bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

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

Short title:

A flexible metabarcoding pipeline based on read correction

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

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

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

[3]Ifremer, Cellule Bioinformatique, Brest, France

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

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

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

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

ABSTRACT

1 Environmental metabarcoding is an increasingly popular tool for studying biodiversity in
2 marine and terrestrial biomes. As metabarcoding with multiple markers, spanning several branches
3 of the tree of life is becoming more accessible, bioinformatic pipelines need to accommodate both
4 micro- and macro biologists. We built and tested a pipeline based on Illumina read correction with
5 DADA2 allowing analysing metabarcode data from prokaryotic and eukaryotic life compartments.
6 We implemented the option to cluster ASVs into Operational Taxonomic Units (OTUs) with
7 swarm v2, a network-based clustering algorithm, and to further curate the ASVs/OTUs based on
8 sequence similarity and co-occurrence rates using a recently developed algorithm, LULU. Finally,
9 a flexible taxonomic assignment of the Amplicon Sequence Variants (ASVs) was added *via* the
10 RDP Bayesian classifier or by BLAST. We validate this pipeline with ribosomal and mitochondrial
11 markers using eukaryotic mock communities and 42 deep-sea sediment samples. The comparison
12 of BLAST and the RDP Classifier underlined the potential of the latter to deliver very good
13 assignments, but highlighted the need for a concerted effort to build comprehensive, yet specific
14 databases adapted to the studied communities. The results underline the advantages of clustering
15 and LULU-curation for producing metazoan biodiversity inventories, and show that LULU is an
16 effective tool for filtering metazoan molecular clusters while avoiding arbitrary relative abundance
17 filters. Overall conservative estimates of diversity can be obtained using DADA2 and LULU
18 correction algorithms alone, or in combination with the clustering algorithm swarm v2 (i.e. to
19 obtain ASVs or OTUs), depending on the objective of the study.

20

21

22 Key words: Biodiversity, bioinformatics, environmental DNA, metabarcoding, mock
23 communities

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

INTRODUCTION

High-throughput sequencing (HTS) technologies are revolutionizing the way we assess biodiversity. By producing millions of DNA sequences per sample, HTS ~~now~~ allows broad taxonomic biodiversity surveys through metabarcoding of bulk DNA from complex communities or DNA directly extracted from soil, water, ~~or~~ air samples, i.e. environmental DNA (eDNA). 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 (Creer et al., 2016; Stat et al., 2017; Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012; Valentini, Pompanon, & Taberlet, 2009). In the last decade, these tools have been used to describe past and present biodiversity in terrestrial (Ji et al., 2013; Pansu et al., 2015; Slon et al., 2017; Yoccoz et al., 2012; Yu et al., 2012), freshwater (Bista et al., 2015; Deiner, Fronhofer, Mächler, Walser, & Altermatt, 2016; Dejean et al., 2011; Evans et al., 2016; Valentini et al., 2016), and marine (Bik et al., 2012; Boussarie et al., 2018; De Vargas et al., 2015; Fonseca et al., 2010; Massana et al., 2015; Pawlowski et al., 2011; Salazar et al., 2016; Sinniger et al., 2016) environments.

As every new technique brings on new challenges, a number of studies have put considerable effort into delineating critical aspects of metabarcoding protocols to ensure robust and reproducible results (see Fig.1 in Fonseca et al, 2018). Recent studies have addressed many issues regarding sampling methods (Dickie et al., 2018), contamination risks (Goldberg et al., 2016), DNA extraction protocols (Brannock & Halanych, 2015; Deiner et al., 2015; Zinger et al., 2016), amplification biases and PCR replication levels (Alberdi, Aizpurua, Gilbert, & Bohmann, 2017; Ficetola et al., 2015; Nichols et al., 2018). Similarly, computational pipelines, through which molecular data are transformed into ecological inventories of putative taxa, have also been in constant improvement. Indeed, PCR-generated errors and sequencing errors are major

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

49 bioinformatic challenges for metabarcoding pipelines, as they can strongly bias biodiversity
50 estimates (Bokulich et al., 2013; Coissac, Riaz, & Puillandre, 2012). A variety of tools have been
51 developed for quality-filtering amplicon data and removing erroneous reads to improve the
52 reliability of Illumina-sequenced metabarcode inventories (Bokulich et al., 2013; Eren, Vineis,
53 Morrison, & Sogin, 2013; Minoche, Dohm, & Himmelbauer, 2011). Studies that evaluated
54 bioinformatic parameters have generally found these quality-filtering steps, as well as arbitrarily
55 set clustering thresholds are the parameters that most strongly affect biodiversity inventories
56 produced by metabarcoding (Brannock & Halanych, 2015; Brown, Chain, Crease, MacIsaac, &
57 Cristescu, 2015; Clare, Chain, Littlefair, & Cristescu, 2016; Xiong & Zhan, 2018).

58 Recent bioinformatic algorithms for the processing of metabarcode data have been
59 developed to alleviate the influence of these two parameters. Amplicon-specific error correction
60 methods, commonly used to correct sequences produced by pyrosequencing (Coissac et al., 2012),
61 have now become available for Illumina-sequenced data. Published in 2016, DADA2 has quickly
62 become a widely used tool for Illumina sequence correction, particularly in the microbial world,
63 producing more accurate biodiversity inventories and resolving fine-scale variations by defining
64 Amplicon Sequence Variants (ASVs) (Callahan et al., 2016; Nearing, Douglas, Comeau, &
65 Langille, 2018).

66 Low abundance molecular clusters remain an issue in metabarcoding biodiversity
67 inventories, as it is challenging to discriminate valid but rare clusters from spurious ones. Singleton
68 removal (clusters with less than 1-2 total reads) is largely advocated in the metabarcoding
69 community (Clare et al., 2016) to limit the inflation of diversity due to the occurrence of spurious
70 sequences. However, this method is arbitrary and potentially hinders the detection of rare species
71 (Frøslev et al., 2017). LULU is a newly developed curation algorithm designed to filter out
72 remaining spurious clusters originating from PCR and sequencing errors, or from intra-individual

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

73 variability (pseudogenes, heteroplasmy) based on objective criteria. Spurious clusters are detected
74 based on their similarity and co-occurrence rate with more abundant clusters, allowing obtaining
75 curated datasets while avoiding arbitrary abundance filters (Frøslev et al., 2017). The authors
76 demonstrated their approach on metabarcoding of plants using ITS2 (nuclear ribosomal internal
77 transcribed spacer region 2) and comparing several pipelines. Their results show that ASV
78 definition with DADA2, subsequent clustering to address intraspecific variation, and final curation
79 with LULU is the safest pathway for obtaining reliable and accurate metabarcoding data. The
80 authors conclude that their validation on plants is relevant to other organism groups and other
81 markers, while recommending future validation of LULU on mock communities.

82 There were historically two reasons for clustering sequences into Operational Taxonomic
83 Units (OTUs). The first was to limit the bias due to PCR and sequencing errors (and to some extent
84 also intra-individual variability linked to the existence of pseudogenes) by clustering erroneous
85 (and non-target) sequences with error free target sequences. The second was to delineate OTUs as
86 clusters of sequences that would best fit a "species level", i.e. the Operational Taxonomic Units
87 defined using a classical phenetic *proxy* (Sokal & Crovello, 1970).

88 The first issue being largely solved by the two correction algorithms DADA2 and LULU,
89 the relevance of the second objective, i.e. the delineation of OTUs, is now being discussed. Indeed,
90 after presenting their new algorithm on prokaryotic communities, the authors of DADA2 proposed
91 that the reproducibility and comparability of ASVs across studies challenge the need for clustering
92 sequences, as OTUs have the disadvantage of being study-specific and defined using arbitrary
93 thresholds (Callahan, McMurdie, & Holmes, 2017).

94 Nevertheless, it is widely recognized that homogeneous entities sharing a set of
95 evolutionary and ecological properties, i.e. species (de Queiroz, 2005; Mayr, 1942), sometimes
96 proposed to be designed as "ecotypes" for prokaryotes (Cohan, 2001; Gevers et al., 2005), represent

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

97 a fundamental category of biological organization that is the cornerstone of most ecological and
98 evolutionary theories and empirical studies. ~~Keeping~~ ASV information for feeding databases and
99 cross-comparing studies is not incompatible with their clustering into OTUs, and this choice
100 depends on the purpose of the study (i.e. providing a census of the extent and distribution of genetic
101 polymorphism for a given gene, or a census of biodiversity to be used and manipulated in ecological
102 or evolutionary studies). In fact, obtaining a biodiversity inventory of metazoan communities
103 without clustering is likely to deliver a dataset ~~hard~~ to manipulate and interpret in a community
104 ecology framework. In such datasets each haplotype of the target gene in a given species will
105 represent an ASV, yet very distinct levels of intraspecific polymorphism can exist in the same gene
106 region due to both evolutionary and biological specificity (Bucklin, Steinke, & Blanco-Bercial,
107 2011; Phillips, Gillis, & Hanner, 2019). ~~For COI for example, this~~ has been reported among species
108 sampled in the same habitats (Plouviez et al., 2009). ASV-based inventories will thus be biased in
109 favour of taxa with high levels of intraspecific diversity, even though the latter are not necessarily
110 the most abundant ones (Bazin, Glémin, & Galtier, 2006). Such bias in biodiversity inventories
111 based on ASVs is likely to be magnified in presence-absence metabarcode datasets, commonly
112 used for metazoan communities (Ji et al., 2013).

113 Clustering sequences while avoiding arbitrary clustering thresholds is possible with tools
114 such as swarm v2, a single-linkage clustering algorithm (Mahe, Rognes, Quince, De Vargas, &
115 Dunthorn, 2015). Based on network theory, this algorithm aggregates sequences iteratively and
116 locally around seed sequences and determines coherent groups of sequences independent of
117 amplicon input order, allowing highly scalable, fine-scale clustering.

118 Here we evaluate the performance of DADA2 and LULU, using them alone and in
119 combination with swarm v2, to test the possibilities offered by these new tools on metazoan
120 communities ~~revealed~~ using both a mitochondrial COI ~~marker~~ (Leray et al., 2013) and ~~the~~ 18S-

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

V1V2 (Sinniger et al., 2016) ~~small subunit~~ ribosomal RNA (~~SSU~~ rRNA) ~~barcode marker~~. For each of the markers, we evaluated the effect of read correction (using DADA2), clustering (using Swarm v2), and LULU curation to select the pipeline delivering the most accurate resolution in two deep-sea mock communities. We then test the different 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-V4V5 rRNA barcode on these natural samples (Parada, Needham, & Fuhrman, 2016).

Our objectives were to (1) ~~select~~ the most appropriate tools ~~allowing avoiding~~ inflating biodiversity estimates while retaining rare biodiversity and (2) discuss the use of ASV and OTU-centred datasets depending on taxonomic compartment of interest and on study objectives.

1 MATERIALS AND METHODS

1.1 Preparation of samples

Mock communities

Genomic-DNA mass-balanced metazoan mock communities 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). 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%.

Environmental DNA

Sediment cores were collected from thirteen deep-sea sites ranging from the Arctic to the Mediterranean during various cruises (Table S2). Sampling was carried out with a multicorer

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

145 (MUC) or with a remotely operated vehicle (ROV). Three ~~tube~~ cores were taken at each sampling
146 station (GPS coordinates in Table S2). The sediment cores were sliced into depth layers, which
147 were transferred into zip-lock bags, homogenised, and frozen at −80°C on board before being
148 shipped on dry ice to the laboratory. The first layer (0-1 cm) was used ~~for~~ the present ~~analysis~~.
149 DNA extractions were performed using approximately 10 g of sediment with the PowerMax Soil
150 DNA Isolation Kit (Qiagen, Hilden, Germany). To increase the DNA yield, the elution buffer was
151 left on the spin filter membrane for 10 min at room temperature before centrifugation. The ~5 mL
152 extract was then split into three parts, one of which was kept in screw-cap tubes for archiving
153 purposes and stored at -80°C. Negative extraction controls were included in each extraction run.
154

155 **1.2 Amplicon library construction and high-throughput sequencing**

156 Two primer pairs were used to amplify the mitochondrial Cytochrome c Oxidase subunit I
157 (COI) and the 18S-V1V2 small-subunit ribosomal RNA (SSU rRNA) barcode genes specifically
158 targeting metazoans, and one pair of primer was used to amplify the prokaryote 16S-V4V5 region
159 (Table S 3). PCR amplifications, library preparation, and sequencing were carried out at Génoscope
160 (Evry, France) as part of the eDNAbyss project.
161

162 *Eukaryotic 18S-V1V2 rRNA gene amplicon generation*

163 Amplifications were performed with the *Phusion* High Fidelity PCR Master Mix with GC
164 buffer (ThermoFisher Scientific, Waltham, MA, USA) and the SSUF04 and SSUR22*mod* primers
165 (Sinniger et al. 2016, Table S 3). The PCR reactions (25 µL final volume) contained 2.5 ng or less
166 of DNA template with 0.4 µM concentration of each primer, 3% of DMSO, and 1X *Phusion* Master
167 Mix. PCR amplifications (98 °C for 30 s; 25 cycles of 10 s at 98 °C, 30 s at 45 °C, 30 s at 72 °C;

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

168 and 72 °C for 10 min) of all samples were carried out in triplicate in order to smooth the intra-
169 sample variance while obtaining sufficient amounts of amplicons for Illumina sequencing.
170

171 *Eukaryotic COI gene amplicon generation*

172 Metazoan COI barcodes were generated using the mlCOIintF and jgHCO2198 primers
173 (Leray et al. 2013, Table S 3). Triplicate PCR reactions (20 µl final volume) contained 2.5 ng or
174 less of total DNA template with 0.5 µM final concentration of each primer, 3% of DMSO, 0.175
175 mM final concentration of dNTPs, and 1X Advantage 2 Polymerase Mix (Takara Bio, Kusatsu,
176 Japan). Cycling conditions included a 10 min denaturation step followed by 16 cycles of 95 °C for
177 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
178 46°C, 68 °C for 60 s and a final extension of 68 °C for 7 min.
179

180 *Prokaryotic 16S rRNA gene amplicon generation*

181 Prokaryotic barcodes were generated using 515F-Y and 926R 16S-V4V5 primers (Parada
182 et al., 2016). Triplicate PCR mixtures were prepared as described above for 18S-V1V2, but cycling
183 conditions included a 30 s denaturation step followed by 25 cycles of 98 °C for 10 s, 53 °C for 30 s,
184 72 °C for 30 s, and a final extension of 72 °C for 10 min.

185 In all cases, amplicon triplicates were then pooled and PCR products purified using 1X
186 AMPure XP beads (Beckman Coulter, Brea, CA, USA) clean up. Aliquots of purified amplicons
187 were run on an Agilent Bioanalyzer using the DNA High Sensitivity LabChip kit (Agilent
188 Technologies, Santa Clara, CA, USA) to check their lengths and quantified with a Qubit
189 fluorimeter (Invitrogen, Carlsbad, CA, USA).
190

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

191 *Amplicon library preparation*

192 One hundred ng of amplicons were directly end-repaired, A-tailed and ligated to Illumina
193 adapters on a Biomek FX Laboratory Automation Workstation (Beckman Coulter, Brea, CA,
194 USA). Library amplification was performed using a Kapa Hifi HotStart NGS library Amplification
195 kit (Kapa Biosystems, Wilmington, MA, USA) with the same cycling conditions applied for all
196 metagenomic libraries and purified using 1X AMPure XP beads.

197

198 *Sequencing library quality control*

199 Libraries were quantified by Quant-iT dsDNA HS assay kits using a Fluoroskan Ascent
200 microplate fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and then by qPCR with
201 the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems, Wilmington, MA,
202 USA) on an MxPro instrument (Agilent Technologies, Santa Clara, CA, USA). Library profiles
203 were assessed using a high-throughput microfluidic capillary electrophoresis system (LabChip GX,
204 Perkin Elmer, Waltham, MA, USA).

205 *Sequencing procedures*

206 Library concentrations were normalized to 10 nM by addition of 10 mM Tris-Cl (pH 8.5)
207 and applied to cluster generation according to the Illumina Cbot User Guide (Part # 15006165).
208 Amplicon libraries are characterized by low diversity sequences at the beginning of the reads due
209 to the presence of the primer sequence. Low-diversity libraries can interfere in correct cluster
210 identification, resulting in a drastic loss of data output. Therefore, loading concentrations of
211 libraries were decreased (8–9 pM instead of 12–14 pM for standard libraries) and PhiX DNA spike-
212 in was increased (20% instead of 1%) in order to minimize the impacts on the run quality.
213 Libraries were sequenced on HiSeq2500 (System User Guide Part # 15035786) instruments
214 (Illumina, San Diego, CA, USA) in a 250 bp paired-end mode.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

215 **1.3 Bioinformatic analyses**

216 All bioinformatic analyses were performed using a Unix shell script on a home-based
217 cluster (DATARMOR, Ifremer), available on Gitlab (https://gitlab.ifremer.fr/abyss-project/). The
218 mock communities were analysed alongside the natural samples, and used to validate the
219 metabarcoding pipeline in terms of detection of correct species and presence of false-positives. The
220 details of the pipeline, along with specific parameters used for both metabarcoding markers, are
221 listed in Table S 4.

222

223 *Reads preprocessing*

224 Our multiplexing strategy relies on ligation of adapters to amplicon pools, meaning that
225 contrary to libraries produced by double PCR, the reads in each paired sequencing run can be
226 forward or reverse. DADA2 correction is based on error distribution differing between R1 and R2
227 reads. We thus developed a custom script (*abyss-preprocessing* in abyss-pipeline) allowing
228 separating forward and reverse reads in each paired run and reformatting the outputs to be
229 compatible with DADA2. Briefly, the script uses cutadapt v1.18 to separate forward and reverse
230 reads in each paired sequence file, producing two pairs of sequence files per sample named
231 R1F/R2R and R2F/R1R, while removing primers based on a maximum error rate (-e 0.17 for 18S-
232 V1 and 0.27 for COI , -O length of primer -1).). Each identified forward and reverse read is then
233 renamed which the correct extension (/1 and /2 respectively), which is a requirement for DADA2
234 to recognize the pairs of reads. Each pair of renamed sequence files is then re-paired with BBMAP
235 Repair v38.22 in order to remove singleton reads (non-paired reads). Optionally, sequence file
236 names can also be renamed if necessary using a CSV correspondence file.

237

238

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

239 *Read correction, amplicon cluster generation and taxonomic assignment*

240 Pairs of Illumina reads were corrected with DADA2 v.1.10 (Callahan et al., 2016) following
241 the online tutorial for paired-end data (https://benjjneb.github.io/dada2/tutorial.html). Reads were
242 filtered and trimmed with the *filterAndTrim* function and all reads containing ambiguous bases
243 removed (truncLen at 220 for 18S and 16S, 200 for COI, maxEE at 2, truncQ at 11, maxN at 0).
244 The error model was calculated for forward and reverse reads (R1F/R2R pairs and then
245 R2F/R1R pairs) with *learnErrors* based on 100 million randomly chosen bases, and reads were
246 dereplicated using *derepFastq*. After read correction with the *dada* function, forward and reverse
247 reads were merged with a minimum overlap of 12 nucleotides, allowing no mismatches. The
248 amplicons were then filtered by size. The size range was set to 330-390 bp for the 18S SSU rRNA
249 marker gene, 300-326 bp for the COI marker gene, and 350-390 bp for the 16S rRNA marker gene.
250 Chimeras were removed with *removeBimeraDenovo* and ASVs were taxonomically
251 assigned via the RDP naïve Bayesian classifier method, the default assignment method
252 implemented in DADA2. A second taxonomic assignment method was optionally implemented in
253 the pipeline, allowing assigning ASVs using BLAST+ (v2.6.0) based on minimum similarity and
254 minimum coverage (-perc_identity 70 and –qcov_hsp 80). The Silva132 reference database was
255 used for the 16S and 18S SSU rRNA marker genes (Quast et al., 2012), and MIDORI-UNIQUE
256 (Machida, Leray, Ho, & Knowlton, 2017) was used for COI. The databases were downloaded from
257 the DADA2 website (https://benjjneb.github.io/dada2/training.html) and from the FROGS website
258 (http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/). We individually barcoded the
259 species present in the mock communities and added their barcode sequences to all the databases.
260 Finally, to evaluate the effect on clustered data when OTUs are to be produced, ASV tables
261 produced by DADA2 were clustered with swarm v2 (Mahe et al., 2015) at *d=4* for 18S, *d=6* for

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

262 COI, and *d=1* for 16S in FROGS (http://frogs.toulouse.inra.fr/) (Escudié et al., 2018). Resulting
263 OTUs were taxonomically assigned via BLAST+ using the databases stated above.
264 Molecular clusters were refined in R v.3.5.1 (R Core Team, 2018). A blank correction was
265 made using the *decontam* package v.1.2.1 (Davis, Proctor, Holmes, Relman, & Callahan, 2018),
266 removing all clusters that were more abundant in negative control samples than in other samples.
267 ASV/OTU tables were refined taxonomically based on their RDP or BLAST taxonomy. For both
268 assignment methods, unassigned clusters were removed. Non-target 18S and COI clusters
269 (bacterial, non-metazoan) as well as all clusters with a terrestrial assignment (taxonomic groups
270 known to be terrestrial-only, such as Insecta, Arachnida, Diplopoda, Amphibia, terrestrial
271 mammals, Stylommatophora, Aves, Onychophora, Succineidae, Cyclophoridae, Diplommatinidae,
272 Megalomastomatidae, Pupinidae, Veronicellidae) were removed. Samples were checked to ensure
273 that a minimum of 10,000 metazoan reads were left after refining. Finally, an abundance
274 renormalization was performed to remove spurious positive results due to random tag switching
275 (Wangensteen & Turon, 2016).
276 To test LULU curation (Frøslev et al., 2017), refined 18S and COI ASVs/OTUs were
277 curated with LULU v.0.1 following the online tutorial (https://github.com/tobiasgf/lulu). The
278 LULU algorithm detects erroneous clusters by comparing their sequence similarities and co-
279 occurrence rate with more abundant ("parent") clusters. LULU was tested with a minimum relative
280 co-occurrence of 0.90 and a minimum similarity (*minimum match*) threshold of 84% and 90%.
281 The vast majority of prokaryotes usually show low levels (< 1% divergence) of intra
282 genomic variability for the 16S SSU rRNA gene (Acinas, Marcelino, Klepac-Ceraj, & Polz, 2004;
283 Pei et al., 2010). Although we acknowledge that for a limited amount of cases, curation with LULU
284 may still be useful to obtain a more rigorous census of biodiversity, this was not tested on the
285 prokaryote communities used in this study. Indeed, parallelization not being currently available for

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

286 LULU curation, the richness of those communities implied an unrealistic amount of calculation
287 time, even on a powerful cluster (several weeks).
288
289 **1.4 Statistical analyses**
290 Sequence tables were analysed using R with the packages phyloseq v1.22.3 (McMurdie &
291 Holmes, 2013) following guidelines on online tutorials
292 (http://joey711.github.io/phyloseq/tutorials-index.html), and vegan v2.5.2 (Oksanen et al., 2018).
293 Each biodiversity inventory and its LULU curated version were merged into a single phyloseq
294 object. The datasets were normalized by rarefaction to their common minimum sequencing depth,
295 before analysis of the mock communities and the natural samples.
296 To evaluate the functionality of the pipeline with the mock communities, taxonomically
297 assigned metazoan clusters were considered as derived from one of the ten species used for the
298 mock communities when the assignment delivered the corresponding species, genus, family, or
299 class. Clusters not fitting the expected taxa were labelled as 'Others'. These non-target clusters
300 may be spurious or reflect contamination by external DNA or associated microfauna, such as
301 commensals or parasites, which might have been present in the extracted tissue.
302 Alpha diversity detected using each pipeline in the natural samples was evaluated with the
303 number of observed target-taxa in the rarefied datasets via analyses of deviance (ANODEV) on
304 generalized linear models based on quasipoisson distribution models. Homogeneity of multivariate
305 dispersions were verified with the *betapart* package v.1.5.1 (Baselga & Orme, 2012). The effect of
306 LULU curation, site and sediment core (nested within site) on community composition was tested
307 by means of PERMANOVA on the rarefied incidence datasets. PERMANOVAs were calculated
308 using the function *adonis* (vegan), with Jaccard dissimilarities, and 9999 permutations, permuting
309 within sites for evaluating the Pipeline and Core effects. Finally, taxonomic compositions in terms

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

310 of cluster abundance were compared between pipelines and with results of a morphological
311 inventory obtained from a first-level sorting in two sites.
312

313 # 2 RESULTS

314 ## 2.1 High throughput DNA sequencing

315 A number of 45,828,979 18S reads, 34,639,914 COI reads and 16,406,877 16S reads were
316 obtained from six Illumina HiSeq runs of pooled amplicon libraries built from 42 sediment
317 samples, 2 mock communities (for 18S and COI), 6 extraction blanks, and 4-10 PCR negative
318 controls (Table 1). Two sediment samples failed amplification for the COI marker gene
319 (PCT_FA_CT2_0_1 and CHR_CT1_0_1). For metazoans, less reads were retained after
320 bioinformatic processing in negative controls (36% kept for 18S, 47% for COI) than in true or
321 mock samples (~60% kept for 18S, 70-80% for COI), while the opposite was observed for 16S
322 (74% of reads retained in control samples against 53% in true samples). In total, 25,773,684 18S
323 reads, 24,244,902 COI reads, and 9,446,242 16S reads remained after processing with DADA2.
324 Negative control samples (extraction and PCR blanks) contained 2,186,230 (~8%) 18S reads,
325 1,015,700 (~4%) COI reads, and 2,618,729 (28%) 16S reads. These reads were mostly originating
326 from the extraction controls (59% for 18S, 65% for COI, and 72% for 16S). The corresponding
327 clusters were removed from real samples if the number of reads in true samples was lower than in
328 the negative controls.

329 After data refining and abundance renormalization, rarefaction curves showed a plateau
330 was achieved for all samples in both clustered and non-clustered datasets, suggesting an overall
331 sequencing depth adequate to capture the diversity present (Fig. S1).
332

333

Table 1. Number of reads, ASVs, and OTUs obtained in samples after each pipeline step. Data refining was performed in R, based on BLAST assignments. Forward slashes separate ASV/OTU datasets (Dada2 without swarm clustering / Dada2 with swarm clustering).

Sample type	Number of samples	Raw reads	Quality-filtered reads	Merged reads	Reads before chimera removal	Non chimeric reads	% reads retained	Number of ASVs/OTUs before refining	Number of samples after refining	Number of target reads after refining	Number of target reads after renormalisation	Final number of target ASVs/OTUs	Number of target OTUs after LULU 84%	Number of target OTUs after LULU 90%
LOCUS														
18S-V1														
Control Sample	14	6 141 567	2 508 908	2 441 821	2 200 132	2 186 230	35,6	57,661 / 31,509	0	10,234,660 / 10,686,911	10,160,603 / 10,541,499	11,304 / 5,877	2,132 / 1,535	3, 639 / 2,889
Mock Sample	2	2 096 631	1 607 219	1 436 773	1 430 823	1 289 608	61,5		2					
True Sample	42	37 590 781	26 828 194	24 826 430	22 636 689	22 297 846	59,3		42					
COI														
Control Sample	16	2 146 476	1 053 997	1 024 547	1 015 821	1 015 700	47,3	78,785 / 52,216	0	7,601,973 / 5,179,905	7,552,406 / 5,129,293	21,663 / 8,249	11,987 / 4,849	17,265 / 7,251
Mock Sample	2	1 482 785	1 261 045	1 252 908	1 251 994	1 224 795	82,6		2					
True Sample	40	31 010 653	26 011 238	25 287 002	22 197 457	22 004 407	71,0		40					
16S - V4V5														
Control Sample	10	3,531,226	2,889,163	2,634,536	2,619,479	2,618,729	74.2	56,577 / 41,746	0	6,809,966 / 6,801,953	6,719,153 / 6,680,238	55,129 / 40,459	-	-
True Sample	42	12,875,651	9,307,729	7,122,154	7,114,195	6,827,513	53		42					

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

334 The 18S ASV dataset comprised 10,160,603 marine metazoan reads, with an average of
335 230,923 per sample (range of 42,119-721,972). When clustered with swarm v2, the final 18S
336 dataset comprised 10,541,499 target reads, with an average of 239,579 per sample (range 45,259-
337 721,753). The final COI ASV dataset comprised 7,552,406 marine metazoan reads, with an average
338 of 179,819 per sample, (range of 54,585-438,324). When clustered with swarm v2, the final COI
339 dataset comprised 5,129,293 target reads, with an average of 122,126 per sample (range of 31,228-
340 349,805). The 16S ASV dataset comprised 6,719,153 prokaryotic reads, with an average of
341 159,979 per sample (range of 71,834 – 251,054). When clustered with swarm v2, the final 16S
342 dataset comprised 6,680,238 prokaryotic reads, with an average of 159,253 per sample (range
343 71,601 - 250,032).

344 From the total 57,661 ASVs detected for 18S, 47,084 (82%) were assigned by BLAST to
345 phylum level or lower. The assigned ASVs accounted for 97% of total 18S reads. BLAST detected
346 11,304 marine metazoan ASVs (Table 1). Samples contained 389 target ASVs on average, with a
347 range of 88-881 per sample. LULU curation of 18S ASVs at 84% *minimum match* resulted in 2,132
348 clusters (134 per sample on average, range of 11-273), while 3,639 clusters remained after LULU
349 curation at 90% *minimum match* (186 per sample on average, range of 14-402) (Table 1). From the
350 total 31,509 18S OTUs obtained after clustering with swarm v2 (Mahe et al., 2015) at $d=4$ (~1%
351 divergence), 22,427 (71%) were assigned to phylum level or lower The assigned OTUs accounted
352 for 93% of 18S reads. This resulted in 5,877 marine metazoan OTUs after data refining (286
353 metazoan clusters per sample on average, range of 29-698). The number of metazoan OTUs was
354 reduced to 1,535 and 2,889 after LULU curation at 84% and 90% *minimum match* respectively
355 (136 and 196 metazoan clusters per sample on average, range of 10-268 and 12-404 respectively).

356 The number of raw ASVs yielded by COI was higher: 78,785 from which 46,301 (59%)
357 were assigned to phylum level or lower. The assigned ASVs accounted for 65% of total COI reads.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

358 After data refining, BLAST identified 21,663 marine metazoan ASVs in the COI dataset (Table 1).
359 Samples contained 914 ASVs on average, with a range of 56-1,955 per sample. LULU curation of
360 COI ASVs at 84% *minimum match* resulted in 11,987 clusters (599 per sample on average, range
361 of 22-1,210), while 17,265 clusters remained after LULU curation at 90% *minimum match* (787
362 per sample on average, range of 23-1,697). From the 52,216 COI OTUs obtained after clustering
363 ASVs with swarm v2 at *d=6* (~2% divergence), 21,924 (42%) were assigned to phylum level or
364 lower. The assigned OTUs represented 52% of COI reads. After data refining, 8,249 marine
365 metazoan COI OTUs remained in the dataset (470 per sample on average, range of 28-1,069). This
366 number was reduced to 4,849 and 7,251 after LULU curation at 84% and 90% *minimum match*
367 respectively (333 and 434 clusters per sample on average, range of 17-671 and 17-990
368 respectively).

369 From the total 56,577 ASVs detected for 16S, 55,804 (98.6%) were assigned by BLAST at
370 phylum level or lower. The assigned ASVs accounted for 99.9% of total 16S reads, resulting in
371 55,129 final ASVs (Table 1). From the total 41,746 16S OTUs obtained after clustering with swarm
372 v2 (Mahe et al., 2015) at *d=1*, 40,768 (97.7%) were assigned to phylum level or lower, resulting
373 in 40,459 final OTUs.

374 Refining the ASV datasets based on RDP taxonomy resulted in decreased metazoan
375 detection levels, but this was not the case for prokaryotes (Table S 5). For 18S, only 45% of ASVs
376 could be assigned to phylum-level or lower, resulting in 8,365 marine metazoan ASVs. For COI,
377 although RDP assigned 76% of ASVS, only 2,526 target ASVs could be retrieved. We therefore
378 reduced our COI database to only marine sequences. This resulted in 11% of assigned ASVs, but
379 increased the number of target clusters to 8,466 (Table S 6).

380

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

381 **2.2 Performance on mock samples**

382 Assigning ASVs with BLAST allowed recovering 7 out of 10 mock species in the 18S
383 dataset and all species in the COI dataset (Table 2), even with minimum relative DNA abundance
384 levels as low as 0.7% (Mock 5).

385 When ASVs were clustered with swarm v2, this generally led to a slight loss of taxonomic
386 resolution (*Chorocaris* sp. was not detected in Mock 3 for 18S and the two bivalves *P. kilmeri* and
387 *C. regab* were taxonomically misidentified for COI). Taxonomically unresolved species were
388 correctly assigned up to their common family or class level. Dominant species generally produced
389 more reads in both the clustered and non-clustered datasets (Table S 7).

390 Clustering sequences with swarm v2 reduced the number of clusters produced per species,
391 but some species still produced multiple (up to 10) OTUs (*A. arbuscula*, *Munidopsis* sp., and *E.*
392 *norvegica* for 18S; *A. muricola, D. dianthus, Chorocaris* sp., and *Paralepetopsis* sp. for COI).
393 Curating with LULU allowed reducing the number of clusters produced per species to nearly one,
394 with and without clustering, and this for both loci. Moreover, LULU curation decreased the number
395 of spurious clusters ("Others"), but this effect was more marked for 18S and at 84% *minimum*
396 *match* (Table 2). However, curating with LULU the 18S data (ASVs or OTUs) led to the loss of
397 one shrimp species (*Chorocaris* sp) when the *minimum match* parameter was at 90% and an
398 additional species (the limpet *Paralepetopsis* sp.) when this parameter was at 84%. LULU
399 consistently merged the shrimp species *Chorocaris* sp with another shrimp species as the latter
400 were always co-occurring in our mock samples.

401

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 2. Number of ASVs/OTUs detected per species in the mock communities using different bioinformatic pipelines. White cells indicate an exact match with the number of OTUs expected, grey cells indicate a number of OTUs differing by ±3 from the number expected, and dark grey cells indicate a number of OTUs >3 from the one expected.

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

402

403

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

404 Assigning ASVs with the RDP Bayesian Classifier allowed recovering 4 out of 10 mock
405 species in the 18S dataset (Fig S-2) and no species in the COI dataset using the full MIDORI
406 database. The six incorrectly resolved species in the 18S dataset could only be resolved
407 taxonomically up to their common class level (venerid bivalves and malacostracan crustaceans).
408 For the COI dataset, using the full MIDORI database resulted in RDP assignments that never
409 matched the expected taxon and were mostly assigned to arthropods (data not shown). When the
410 database was reduced to marine-only taxa, all 10 species were detected (Fig S 2), although the
411 dataset contained a considerable amount of spurious assignments (29 clusters assigned up to
412 Arthropoda and Chordata). The latter were however always associated to a phylum bootstrap level
413 < 98. As the taxonomic resolution using RDP was poorer in the mock communities using 18S, the
414 remaining work was performed using BLAST assignments.
415

416 **2.3 Alpha-diversity patterns between pipelines**

417 *Eukaryotes*

418 The number of metazoan clusters detected in the deep-sea sediment samples varied
419 significantly between bioinformatic pipelines chosen (ANODEV: 18S, $F(5,175)=599.91$, $p<0.001$
420 and COI, $F(5,195)=1,320.32$, $p<0.001$, 16S, $F(51,41)=2008.76$, $p<0.001$, see Table S 8).
421 Expectedly, clustering and LULU curation significantly reduced the number of detected clusters
422 per sample for all loci. The reduction due to clustering was much more pronounced for metazoans,
423 particularly for COI, than for 16S data (Fig. 1). DADA2 detected on average 389 (SE=28) and 863
424 (SE=61) metazoan 18S and COI ASVs per sample respectively, while clustering ASVs (at $d=4$ for
425 18S, $d=6$ for COI, and $d=1$ for 16S) reduced the number of metazoan OTUs detected to 289
426 (SE=21) for 18S and 467 (SE=34) for COI. For prokaryotes, the number of ASVs was on average
427 3,567 (SE=480) per sample, clustering decreased this mean to 3,138 (SE=413) OTUs per sample.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

428

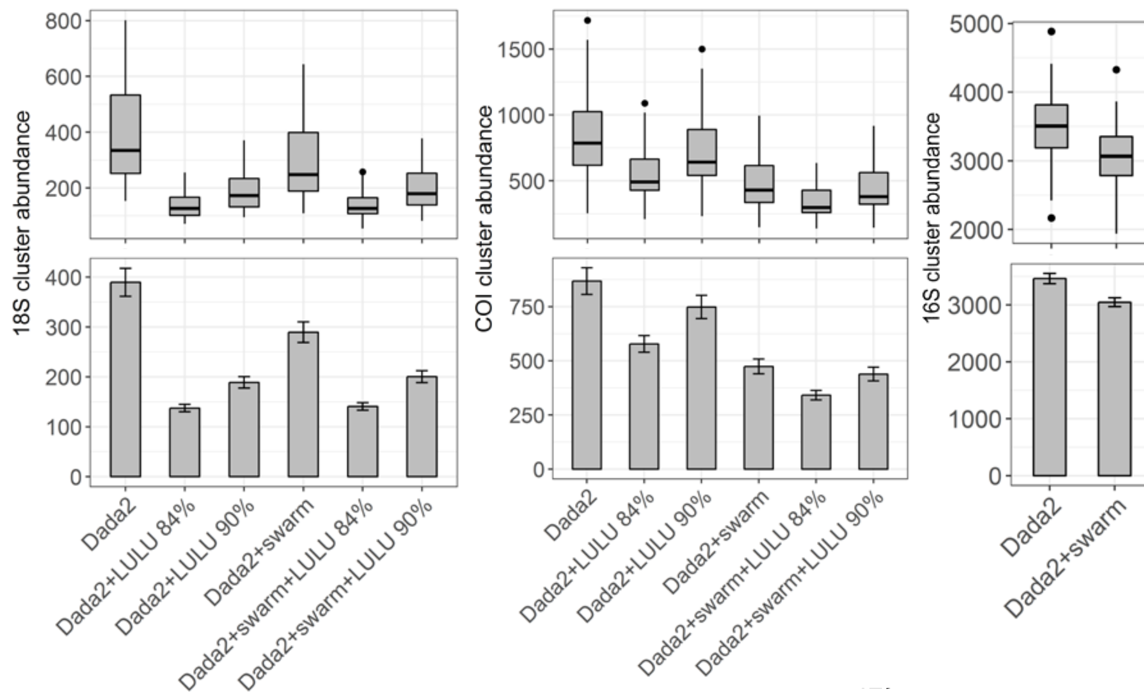

Figure 1. Number of clusters detected in sediment of 14 deep-sea sites with the Dada2 metabarcoding pipeline with or without LULU curation at 84% and 90% *minimum match* and swarm v2 clustering, using the 18S (left) and COI (centre) and 16S (right) marker genes. Cluster abundance was obtained after rarefaction to minimal sequencing depth. Boxplots represent medians with first and third quartiles. Barplots show means and standard errors.

430

431 LULU significantly decreased the number of metazoan clusters detected in both the ASV
432 and OTU datasets. The effect was stronger at a lower *minimum match* parameter. It was also more
433 pronounced in the ASV datasets and for the 18S locus (Fig. 1). At 90% minimum match, LULU
434 decreased by 51% the number of 18S and by 14% the number of COI ASVs, while this decrease
435 was only of 31% for 18S OTUs and 7% for COI OTUs. When the *minimum match* parameter was
436 at 84%, LULU decreased the number of detected metazoan clusters by 65% for 18S ASVs and
437 33% for COI ASVs, while in the clustered dataset this decrease was of 51% and 28% for 18S and
438 COI OTUs respectively. LULU curation of ASVs or OTUs produced comparable number of
439 clusters in the 18S dataset. At 84% *minimum match*, LULU curation produced on average 137 ± 7
440 and 140 ± 8 clusters per sample after application on ASVs and OTUs respectively. At 90%, these

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

441 numbers were at 189 ± 11 and 200 ± 12 (Fig. 1). This was not the case for COI, where LULU
442 curation of ASVs resulted in significantly more clusters (574 ± 38 at 84% and 742 ± 53 at 90%)
443 than LULU curation of OTUs (334 ± 21 and 433 ± 31).
444 The number of clusters detected also varied significantly among sites (ANODEV: 18S,
445 $F(11,175)=283.57$, $p<0.001$; COI, $F(13,195)=761.19$, $p<0.001$; 16S, $F(13,41)=507.37$, $p<0.01$),
446 and cores nested within sites (ANODEV: 18S, $F(24,175)=32.21$, $p<0.001$; COI, $F(26,195)=72.91$,
447 $p<0.001$; 16S, $F(28,41)=241.73$, $p<0.01$). However, while the mean number of clusters detected
448 per sample spanned a wide range in all loci (100-800 for 18S, 150-1,500 for COI datasets, and
449 2,000-5,000 for 16S), the pipeline effect was consistent across sites (Fig. S 3).
450

451 **2.4 Taxonomic assignments and patterns of beta-diversity between pipelines**

452 Sequence identity varied strongly depending on phyla and marker gene (Fig. 2). For 18S,
453 most clusters had hit identities ≥ 90%. Poorly assigned clusters (hit identity < 90%) represented
454 less than 20% of the dataset and were mostly assigned to Nematoda, Cnidaria, Tardigrada, Porifera,
455 and Xenacoelomorpha. For COI, nearly all clusters had similarities to sequences in databases lower
456 than 90%. Overall, arthropods and echinoderms were detected at similar levels by both markers.
457 The 18S barcode marker performed better in the detection of nematodes, annelids, platyhelminths,
458 and xenacoelomorphs while COI mostly detected cnidarians, molluscs, and poriferans (Fig. 2),
459 highlighting the complementarity of these two loci. Sequence identity was much higher for
460 prokaryotes, with most clusters assigned above 90%.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

461

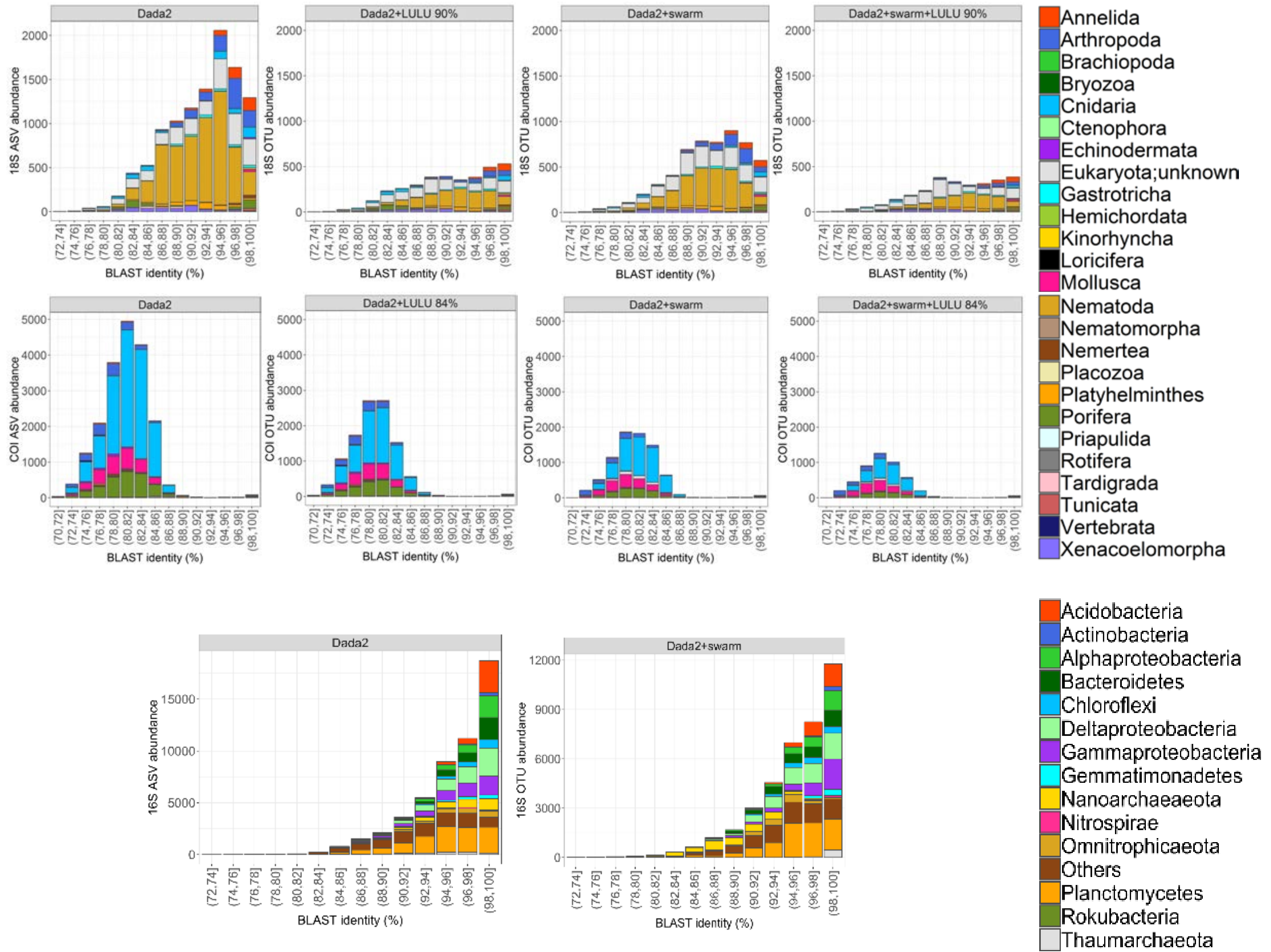

Figure 2. Taxonomic resolution in in metabarcoding datasets of 14 deep-sea sediment sites with four bioinformatic pipelines. Metazoan taxonomic assignment quality based on the 18S (top), COI (centre) and 16S (bottom) marker genes. BLAST hit identity of all metazoan clusters detected is given for four bioinformatic pipelines: DADA2, DADA2 curated with LULU at 84/90% *minimum match*, DADA2 clustered with swarm v2, and DADA2 clustered with swarm v2 and curated with LULU at 84/90% *minimum match*. BLAST hit identity for prokaryotes is given for two pipelines: DADA2 and DADA2 with swarm v2.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

462 For metazoan loci, while clustering significantly decreased the number of OTUS detected,
463 it increased the amount of clusters not assigned up to the phylum level in both loci (~10-20%
464 increase, Fig. 2). In the 18S dataset, clustering led to the decrease in abundance of dominant taxa
465 such as nematodes and non-dominant taxa such as cnidarians and poriferans (Fig. 2, Fig. 3).
466 Similarly, for COI, clustering led to a decreased abundance of dominant taxa such as poriferans
467 and cnidarians, while the number of clusters assigned to arthropods and molluscs increased (Fig.
468 2, Fig. 3). Changes were less marked for 16S data (Fig. 2), yet the number of some taxa clearly
469 increased (i.e. Thaumarchaeota and Gammaproteonbacteria) whereas others decreased (i.e.
470 Omnitrophicaeota).
471 For COI and 18S datasets, PERMANOVAs were performed to evaluate the effect of LULU
472 curation at two *minimum match* thresholds. Multivariate analyses on clustered and non-clustered
473 datasets showed significant differences in community structure between bioinformatic pipeline (i.e.
474 with or without LULU), sites, and cores nested within sites (Table 3). LULU had a significant
475 effect on taxonomic structure resolved, even though the percentage variation it explained was only
476 around 1.3% for 18S and 0.5% for COI (R^2 values in Table 3), compared to 40-50% variation
477 explained by sites, reflecting the predominant effect of biological signatures over bioinformatic
478 processing in the resolution of community structure. Comparing the taxonomic composition
479 resolved by all pipelines showed that LULU curation of ASVs or OTUS resulted in detected
480 community compositions similar to non-curated datasets, although it increased the relative
481 abundance of non-dominant taxa by decreasing the abundance of dominant phyla such as
482 nematodes in 18S and cnidarians in COI (Fig. 3).
483

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license.

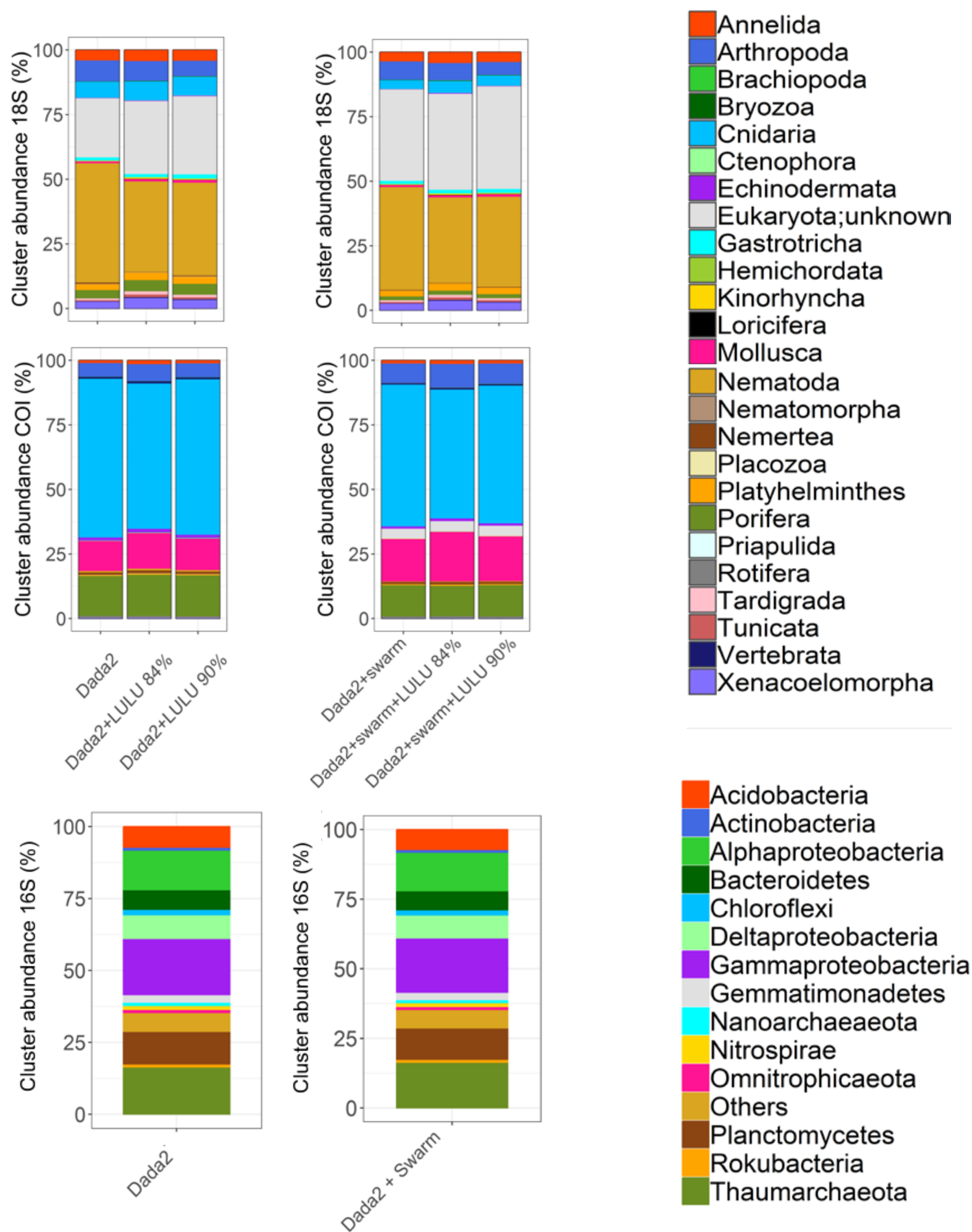

Figure 3. Patterns of relative cluster abundance resolved by different bioinformatic pipelines (ASV-centred on the left, OTU-centred on the right) in 14 deep-sea sites, using the 18S (top), COI (centre), and 16S (bottom) marker genes. LULU curation and clustering increase the abundance of non-dominant taxonomic groups in both metazoan loci, while this is not the case for prokaryotes.

484

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

485 Overall, community differences were visualized using PCoA ordinations of Jaccard
486 distance matrices and showed that the different pipelines resolved the same ecological patterns, in
487 which, consistently with the PERMANOVAs, the site effect was predominant (Fig. S 4).
488

Table 3. Effect of LULU curation on community structure detected in 14 deep-sea sites. Results of the permutational analysis of variance (PERMANOVA) of the rarefied OTU richness in clustered (Dada2+swarm+LULU) and non-clustered (Dada2+LULU) datasets, for the two genes studied. The tests were performed by permuting 9999 times using Jaccard distances. The pipeline and core effects were evaluated by permuting within sites.

	Dada2+swarm+LULU					**Dada2+LULU**					
LOCUS	***df***	**SS**	**Pseudo-F**	**P(>F)**	**R2**		***df***	**SS**	**Pseudo-F**	**P(>F)**	**R2**
18S-V1						**18S-V1**					
Pipeline	2	0.755	5.62	0.001	0.014	Pipeline	2	0.695	2.97	0.0001	0.012
Site	13	24.238	27.79	0.001	0.455	Site	13	23.658	15.57	0.0001	0.410
Site:Core	28	22.734	12.10	0.001	0.427	Site:Core	28	23.74	7.25	0.0001	0.412
Residuals	82	5.505			0.103	Residuals	82	9.584			0.166
Total	125	53.228			1.000	Total	125	57.677			1.000
COI						**COI**					
Pipeline	2	0.262	4.75	0.0001	0.005	Pipeline	2	0.244	2.68	0.0001	0.004
Site	13	29.555	82.47	0.0001	0.557	Site	13	27.525	46.61	0.0001	0.498
Site:Core	26	21.069	29.40	0.0001	0.397	Site:Core	26	24.984	20.31	0.0001	0.434
Residuals	78	2.15			0.041	Residuals	78	3.543			0.064
Total	119	53.036			1.000	Total	119	55.296			1.000

489
490
491 # 3 DISCUSSION
492 ## 3.1 ASVs or OTUs for metazoans?
493 The rise of HTS and the subsequent use of metabarcoding have revolutionized
494 microbiology by unlocking the access to uncultivable microorganisms, which represent by far the
495 great majority of prokaryotes (Klappenbach, Saxman, R., & Schmidt, 2001). The development and
496 improvement of molecular and bioinformatic methods to perform inventories were historically
497 primarily developed for 16S rRNA barcode loci, before being transferred to the eukaryotic

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

498 kingdom based on the use of barcode markers such as 18S rRNA, ITS, or mitochondrial markers
499 such as COI (Bellemain et al., 2010; Valentini et al., 2009). Thus, most bioinformatics pipelines
500 were initially developed accounting for intrinsic properties of prokaryotes and concepts inherent to
501 microbiology (Boyer et al., 2016; Caporaso et al., 2010; Schloss et al., 2009), before being
502 transferred to eukaryotes in general or metazoans in particular. Such transfers are not always
503 straightforward, and require adaptations accounting for differences in both concepts and basic
504 biological features. One example is the question of the relevance of the use of amplicon sequence
505 variants (ASVs), advocated to replace OTUs "… *as the standard unit of marker-gene analysis and*
506 *reporting*" (Callahan et al., 2017): an advice for microbiologists that may not apply when working
507 on metazoans.

508 The results on the mock samples showed that ASV-centred datasets produced using
509 DADA2-alone are likely to be unsuited for metazoan metabarcoding using the 18S and COI
510 barcode markers, as single individuals produced very different numbers of ASVs, therefore not
511 reflecting actual species composition. Clustering ASVs into OTUs using swarm v2 still led to
512 inflated diversity estimates, as despite a unique specimen of each species was used in the mock
513 communities, some still produced up to ten OTUs for both loci (Table 2). This result suggests that
514 even in quality-filtered and clustered datasets, diversity of some taxa will still be overestimated
515 unless high clustering thresholds are used, which may in turn lead to the loss of diversity through
516 the merging of distinct taxa. Intra-individual variation is a recognised problem in metabarcoding,
517 known to generate spurious clusters (Brown et al., 2015), especially in the COI barcode marker.
518 Indeed, this gene region has increased intra-individual variation due to heteroplasmy and the
519 abundance of pseudogenes, such as NUMTs, playing an important part of the supernumerary OTU
520 richness in COI-metabarcoding (Bensasson, Zhang, Hartl, & Hewitt, 2001; Song, Buhay, Whiting,
521 & Crandall, 2008). Together with clustering, LULU curation proved effective in limiting the

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

522 number of multiple clusters produced by single individuals, confirming its efficiency to correct for
523 intra-individual diversity (Table 2).

524 **3.2 Adapting pipelines to marker properties**

525 As seen above for COI, when considering a single marker, the biology of the organisms
526 together with the properties of the gene itself determine its level of intra-individual and intraspecific
527 diversity. Concerted evolution is a common feature of SSU rRNA markers such as 16S (Hashimoto,
528 Stevenson, & Schmidt, 2003; Klappenbach et al., 2001) and 18S (Carranza, Giribet, Ribera,
529 Baguñà, & Riutort, 1996) that limits the amount of intra individual polymorphism. Despite a
530 number of 16S rRNA variants estimated to be 2.5-fold greater than the number of bacterial species
531 (Acinas et al., 2004), the reproductive mode and pace of selection in microbial populations is likely
532 to lead to locally lower level of intraspecific variation than the one expected for 18S and COI in
533 metazoans for example. In addition, in metazoans, a lower level of diversity is expected for the
534 slower evolving 18S (Carranza et al., 1996), than for COI. This may explain the lower ASV
535 (DADA2) to OTU (DADA2+swarm) ratios observed here for 16S (~1.4) compared to 18S (~1.9)
536 and COI (~2.6) data, underlining the different influence –and importance- of clustering on these
537 loci, and the need for a versatile marker by marker choice for clustering parameters.

538 The COI locus allowed the detection of all ten species present in the mock samples,
539 compared to seven in the 18S dataset (Table 2). This locus also provided much more accurate
540 assignments, most of them correct at the genus (and species) level, confirming that COI uncovers
541 more metazoan species and offers a better taxonomic resolution than 18S (Clarke, Beard, Swadling,
542 & Deagle, 2017; Tang et al., 2012). The results also confirm an important variation in the
543 amplification success across taxa (Bhadury et al., 2006; Carugati, Corinaldesi, Dell’Anno, &
544 Danovaro, 2015), supporting the present approach combining nuclear and mitochondrial markers

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

545 to achieve more comprehensive biodiversity inventories (Cowart et al., 2015; Drummond et al.,
546 2015; Zhan, Bailey, Heath, & Macisaac, 2014).
547 While clustering and LULU curation improved COI results in the mock communities
548 (where species always co-occurred), they were associated with a decrease in taxonomic resolution
549 for 18S data, as some closely related species were merged, i.e. the vesicomyid bivalves, the
550 gastropod, and the shrimp species (Table 2). When studying natural habitats, very likely to harbour
551 closely related co-occurring species, both LULU curation and clustering are likely to lead to the
552 loss of true species diversity for low-resolution markers such as 18S. Optimal results in the mock
553 samples, i.e. delivering the best balance between the limitation of spurious clusters and the loss of
554 true diversity, were obtained with LULU curation at 90% for 18S and 84% for COI, highlighting
555 the importance of adjusting bioinformatic correction tools to each barcode marker, a step for which
556 mock communities are most adequate.
557

558 **3.3 Application to real communities**

559 The mock communities we used here did not contain several haplotypes of the same species
560 (intraspecific variation), as is most often the case in environmental samples. This prevents us from
561 generalizing the comparable results of LULU obtained with or without clustering to more complex
562 communities. As distinct haplotypes do not always co-occur in nature, LULU curation of ASVs
563 alone is not suited to correct for haplotype diversity, and clustering ASVs may therefore still be
564 necessary to produce datasets that reflect species rather than gene diversity. As expected, results
565 on natural samples showed distinct answers to this question for 18S and COI. When applying
566 LULU to ASVs (DADA2) *versus* OTUs (DADA2+swarm) on 18S, similar numbers of detected
567 clusters were obtained (e.g. average of 137 ± 7 and 140 ± 8 clusters per sample after application at
568 84% on ASVs and OTUs respectively), again suggesting a limited added effect of clustering for

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

569 this marker once DADA2 and LULU are applied (Fig. 1). This is in line with its slow evolutionary
570 rate (Carranza et al., 1996) leading to a limited number of haplotypes per species compared to COI.
571 In contrast, after LULU curation of the COI ASV dataset, nearly twice the number of clusters were
572 obtained (574 ± 38 at 84% and 742 ± 53 at 90%) compared to the LULU-curated OTU dataset (334
573 ± 21 for 84% and 433 ± 31 for 90%). This confirms the need for clustering on COI and the fact
574 that LULU curation of ASVs is not sufficient to account for intraspecific diversity in natural
575 samples for such a highly polymorphic marker.

576 Finally, prokaryotic alpha diversity was less affected by the clustering of ASVs (Table 1,
577 Fig. 1), illustrating their lower intra-genomic variability (Pei et al., 2010) and the possibly lower
578 diversity within ecotypes. Nevertheless, the differences suggest the occurrence of very closely
579 related sequences of 16S rRNA, possibly belonging to the same ecotype/species. Such entities may
580 still be important to delineate in studies aiming for example at identifying species associations (i.e.
581 symbiotic relationships) across large distances and ecosystems, where drift or selection can lead to
582 slightly different ASVs in space and time, with their function and association remaining stable.
583

584 **3.4 Influence on beta diversity**

585 After focusing on alpha diversity estimates and the accuracy of inventories, the analysis of
586 taxonomic structure showed that the non-clustered, clustered, and LULU-curated datasets resolved
587 similar ecological patterns (Fig. S 4) and community compositions (Fig. 3), although differences
588 in abundance were observed (Fig. 2). This is in accordance with other studies reporting severe
589 impacts of bioinformatic parameters on alpha diversity while comparable patterns of beta diversity
590 were observed, at least down to a minimum level of clustering stringency (Bokulich et al., 2013;
591 Xiong & Zhan, 2018).

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

592 Clustering and LULU curation mainly led to the decrease of the number of clusters assigned
593 to dominant taxa in both loci, i.e. nematodes for 18S, cnidarians and to lesser extent molluscs for
594 COI. This is likely attributable to the low resolutive power of 18S, already acknowledged in general
595 and for nematodes in particular (Derycke, Vanaverbeke, Rigaux, Backeljau, & Moens, 2010).
596 Similarly the lack of resolution of COI for cnidarians has long been known (Hebert, Ratnasingham,
597 & de Waard, 2003). Clustering also introduced more OTUs that could not be assigned at the phylum
598 level with BLAST (Fig. 3), confirming the limitations of assigning taxonomy at the OTU level, as
599 the representative sequence chosen for taxonomic assignment can lead to taxonomic ambiguity.
600

601 **3.5 Assignment comparison**

602 Finally, compared to BLAST assignment, lower taxonomic resolution was observed using
603 the RDP Bayesian Classifier on the mock samples for 18S (Fig. S 2) and for COI when using the
604 full MIDORI database. With this database, only five phyla were detected in the whole dataset:
605 Arthropoda, Chordata, Mollusca, Nemertea, Porifera (data not shown). This is likely due to the size
606 of the RDP training sets available for this study, and to the low coverage of deep-sea species in
607 public databases. Small databases, taxonomically similar to the targeted communities, and with
608 sequences of the same length as the amplified fragment of interest, are known to maximise accurate
609 identification (Macheriotou et al., 2019). This limitation of databases, rather than the method itself,
610 was confirmed by results using a reduced marine-only COI database. The latter (containing the
611 barcodes of the mock species) resulted in accurate RDP assignments in the mock samples when
612 the phylum bootstrap level was ≥ 98 (Fig. S 2), although the majority of clusters remained
613 unassigned in the full dataset (89% compared to 45% with BLAST). The development of custom-
614 built marine RDP training sets, without overrepresentation of terrestrial species, is therefore needed
615 for this Bayesian assignment method to be effective on deep-sea datasets. With reduced trainings

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

sets, removing clusters with phylum bootstrap-level < 98 could be an efficient way to increase taxonomic quality of deep-sea metabarcoding datasets. At present, BLAST seems however the most efficient assignment method for deep-sea metabarcoding data, even though it has to be kept in mind that hit identities tend to be low, especially for COI, making it hard to work at taxonomic levels beyond phylum or class (Fig. 2).

CONCLUSIONS AND PERSPECTIVES

~~In this work based on~~ mock communities and natural samples, we propose a new pipeline ~~using~~ several recent algorithms allowing to improve the quality of biodiversity inventories based on metabarcoding data. Results showed that ASV data should be produced and communicated for reusability and reproducibility following the recommendations of Callahan et al. (2017). This is especially useful in large projects spanning wide geographic zones and time scales, as different ASV datasets can be easily merged *a posteriori,* and clustered if necessary afterwards. Nevertheless, clustering ASVs into OTUs will be required to obtain accurate inventories, at least for metazoan communities. Considering 16S polymorphism observed in prokaryotic species (Acinas et al., 2004) and the possible geographic segregation of their populations, clustering may also be required in prokaryotic datasets, for example in studies screening for species associations (i.e. symbiotic or parasitic relationships, considering that symbionts may be prone to differential fixation through enhanced drift; Shapiro, Leducq, & Mallet, 2016).

Results also demonstrated that LULU curation is a good alternative to arbitrary relative abundance filters in metabarcoding pipelines. They also underline the need to adapt parameters for curation (e.g. LULU curation at 90% for 18S and 84% for COI) and clustering to each gene used and taxonomic compartment targeted, in order to identify an optimal balance between the correction for spurious clusters and the merging of closely related species.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

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

The pipeline is publicly available on Gitlab (https://gitlab.ifremer.fr/abyss-project/), and allows the use of sequence data obtained from libraries produced by double PCR or adaptor ligation methods, as well as having built-in options for using six commonly used metabarcoding primers.

ACKNOWLEDGEMENTS

This work is part of the "*Pourquoi Pas les Abysses?*" project, funded by Ifremer and the project eDNAbyss (AP2016 -228) funded by France Génomique (ANR-10-INBS-09) and Genoscope-CEA. We wish to thank Caroline Dussart for her contribution to the early developments of bioinformatic scripts and Patrick Durand for bioinformatic support, Cathy Liautard-Haag for her help with lab management, Babett Guenther for barcoding. We also wish to express our gratitude to the participants and mission chiefs of the Essnaut (Marie-Anne Cambon Bonavita), MarMine (Eva Ramirez Llodra, project 247626/O30 (NRC-BI), PEACETIME (Cécile Guieu and Christian Tamburini), CanHROV (Marie Claire Fabri) and MEDWAVES (Covadonga Orejas) cruises. The MEDWAVES cruise was organised in the framework of the ATLAS Project, supported by the European Union 2020 Research and Innovation Programme under grant agreement number: 678760.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

661 **REFERENCES**

662 Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and
663 Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons. *Journal of*
664 *Bacteriology*, *186*(9), 2629–2635. https://doi.org/10.1128/JB.186.9.2629-2635.2004
665 Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2017). Scrutinizing key steps for
666 reliable metabarcoding of environmental samples. (A. Mahon, Ed.), *Methods in Ecology and*
667 *Evolution*. https://doi.org/10.1111/2041-210X.12849
668 Baselga, A., & Orme, C. D. L. (2012). betapart : an R package for the study of beta diversity.
669 *Methods in Ecology and Evolution*, *3*(5), 808–812. https://doi.org/10.1111/j.2041-
670 210X.2012.00224.x
671 Bazin, E., Glémin, S., & Galtier, N. (2006). Population size does not influence mitochondrial
672 genetic diversity in animals. *Science*, *312*(5773), 570–572.
673 https://doi.org/10.1126/science.1122033
674 Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., & Kauserud, H. (2010). ITS
675 as an environmental DNA barcode for fungi: An in silico approach reveals potential PCR
676 biases. *BMC Microbiology*, *10*, 189. https://doi.org/10.1186/1471-2180-10-189
677 Bensasson, D., Zhang, D. X., Hartl, D. L., & Hewitt, G. M. (2001, June 1). Mitochondrial
678 pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology and Evolution*.
679 https://doi.org/10.1016/S0169-5347(01)02151-6
680 Bhadury, P., Austen, M. C., Bilton, D. T., Lambshead, P. J. D., Rogers, A. D., & Smerdon, G. R.
681 (2006). Molecular detection of marine nematodes from environmental samples: overcoming
682 eukaryotic interference. *Aquatic Microbial Ecology*, *44*(1), 97–103. https://doi.org/Doi
683 10.3354/Ame044097

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

684 Bik, H. M., Sung, W., De Ley, P., Baldwin, J. G., Sharma, J., Rocha-Olivares, A., & Thomas, W.
685 K. (2012). Metagenetic community analysis of microbial eukaryotes illuminates
686 biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology*, *21*(5),
687 1048–1059. https://doi.org/10.1111/j.1365-294X.2011.05297.x
688 Bista, I., Carvalho, G., Walsh, K., Christmas, M., Hajibabaei, M., Kille, P., … Creer, S. (2015).
689 Monitoring lake ecosystem health using metabarcoding of environmental DNA: temporal
690 persistence and ecological relevance. *Genome*, *58*(5), 197.
691 Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., … Caporaso,
692 J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon
693 sequencing. *Nature Methods*, *10*(1), 57–59. https://doi.org/10.1038/nmeth.2276
694 Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J.-B., … Mouillot,
695 D. (2018). Environmental DNA illuminates the dark diversity of sharks. *Science Advances*,
696 *4*(5), eaap9661. https://doi.org/10.1126/sciadv.aap9661
697 Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). OBITOOLS: a
698 UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*,
699 *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428
700 Brannock, P. M., & Halanych, K. M. (2015). Meiofaunal community analysis by high-throughput
701 sequencing: Comparison of extraction, quality filtering, and clustering methods. *Marine*
702 *Genomics*, *23*, 67–75. https://doi.org/10.1016/j.margen.2015.05.007
703 Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J., & Cristescu, M. E. (2015).
704 Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably
705 describe zooplankton communities? *Ecology and Evolution*, *5*(11), 2234–2251.
706 https://doi.org/10.1002/ece3.1485
707 Bucklin, A., Steinke, D., & Blanco-Bercial, L. (2011). DNA Barcoding of Marine Metazoa.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

708 *Annual Review of Marine Science*, *3*(1), 471–508. https://doi.org/10.1146/annurev-marine-
709 120308-080950
710 Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace
711 operational taxonomic units in marker-gene data analysis. *ISME Journal*, *11*(12), 2639–
712 2643. https://doi.org/10.1038/ismej.2017.119
713 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P.
714 (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature*
715 *Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869
716 Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., …
717 Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data.
718 *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303
719 Carranza, S., Giribet, G., Ribera, C., Baguñà, J., & Riutort, M. (1996). Evidence that two types of
720 18S rDNA coexist in the genome of Dugesia (Schmidtea) mediterranea (Platyhelminthes,
721 Turbellaria, Tricladida). *Molecular Biology and Evolution*, *13*(6), 824–832.
722 https://doi.org/10.1093/oxfordjournals.molbev.a025643
723 Carugati, L., Corinaldesi, C., Dell'Anno, A., & Danovaro, R. (2015). Metagenetic tools for the
724 census of marine meiofaunal biodiversity: An overview. *Marine Genomics*, *24*, 11–20.
725 https://doi.org/10.1016/j.margen.2015.04.010
726 Clare, E. L., Chain, F. J. J., Littlefair, J. E., & Cristescu, M. E. (2016). The effects of parameter
727 choice on defining molecular operational taxonomic units and resulting ecological analyses
728 of metabarcoding data. *Genome*, *59*(11), 981–990. https://doi.org/10.1139/gen-2015-0184
729 Clarke, L. J., Beard, J. M., Swadling, K. M., & Deagle, B. E. (2017). Effect of marker choice and
730 thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and*
731 *Evolution*, *7*(3), 873–883. https://doi.org/10.1002/ece3.2667

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

732 Cohan, F. M. (2001). Bacterial species and speciation. *Systematic Biology*, *50*(4), 513–524.
733 https://doi.org/10.1080/10635150118398
734 Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding
735 of plants and animals. *Molecular Ecology*, *21*(8), 1834–1847.
736 https://doi.org/10.1111/j.1365-294X.2012.05550.x
737 Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., & Arnaud-Haond, S.
738 (2015). Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of
739 Morphological and Molecular Surveys of Seagrass Communities. *PLoS One*, *10*(2),
740 e0117562. https://doi.org/10.1371/journal.pone.0117562
741 Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., … Bik, H. M.
742 (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods*
743 *in Ecology and Evolution*, *7*(9), 1008–1018. https://doi.org/10.1111/2041-210X.12574
744 Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple
745 statistical identification and removal of contaminant sequences in marker-gene and
746 metagenomics data. *Microbiome*, *6*(1), 226. https://doi.org/10.1186/s40168-018-0605-2
747 de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the*
748 *National Academy of Sciences*, *102*(Supplement 1), 6600–6607.
749 https://doi.org/10.1073/pnas.0502030102
750 De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., … Velayoudon, D.
751 (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237).
752 https://doi.org/10.1126/science.1261605
753 Deiner, K., Fronhofer, E. A., Mächler, E., Walser, J.-C., & Altermatt, F. (2016). Environmental
754 DNA reveals that rivers are conveyer belts of biodiversity information. *Nature*
755 *Communications*, *7*(1), 12544. https://doi.org/10.1038/ncomms12544

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

756 Deiner, K., Walser, J.-C. C., Machler, E., Altermatt, F., Mächler, E., & Altermatt, F. (2015).
757 Choice of capture and extraction methods affect detection of freshwater biodiversity from
758 environmental DNA. *Biological Conservation*, *183*, 53–63.
759 https://doi.org/10.1016/j.biocon.2014.11.018
760 Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., & Miaud, C.
761 (2011). Persistence of Environmental DNA in Freshwater Ecosystems. *PLoS One*, *6*(8).
762 https://doi.org/10.1371/journal.pone.0023398
763 Derycke, S., Vanaverbeke, J., Rigaux, A., Backeljau, T., & Moens, T. (2010). Exploring the use
764 of cytochrome oxidase c subunit 1 (COI) for DNA barcoding of free-living marine
765 nematodes. *PLoS ONE*, *5*(10), e13716. https://doi.org/10.1371/journal.pone.0013716
766 Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., … Weaver, L.
767 (2018). Towards robust and repeatable sampling methods in eDNA-based studies.
768 *Molecular Ecology Resources*, *18*(5), 940–952. https://doi.org/10.1111/1755-0998.12907
769 Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C. M., …
770 Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity
771 assessment. *Gigascience*, *4*. https://doi.org/ARTN 4610.1186/s13742-015-0086-1
772 Eren, A. M., Vineis, J. H., Morrison, H. G., & Sogin, M. L. (2013). A Filtering Method to
773 Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLoS ONE*,
774 *8*(6), e66643. https://doi.org/10.1371/journal.pone.0066643
775 Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., … Berger, B.
776 (2018). FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, *34*(8), 1287–
777 1294. https://doi.org/10.1093/bioinformatics/btx791
778 Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y. Y., Jerde, C. L., … Lodge, D. M.
779 (2016). Quantification of mesocosm fish and amphibian species diversity via environmental

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

780 DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 29–41.
781 https://doi.org/10.1111/1755-0998.12433
782 Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., … Taberlet, P.
783 (2015). Replication levels, false presences and the estimation of the presence/absence from
784 eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556.
785 https://doi.org/10.1111/1755-0998.12338
786 Fonseca, V. G. (2018). Pitfalls in relative abundance estimation using edna metabarcoding.
787 *Molecular Ecology Resources*, *18*(5), 923–926. https://doi.org/10.1111/1755-0998.12902
788 Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power, D. M., Neill, S. P., … Creer,
789 S. (2010). Second-generation environmental sequencing unmasks marine metazoan
790 biodiversity. *Nature Communications*, *1*. https://doi.org/9810.1038/ncomms1095
791 Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A.
792 J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable
793 biodiversity estimates. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-
794 01312-x
795 Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., … Swings, J.
796 (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, *3*(9), 733–739.
797 https://doi.org/10.1038/nrmicro1236
798 Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., …
799 Taberlet, P. (2016). Critical considerations for the application of environmental DNA
800 methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307.
801 https://doi.org/10.1111/2041-210X.12595
802 Hashimoto, J. G., Stevenson, B. S., & Schmidt, T. M. (2003). Rates and consequences of
803 recombination between rRNA operons. *Journal of Bacteriology*, *185*(3), 966–972.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

804 https://doi.org/10.1128/JB.185.3.966-972.2003

805 Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: cytochrome
806 c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal*
807 *Society of London. Series B: Biological Sciences*, *270*(suppl_1), S96-9.
808 https://doi.org/10.1098/rsbl.2003.0025

809 Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … Yu, D. W. (2013).
810 Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology*
811 *Letters*, *16*(10), 1245–1257. https://doi.org/10.1111/ele.12162

812 Klappenbach, J. A., Saxman, P. R., R., C. J., & Schmidt, T. M. (2001). rrndb: the Ribosomal
813 RNA Operon Copy Number Database. *Nucleic Acids Research*, *29*(1), 181–184.
814 https://doi.org/10.1093/nar/29.1.181

815 Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J.
816 (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI
817 region for metabarcoding metazoan diversity: application for characterizing coral reef fish
818 gut contents. *Front Zool*, *10*, 34. https://doi.org/10.1186/1742-9994-10-34

819 Macheriotou, L., Guilini, K., Bezerra, T. N., Tytgat, B., Nguyen, D. T., Phuong Nguyen, T. X.,
820 … Rigaux, A. (2019). Metabarcoding free-living marine nematodes using curated 18S and
821 CO1 reference sequence databases for species-level taxonomic assignments. *Ecology and*
822 *Evolution*, *9*(1), 1–16. https://doi.org/10.1002/ece3.4814

823 Machida, R. J., Leray, M., Ho, S. L., & Knowlton, N. (2017). Data Descriptor: Metazoan
824 mitochondrial gene sequence reference datasets for taxonomic assignment of environmental
825 samples. *Scientific Data*, *4*. https://doi.org/10.1038/sdata.2017.27

826 Mahe, F., Rognes, T., Quince, C., De Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-
827 scalable and high-resolution amplicon clustering. *PeerJ*, *3*. https://doi.org/Artn

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

828 E142010.7717/Peerj.1420
829 Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., … De Vargas, C. (2015).
830 Marine protist diversity in European coastal waters and sediments as revealed by high-
831 throughput sequencing. *Environmental Microbiology*, *17*(10), 4035–4049.
832 https://doi.org/10.1111/1462-2920.12955
833 Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. New
834 York, NY: Columbia University Press. Retrieved from
835 http://www.hup.harvard.edu/catalog.php?isbn=9780674862500
836 McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive
837 Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, *8*(4), e61217.
838 https://doi.org/10.1371/journal.pone.0061217
839 Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-
840 throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems.
841 *Genome Biology*, *12*(11), R112. https://doi.org/10.1186/gb-2011-12-11-r112
842 Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the
843 Denoisers: an independent evaluation of microbiome sequence error-correction approaches.
844 *PeerJ*, *6*, e5364. https://doi.org/10.7717/peerj.5364
845 Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., …
846 Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology*
847 *Resources*, *18*(5), 927–939. https://doi.org/10.1111/1755-0998.12895
848 Oksanen, J., Blanchet, Guillaume F. Friendly, M., Kindt, R., Legendre, P., McGlinn, D.,
849 Minchin, R. P., … Wagner, H. (2018). vegan: Community Ecology Package. Retrieved from
850 https://cran.r-project.org/package=vegan
851 Pansu, J., Giguet-Covex, C., Ficetola, G. F., Gielly, L., Boyer, F., Coissac, E., … Arnaud, F.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

852 (2015). Environmental DNA metabarcoding to investigate historic changes in biodiversity.
853 *Genome*, *58*(5), 264.
854 Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: assessing small
855 subunit rRNA primers for marine microbiomes with mock communities, time series and
856 global field samples. *Environ Microbiol*, *18*(5), 1403–1414. https://doi.org/10.1111/1462-
857 2920.13023
858 Pawlowski, J. W., Christen, R., Lecroq, B., Bachar, D., Shahbazkia, H. R., Amaral-Zettler, L., &
859 Guillou, L. (2011). Eukaryotic Richness in the Abyss: Insights from Pyrotag Sequencing.
860 *PLoS One*, *6*(4). https://doi.org/e1816910.1371/journal.pone.0018169
861 Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., … Pei, Z.
862 (2010). Diversity of 16S rRNA genes within individual prokaryotic genomes. *Applied and*
863 *Environmental Microbiology*, *76*(12), 3886–3897. https://doi.org/10.1128/AEM.02953-09
864 Phillips, J. D., Gillis, D. J., & Hanner, R. H. (2019, March 1). Incomplete estimates of genetic
865 diversity within species: Implications for DNA barcoding. *Ecology and Evolution*. John
866 Wiley & Sons, Ltd. https://doi.org/10.1002/ece3.4757
867 Plouviez, S., Shank, T. M., Faure, B., Daguin-Thiebaut, C., Viard, F., Lallier, F. H., & Jollivet,
868 D. (2009). Comparative phylogeography among hydrothermal vent species along the East
869 Pacific Rise reveals vicariant processes and population expansion in the South. *Molecular*
870 *Ecology*, *18*(18), 3903–3917. https://doi.org/10.1111/j.1365-294X.2009.04325.x
871 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2012).
872 The SILVA ribosomal RNA gene database project: improved data processing and web-
873 based tools. *Nucleic Acids Research*, *41*(D1), D590–D596.
874 https://doi.org/10.1093/nar/gks1219
875 R Core Team. (2018). R: A language and environment for statistical computing. R Foundation

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

876 for Statistical Computing, Vienna, Austria.

877 Salazar, G., Cornejo-Castillo, F. M., Benitez-Barrios, V., Fraile-Nuez, E., Alvarez-Salgado, X.
878 A., Duarte, C. M., … Acinas, S. G. (2016). Global diversity and biogeography of deep-sea
879 pelagic prokaryotes. *Isme Journal*, *10*(3), 596–608. https://doi.org/10.1038/ismej.2015.137

880 Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber,
881 C. F. (2009). Introducing mothur: Open-source, platform-independent, community-
882 supported software for describing and comparing microbial communities. *Applied and*
883 *Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

884 Shapiro, B. J., Leducq, J. B., & Mallet, J. (2016). What Is Speciation? *PLoS Genetics*, *12*(3),
885 e1005860. https://doi.org/10.1371/journal.pgen.1005860

886 Sinniger, F., Pawlowski, J. W., Harii, S., Gooday, A. J., Yamamoto, H., Chevaldonné, P., …
887 Creer, S. (2016). Worldwide analysis of sedimentary DNA reveals major gaps in taxonomic
888 knowledge of deep-sea benthos. *Frontiers in Marine Science*, *3*(June), 92.
889 https://doi.org/10.3389/FMARS.2016.00092

890 Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., de la Rasilla, M., Lalueza-Fox, C., … Meyer, M.
891 (2017). Neandertal and Denisovan DNA from Pleistocene sediments. *Science (New York,*
892 *N.Y.)*, *356*(6338), 605–608. https://doi.org/10.1126/science.aam9695

893 Sokal, R. R., & Crovello, T. J. (1970). The Biological Species Concept : A Critical Evaluation.
894 *The American Naturalist*, *104*(936), 127–153.

895 Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA
896 barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are
897 coamplified. *Proceedings of the National Academy of Sciences of the United States of*
898 *America*, *105*(36), 13486–13491. https://doi.org/10.1073/pnas.0803076105

899 Stat, M., Huggett, M. J., Bernasconi, R., Dibattista, J. D., Berry, T. E., Newman, S. J., … Bunce,

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

900 M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a
901 tropical marine environment. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-
902 12501-5
903 Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA.
904 *Molecular Ecology*, *21*(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x
905 Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T. G., & Fontaneto, D. (2012).
906 The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in
907 biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences of*
908 *the United States of America*, *109*(40), 16208–16212.
909 https://doi.org/10.1073/pnas.1209160109
910 Valentini, A., Pompanon, F., & Taberlet, P. (2009, February 1). DNA barcoding for ecologists.
911 *Trends in Ecology and Evolution*. Elsevier Current Trends.
912 https://doi.org/10.1016/j.tree.2008.09.011
913 Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., … Dejean, T.
914 (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA
915 metabarcoding. *Molecular Ecology*, *25*(4), 929–942. https://doi.org/10.1111/mec.13428
916 Wangensteen, O. S., & Turon, X. (2016). Metabarcoding Techniques for Assessing Biodiversity
917 of Marine Animal Forests. In S. Rossi, L. Bramanti, A. Gori, & C. Orejas Saco del Valle
918 (Eds.), *Marine Animal Forests* (pp. 1–29). Cham: Springer International Publishing.
919 https://doi.org/10.1007/978-3-319-17001-5_53-1
920 Xiong, W., & Zhan, A. (2018). Testing clustering strategies for metabarcoding-based
921 investigation of community–environment interactions. *Molecular Ecology Resources*, *18*(6),
922 1326–1338. https://doi.org/10.1111/1755-0998.12922
923 Yoccoz, N. G., Brathen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., … Taberlet, P.

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

924 (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular*
925 *Ecology*, *21*(15), 3647–3655. https://doi.org/10.1111/j.1365-294X.2012.05545.x
926 Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity
927 soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring.
928 *Methods in Ecology and Evolution*, *3*(4), 613–623. https://doi.org/10.1111/j.2041-
929 210X.2012.00198.x
930 Zhan, A., Bailey, S. A., Heath, D. D., & Macisaac, H. J. (2014). Performance comparison of
931 genetic markers for high-throughput sequencing-based biodiversity assessment in complex
932 communities. *Molecular Ecology Resources*, *14*(5), 1049–1059.
933 https://doi.org/10.1111/1755-0998.12254
934 Zinger, L., Chave, J., Coissac, E., Iribar, A., Louisanna, E., Manzi, S., … Taberlet, P. (2016).
935 Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys
936 based on soil DNA. *Soil Biology and Biochemistry*, *96*, 16–19.
937 https://doi.org/10.1016/j.soilbio.2016.01.008
938

939 **DATA ACCESSIBILITY**

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

bioRxiv preprint first posted online Aug. 1, 2019; doi: http://dx.doi.org/10.1101/717355. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

946 **AUTHOR CONTRIBUTIONS**

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