxy Solution." Edited by Berger, B. *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, G.F., J. Pansu, A. Bonin, E. Coissac, C. Giguet-Covex, M. De Barba, L. 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, V.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, V.G., G.R. Carvalho, W. Sung, H.F. Johnson, D.M. Power, S.P. Neill, M. Packer, et al. 2010. "Second-Generation Environmental Sequencing Unmasks Marine Metazoan Biodiversity."

Nature Communications 1. https://doi.org/9810.1038/ncomms1095.

Freudenstein, J. V., M.B. Broe, R.A. Folk, and B.T. Sinn. 2017. "Biodiversity and the Species Concept - Lineages Are Not Enough." *Systematic Biology* 66 (4): 644–56. https://doi.org/10.1093/sysbio/syw098.

Frøslev, T.G., R. Kjøller, H.H. Bruun, R. Ejrnæs, A.K. Brunbjerg, C. Pietroni, and A.J. 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, D., F.M. Cohan, J.G. Lawrence, B.G. Spratt, T. Coenye, E.J. Feil, E. Stackebrandt, et al. 2005. "Re-Evaluating Prokaryotic Species." *Nature Reviews Microbiology* 3 (9): 733–39. https://doi.org/10.1038/nrmicro1236.

Goldberg, C.S., C.R. Turner, K. Deiner, K.E. Klymus, P.F. Thomsen, M.A. Murphy, S.F. Spear, et al. 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., B.S. Stevenson, and T.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, P.D.N., S. Ratnasingham, and J.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, S., H. Watanabe, and T.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, Y., L. Ashton, S.M. Pedley, D.P. Edwards, Y. Tang, A. Nakamura, R. 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.

Jost, L. 2007. "Partitioning Diversity into Independent Alpha and Beta Components." *Ecology* 88 (10): 2427–39. https://doi.org/10.1890/06-1736.1.

———. 2008. "GST and Its Relatives Do Not Measure Differentiation." *Molecular Ecology* 17 (18): 4015–26. https://doi.org/10.1111/j.1365-294X.2008.03887.x.

Klappenbach, J.A., P.R. Saxman, C.J. R., and T.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.

Laroche, O., S.A. Wood, L.A. Tremblay, J.I. Ellis, G. Lear, and X. Pochon. 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., J.Y. Yang, C.P. Meyer, S.C. Mills, N. Agudelo, V. Ranwez, J.T. Boehm, and R.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, L., K. Guilini, T.N. Bezerra, B. Tytgat, D.T. Nguyen, T.X. Phuong Nguyen, F. 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, R.J., and N. Knowlton. 2012. "PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences." Edited by Gilbert, J.A. *PLoS ONE* 7 (9): e46180. https://doi.org/10.1371/journal.pone.0046180.

Machida, R.J., M. Kweskin, and N. Knowlton. 2012. "PCR Primers for Metazoan Mitochondrial 12S Ribosomal DNA Sequences." *PLoS ONE* 7 (4). https://doi.org/10.1371/journal.pone.0035887.

Machida, R.J., M. Leray, S.L. Ho, and N. 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.

Mahé, F., T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. 2015. "Swarm v2: Highly-Scalable and High-Resolution Amplicon Clustering." *PeerJ* 3 (December): e1420. https://doi.org/10.7717/peerj.1420.

Massana, R.R., A. Gobet, S. Audic, D. Bass, L. Bittner, C. Boutte, A. 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, E. 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, P.J., and S. Holmes. 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., J.C. Dohm, and H. 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, J.T., G.M. Douglas, A.M. Comeau, and M.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, R. V., C. Vollmers, L.A. Newsom, Y. Wang, P.D. Heintzman, M. Leighton, R.E. Green, and B. Shapiro. 2018. "Minimizing Polymerase Biases in Metabarcoding." *Molecular Ecology Resources* 18 (5): 927–39. https://doi.org/10.1111/1755-0998.12895.

Oksanen, J., M. Blanchet, Guillaume F. Friendly, R. Kindt, P. Legendre, D. McGlinn, R.P. Minchin, R.B. O'Hara, et al. 2018. "Vegan: Community Ecology Package." https://cran.r-project.org/package=vegan.

Pansu, J., C. Giguet-Covex, F. Ficetola, L. Gielly, F. Boyer, E. Coissac, I. Domaizon, L. Zinger, J. Poulenard, and F. 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, J.W., R. Christen, B. Lecroq, D. Bachar, H.R. Shahbazkia, L. Amaral-Zettler, and L. Guillou. 2011. "Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing." *PLoS*

One 6 (4). https://doi.org/e1816910.1371/journal.pone.0018169.

Pei, A.Y., W.E. Oberdorf, C.W. Nossa, A. Agarwal, P. Chokshi, E.A. Gerz, Z. 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, J.D., D.J. Gillis, and R.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, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F.O. 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, K. 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. https://www.r-project.org/.

Ritari, J., J. Salojärvi, L. Lahti, and W.M. de Vos. 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., F.M. Cornejo-Castillo, V. Benitez-Barrios, E. Fraile-Nuez, X.A. Alvarez-Salgado, C.M. Duarte, J.M. Gasol, and S.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.

Sawaya, N.A., A. Djurhuus, C.J. Closek, M. Hepner, E. Olesin, L. Visser, C. Kelble, K. Hubbard, and M. Breitbart. 2019. "Assessing Eukaryotic Biodiversity in the Florida Keys National Marine Sanctuary through Environmental DNA Metabarcoding." *Ecology and Evolution* 9 (3): 1029–40. https://doi.org/10.1002/ece3.4742.

Schnell, I.B., K. Bohmann, and M.T.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, T.M., M.B. Black, K.M. Halanych, R.A. Lutz, and R.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.J., J.B. Leducq, and J. Mallet. 2016. "What Is Speciation?" Edited by Matic, I. *PLoS Genetics* 12 (3): e1005860. https://doi.org/10.1371/journal.pgen.1005860.

Sinniger, F., J.W. Pawlowski, S. Harii, A.J. Gooday, H. Yamamoto, P. Chevaldonné, T. Cedhagen, G. Carvalho, and S. 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, V., C. Hopfe, C.L. Weiß, F. Mafessoni, M. De La Rasilla, C. Lalueza-Fox, A. Rosas, et al. 2017. "Neandertal and Denisovan DNA from Pleistocene Sediments." *Science* 356 (6338): 605–8. https://doi.org/10.1126/science.aam9695.

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

Song, H., J.E. Buhay, M.F. Whiting, and K.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, M., M.J. Huggett, R. Bernasconi, J.D. Dibattista, T.E. Berry, S.J. Newman, E.S. Harvey, and M. 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, S., D. Stanković, D. Borme, A. de Olazabal, T. Juretić, A. Pallavicini, and V. 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, P., E. Coissac, M. Hajibabaei, and L.H. Rieseberg. 2012. "Environmental DNA." *Molecular Ecology* 21 (8): 1789–93. https://doi.org/10.1111/j.1365-294X.2012.05542.x.

Tang, C.Q., F. Leasi, U. Obertegger, A. Kieneke, T.G. Barraclough, and D. 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, S., K. Olu, C. Decker, R.L. Cunha, S. Fuchs, S. Hourdez, E.A. Serrão, and S. 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.

Turon, X., A. Antich, C. Palacín, K. Præbel, and O.S. Wangensteen. 2020. "From Metabarcoding to Metaphylogeography: Separating the Wheat from the Chaff." *Ecological Applications* 30 (2). https://doi.org/10.1002/eap.2036.

Valentini, A., F. Pompanon, and P. 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, A., P. Taberlet, C. Miaud, R.R. Civade, J. Herder, P.F. Thomsen, E. 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, C., S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, et al. 2015. "Eukaryotic Plankton Diversity in the Sunlit Ocean." *Science* 348 (6237). https://doi.org/10.1126/science.1261605.

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

Wood, S.A., X. Pochon, O. Laroche, U. von Ammon, J. Adamson, and A. Zaiko. 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): 1407–19. https://doi.org/10.1111/1755-0998.13055.

Xiong, W., and A. 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, D.W., Y. Ji, B.C. Emerson, X. Wang, C. Ye, C. Yang, and Z. 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, J., K.M. Halanych, T.H. Struck, and K. 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, A., S.A. Bailey, D.D. Heath, and H.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, L., J. Chave, E. Coissac, A. Iribar, E. Louisanna, S. Manzi, V. Schilling, H. Schimann, G. Sommeria-Klein, and P. 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.

AUTHOR CONTRIBUTIONS

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

TABLE LEGENDS

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 (i.e. 1 OTU for each mock species), light grey cells indicate a number of OTUs differing by ±3 from the number expected, dark grey cells indicate a number of OTUs > 3 times the one expected, and black cells a number ≥ 10 times the one expected. Ø indicates absence of expected OTU. Taxonomy is given up to the lowest common rank assigned to OTUs from mock species. "Others" represents unexpected OTUs, i.e. with assignments not related to any species in the mocks. These may represent contamination or symbionts of the mock species. LULU was run at *minimum ratio* = 100 for 18S and *minimum ratio* = 1 for COI.

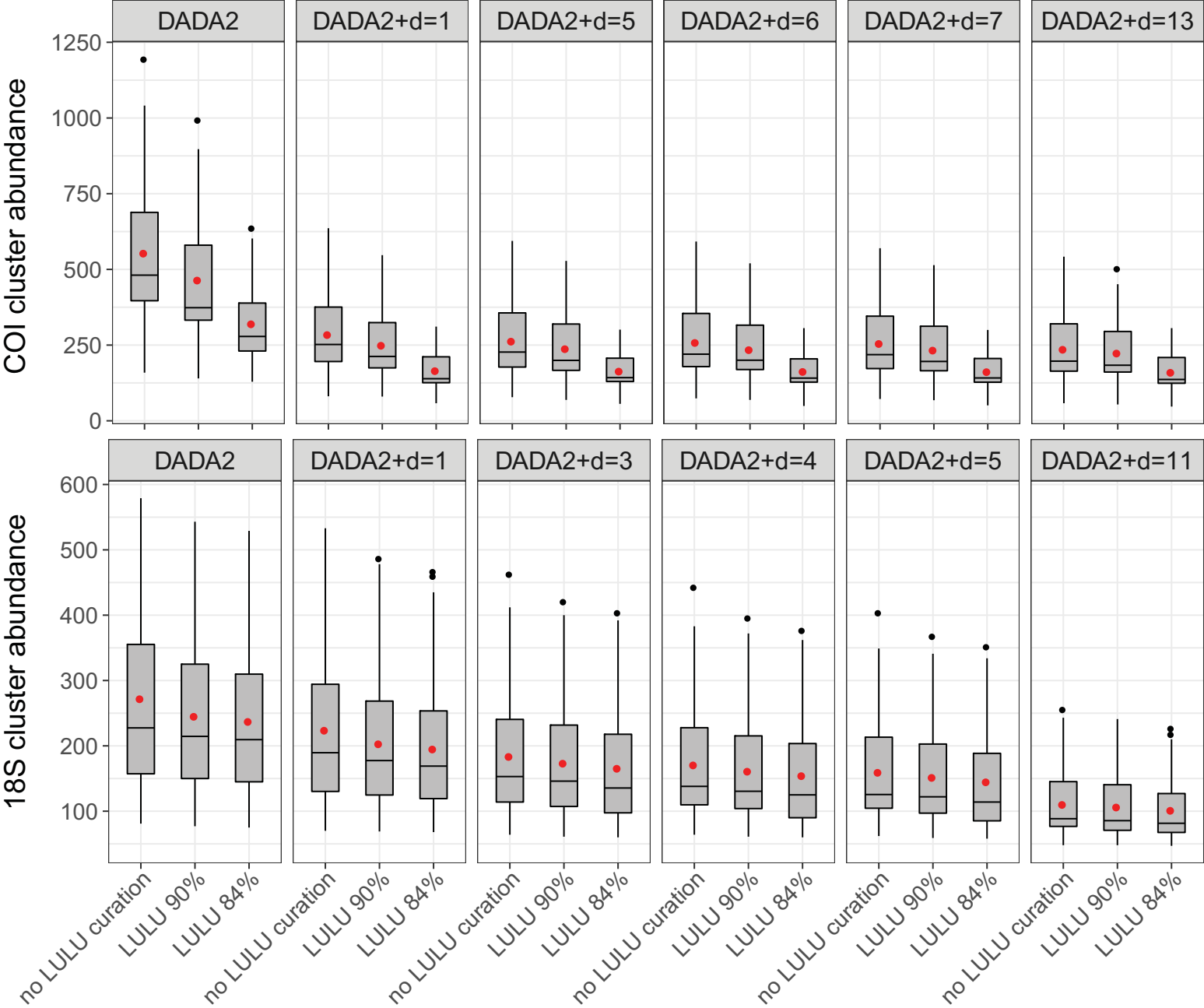
Table 2. Effect of pipeline and site on the number of metazoan and prokaryote clusters. Results of the analysis of variance (ANOVA) of the rarefied cluster richness for the three genes studied. Pairwise comparisons were performed with Tukey's HSD tests. DS: Dada2+swarm; DSL: Dada2+swarm+LULU; d: swarm *d-value*. 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: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$.

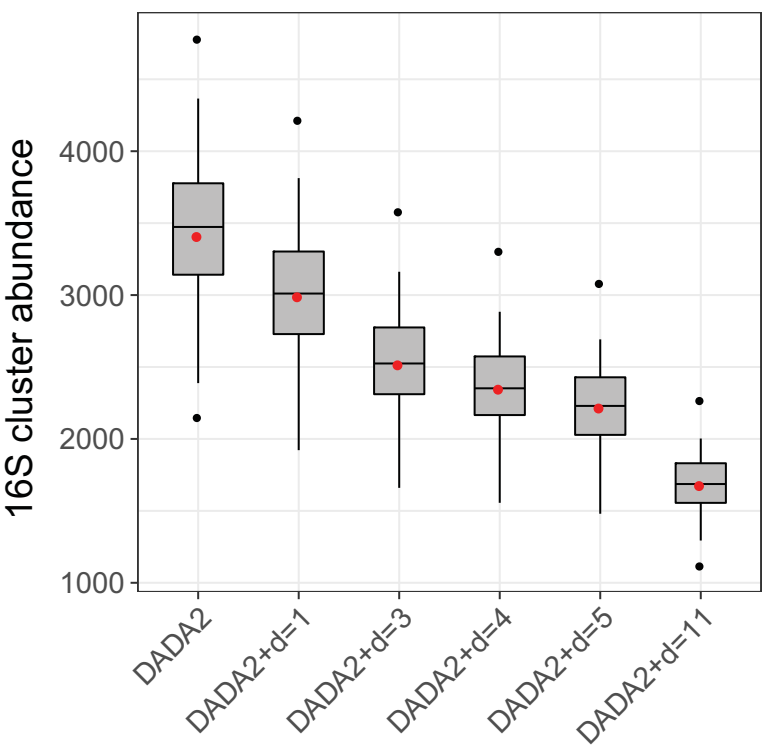
FIGURE LEGENDS

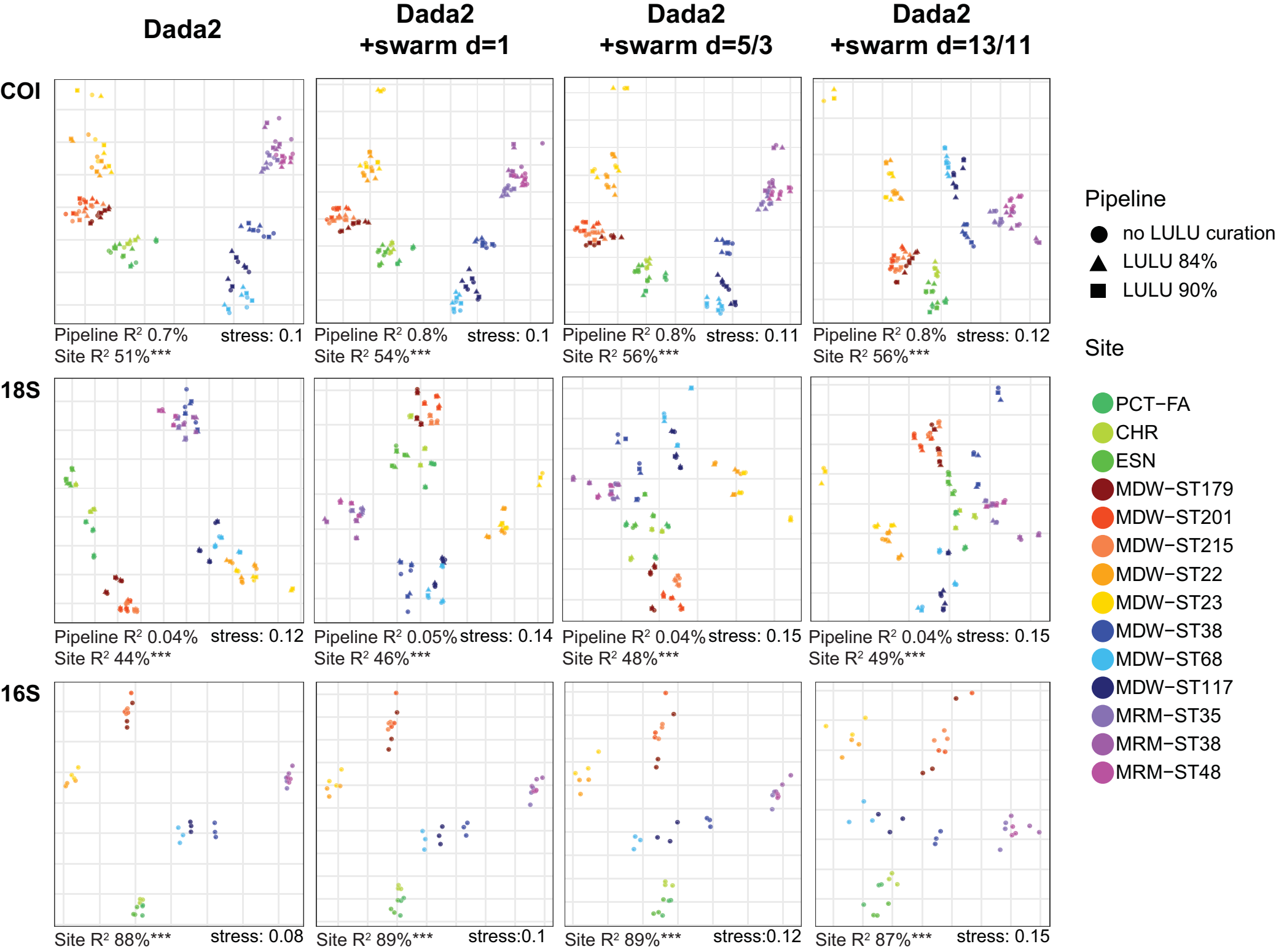
Figure 1. Number of metazoan (COI, 18S) and prokaryote (16S) clusters detected in sediment of 14 deep-sea sites with ASV *vs* OTU-centred datasets. ASVs were obtained with the DADA2 metabarcoding pipeline, and clustered with swarm at different *d* values. Metazoan ASVs and OTUs were curated with LULU 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 datasets rarefied to same sequencing depth. Boxplots represent medians with first and third quartiles. Red dots indicate means.

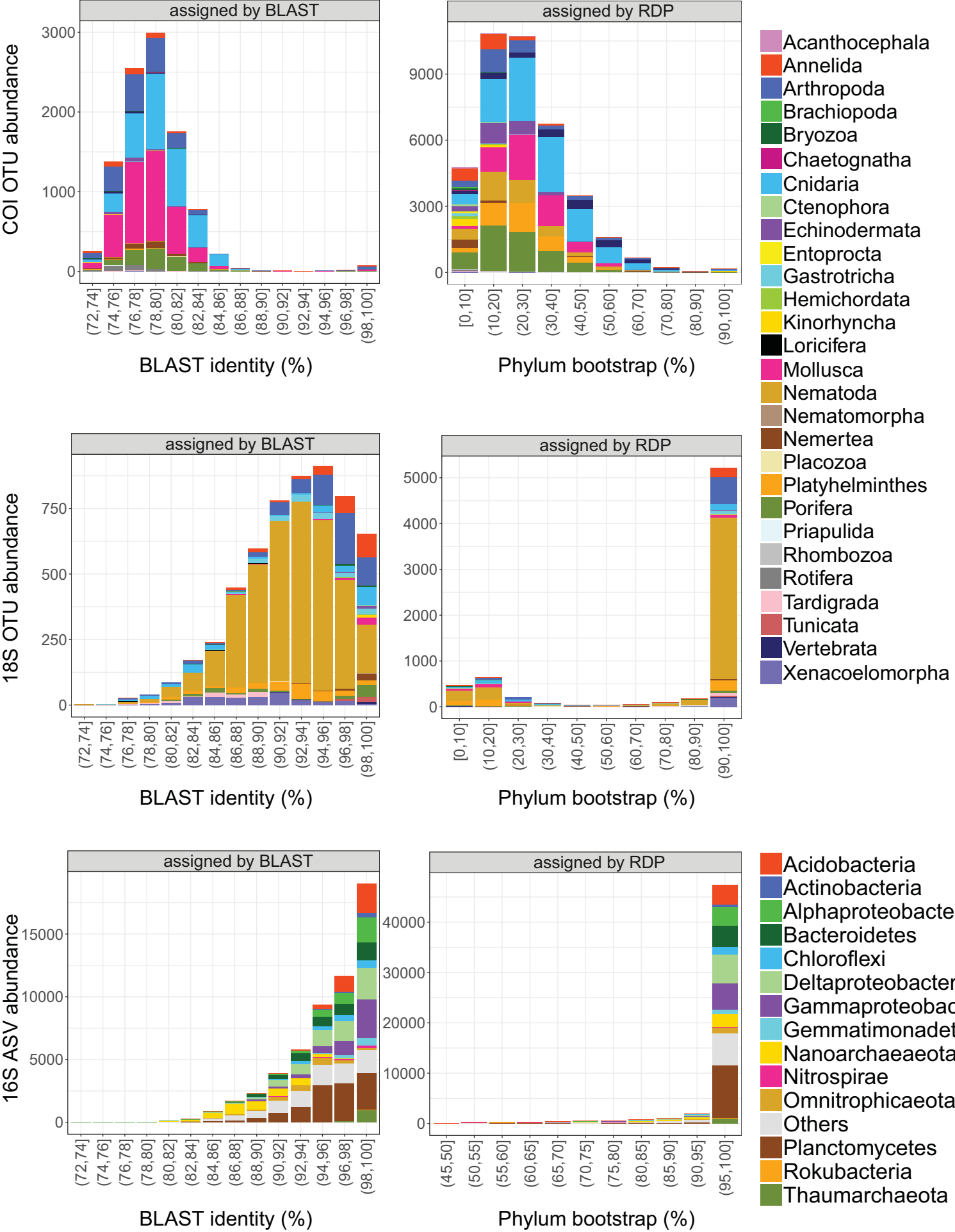
Figure 2. Metazoan (COI, 18S) and prokaryote (16S) beta-diversity patterns in ASV and OTU-centred datasets. Nonmetric multidimensional scaling (NMDS) ordinations showing community differentiation observed between sites with different clustering scenarios. ASVs were obtained with the DADA2 metabarcoding 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: ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$. 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.

Figure 3. Taxonomic assignment quality of BLAST and RDP methods on metazoan (COI, 18S) and prokaryote (16S) metabarcoding datasets 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.


COI cluster abundance
DADA2
DADA2+d=1
DADA2+d=5
DADA2+d=6
DADA2+d=7
DADA2+d=13
1250
1000
750
500
250
0
18S cluster abundance
DADA2
DADA2+d=1
DADA2+d=3
DADA2+d=4
DADA2+d=5
DADA2+d=11
600
500
400
300
200
100
no LULU curation
LULU 90%
LULU 84%


16S cluster abundance
4000
3000
2000
1000
DADA2
DADA2+d=1
DADA2+d=3
DADA2+d=4
DADA2+d=5
DADA2+d=11


Dada2
Dada2 +swarm d=1
Dada2 +swarm d=5/3
Dada2 +swarm d=13/11
COI
18S
16S
Pipeline R² 0.7% stress: 0.1
Site R² 51%***
Pipeline R² 0.8% stress: 0.1
Site R² 54%***
Pipeline R² 0.8% stress: 0.11
Site R² 56%***
Pipeline R² 0.8% stress: 0.12
Site R² 56%***
Pipeline R² 0.04% stress: 0.12
Site R² 44%***
Pipeline R² 0.05% stress: 0.14
Site R² 46%***
Pipeline R² 0.04% stress: 0.15
Site R² 48%***
Pipeline R² 0.04% stress: 0.15
Site R² 49%***
Site R² 88%*** stress: 0.08
Site R² 89%*** stress:0.1
Site R² 89%*** stress:0.12
Site R² 87%*** stress: 0.15
Pipeline
no LULU curation
LULU 84%
LULU 90%
Site
PCT−FA
CHR
ESN
MDW−ST179
MDW−ST201
MDW−ST215
MDW−ST22
MDW−ST23
MDW−ST38
MDW−ST68
MDW−ST117
MRM−ST35
MRM−ST38
MRM−ST48


assigned by BLAST
assigned by RDP
COI OTU abundance
18S OTU abundance
16S ASV abundance
BLAST identity (%)
Phylum bootstrap (%)
(72,74]
(74,76]
(76,78]
(78,80]
(80,82]
(82,84]
(84,86]
(86,88]
(88,90]
(90,92]
(92,94]
(94,96]
(96,98]
(98,100]
[0,10]
(10,20]
(20,30]
(30,40]
(40,50]
(50,60]
(60,70]
(70,80]
(80,90]
(90,100]
(45,50]
(50,55]
(55,60]
(60,65]
(65,70]
(70,75]
(75,80]
(80,85]
(85,90]
(90,95]
(95,100]
Acanthocephala
Annelida
Arthropoda
Brachiopoda
Bryozoa
Chaetognatha
Cnidaria
Ctenophora
Echinodermata
Entoprocta
Gastrotricha
Hemichordata
Kinorhyncha
Loricifera
Mollusca
Nematoda
Nematomorpha
Nemertea
Placozoa
Platyhelminthes
Porifera
Priapulida
Rhombozoa
Rotifera
Tardigrada
Tunicata
Vertebrata
Xenacoelomorpha
Acidobacteria
Actinobacteria
Alphaproteobacteria
Bacteroidetes
Chloroflexi
Deltaproteobacteria
Gammaproteobacteria
Gemmatimonadetes
Nanoarchaeaeota
Nitrospirae
Omnitrophicaeota
Others
Planctomycetes
Rokubacteria
Thaumarchaeota