



HAL
open science

Enhancing DNA metabarcoding performance and applicability with bait capture enrichment and DNA from conservative ethanol

Maïlys Gauthier, Lara Konecny-Dupré, Agnès Nguyen, Vasco Elbrecht, T. Datry, Christophe Jean Douady, Tristan Lefébure

► To cite this version:

Maïlys Gauthier, Lara Konecny-Dupré, Agnès Nguyen, Vasco Elbrecht, T. Datry, et al.. Enhancing DNA metabarcoding performance and applicability with bait capture enrichment and DNA from conservative ethanol. *Molecular Ecology Resources*, 2020, 20 (79-96), pp.79-96. 10.1111/1755-0998.13088 . hal-02328256

HAL Id: hal-02328256

<https://univ-lyon1.hal.science/hal-02328256>

Submitted on 3 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 TITLE: Enhancing DNA metabarcoding performance and applicability with bait capture enrichment
2 and DNA from conservative ethanol

3

4 RUNNING TITLE:

5 Capture enrichment & etDNA for metabarcoding

6

7 AUTHORS:

8 Gauthier M.^{1, 2*}, Konecny-Dupré L.¹, Nguyen A.³, Elbrecht V.⁴, Datry T.², Douady C.J.¹, Lefébure
9 T.^{1*}

10

11 AUTHORS AFFILIATION

12 1. Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5023, ENTPE, Laboratoire
13 d'Ecologie des Hydrosystèmes Naturels et Anthropisés, F-69622 Villeurbanne, France

14 2. IRSTEA, UR-RiverLy, Centre de Lyon-Villeurbanne, Villeurbanne Cedex, France

15 3. Biofidal, Vaulx-en-Velin, France

16 4. Centre for Biodiversity Genomics, University of Guelph, 50 Stone Road East, Guelph, Ontario,
17 N1G 2W1, Canada

18 *Corresponding authors: mailys.gauthier@irstea.fr, tristan.lefebure@univ-lyon1.fr

19

20 ABSTRACT

21 Metabarcoding is often presented as an alternative identification tool to compensate for coarse
22 taxonomic resolution and misidentification encountered with traditional morphological approaches.

23 However, metabarcoding comes with two major impediments which slow down its adoption. First, the
24 picking and destruction of organisms for DNA extraction are time and cost consuming and do not allow

25 organism conservation for further evaluations. Second, current metabarcoding protocols include a PCR
26 enrichment step which induces errors in the estimation of species diversity and relative biomasses. In

27 this study, we first evaluated the capacity of capture enrichment to replace PCR enrichment using

28 controlled freshwater macrozoobenthos mock communities. Then, we tested if DNA extracted from the
29 fixative ethanol (etDNA) of the same mock communities can be used as an alternative to DNA extracted
30 from pools of whole organisms (bulk DNA). We show that capture enrichment provides more reliable
31 and accurate representation of species occurrences and relative biomasses in comparison with PCR
32 enrichment for bulk DNA. While etDNA does not permit to estimate relative biomasses, etDNA and
33 bulk DNA provide equivalent species detection rates. Thanks to its robustness to mismatches, capture
34 enrichment is already an efficient alternative to PCR enrichment for metabarcoding and, if coupled to
35 etDNA, is a time-saver option in studies where presence information only is sufficient.

36

37 KEYWORDS

38 bait capture, ethanol DNA, DNA enrichment, metabarcoding, biomass estimation, macrozoobenthos

39

40 INTRODUCTION

41 Reliable and accurate taxa identification is fundamental in biological sciences. Poor taxonomic
42 identification can lead to cascades of error affecting our knowledge and understanding not only in
43 theoretical and fundamental biology, but also in applied fields leading to poor management decisions
44 (Bortolus, 2008). It distorts our ability to infer processes in ecology and evolution, to manage and
45 conserve human-impacted systems and to carry out human health and resource programs (Bortolus,
46 2008; Leys, Keller, Räsänen, Gattolliat, & Robinson, 2016; Prié, Puillandre, & Bouchet, 2013). Poor
47 taxonomic identification occurs when identification to species level is not possible (coarse taxonomic
48 resolution) or when identification is incorrect (misidentification). For example, Martin, Adamowicz,
49 and Cottenie (2016) investigated macrozoobenthos community distribution in freshwater streams at
50 different taxonomic level (family, genus and species) and found a spatial structure only when the species
51 identification level was reached. In Vietnam, misidentification in *Anopheles* species lead to
52 mismanagement in control program of the Malaria disease as the main target species was a non-vector
53 one (Van Bortel et al., 2001). This kind of difficulties leading to poor taxonomic identification and thus
54 to misinterpretation and erroneous conclusions is often associated to morphological based taxonomy

55 (Baird & Hajibabaei, 2012; Creer et al., 2016). Cryptic species, limited expertise, damaged or juvenile
56 specimens and even cost and time constraints in the case of applied research are all problems
57 encountered in morphological identification leading to misidentification or coarse taxonomic resolution
58 (Bringloe, Cottenie, Martin, & Adamowicz, 2016; Hajibabaei, Baird, Fahner, Beiko, & Golding, 2016;
59 Ji et al., 2013; Sweeney et al., 2011).

60 In the past decade, DNA-based identification has been proposed as an alternative to
61 morphological approaches. DNA barcoding uses the sequence of a genetic marker of one specimen,
62 hereafter called a DNA barcode, usually of an organelle genome for eukaryotes (e.g. mitochondria for
63 animals and chloroplast for plants, see Creer et al. (2016) for overview) and assign it to a species name
64 within a reference database (Hebert, Ratnasingham, & de Waard, 2006). DNA barcoding supposedly
65 addresses limitations in morphological identification by accurately discriminating species regardless of
66 their morphology (Sweeney et al., 2011), development stages (Hubert, Delrieu-Trottin, Irisson, Meyer,
67 & Planes, 2010) or sex (Forshaw, 2010). Recent advances in high throughput technologies enabled the
68 emergence of DNA metabarcoding, the barcoding of pool of specimens in a single reaction which permit
69 to work on whole community at once (Pompanon et al, 2011). Metabarcoding has already been adopted
70 as a routine identification tool in microbial community ecology (Abdelfattah, Malacrino, Wisniewski,
71 Cacciola, & Schena, 2018). In macro-organism community studies, the last ten years have seen the
72 development of invertebrate derived DNA (iDNA, (Calvignac-Spencer et al., 2013; Schnell et al., 2012))
73 and the analysis of DNA associated with environmental matrices such as soil or water (eDNA, (Harper
74 et al., 2018; Thomsen & Willerslev, 2015)). Metabarcoding was instrumental in the emergence of these
75 new approaches for which, by definition, there is no non-molecular alternative. Conversely, in
76 conventional macro-organism community studies where the traditional identification using
77 morphological criteria has been used for many years, metabarcoding is still currently marginally used
78 as a routine identification tool. Yet, metabarcoding could enhance the capacity to characterize these
79 communities through iDNA, eDNA or through the extraction of DNA from a homogenate of whole
80 organisms collected together (bulk DNA; (Deiner et al., 2017)). A major limitation to the
81 democratization of metabarcoding is the absence of standardized metabarcoding protocol that has been

82 established and validated by the scientific community (Leese et al., 2018). In particular, methodological
83 roadblocks are encountered throughout each step of sample processing (i.e. DNA extraction, enrichment
84 of the targeted DNA barcode, sequencing, bioinformatic treatment and taxonomic assignment).
85 Hereafter, we focused on two major metabarcoding roadblocks: specimen picking before extraction and
86 DNA barcode enrichment.

87

88 Specimen picking is a major concern for bulk DNA samples and is particularly critical when the ratio
89 of targeted organism over substrate is low (e.g. invertebrates sampled in streambed). It aims to separate
90 individuals from substrate (e.g. leaves, sand...) prior to DNA extraction to avoid PCR inhibition
91 (Elbrecht, Vamos, Meissner, Aroviita, & Leese, 2017) and to limit the quantity of material to be
92 processed during DNA extraction. This step is time and cost consuming. Direct extraction from fixative
93 agent, usually ethanol, has been proposed as a time-saver alternative to specimen picking (Hajibabaei,
94 Spall, Shokralla, & van Konynenburg, 2012; Zizka, Leese, Peinert, & Geiger, 2018). When fixative
95 agent is used as a DNA template, organisms are not destroyed and are conserved for further taxonomic
96 work or downstream analyses (Leese et al., 2016). Similar techniques have already been employed by
97 the ancient DNA community where the destruction of samples such as museum specimens is to be
98 avoided (Paijmans, Fickel, Courtiol, Hofreiter, & Forster, 2016; van der Valk, Lona Durazo, Dalen, &
99 Guschanski, 2017). Yet, little research on bulk DNA has been conducted on this alternative (but see
100 Hajibabaei et al. (2012); Zizka et al. (2018)) and a rigorous comparison with traditional specimen
101 picking is warranted before it can be used as a standard template of DNA in community studies.

102

103 PCR enrichment bias is often considered as the most problematic roadblock in metabarcoding
104 because it may alter species detection and relative abundance recovery of species (Elbrecht & Leese,
105 2017; Leese et al., 2018; Piñol, Senar, & Symondson, 2018). Prior to sequencing, DNA barcodes are
106 first amplified by PCR using primers that may not have the same number of mismatches with the
107 targeted sequences across taxa (Piñol, Mir, Gomez-Polo, & Agustí, 2015). During a PCR, if DNA
108 barcode sequences from two different species are in equimolar concentration, but that the first species

109 presents less mismatches with the primers, this species' sequence will be amplified preferentially.
110 Consequently, amplification efficiency is expected to be non-equal across taxa, leading from under-
111 amplification to no amplification of some taxa in the worst case scenario. PCR bias was demonstrated
112 for fungi (Bellemain et al., 2010), bacteria (Frank et al., 2008), invertebrates (Piñol et al., 2015) and
113 vertebrates (Arif, Khan, Al Sadoon, & Shobrak, 2011). Other factors affect PCR like GC content (Aird
114 et al., 2011) or inhibitors that can remain after DNA extraction but primer bias is commonly presented
115 as the major cause of biases in metabarcoding (e.g. (Elbrecht & Leese, 2015; Piñol et al., 2015; Pinto &
116 Raskin, 2012)). In consequence, a lot of work focused on primer design to decrease PCR biases with,
117 for instance, the use of several primer pairs, degenerated primers or amplification of several DNA
118 barcodes (Drummond et al., 2015; Elbrecht & Leese, 2017; Elbrecht et al., 2016; Gibson et al., 2015;
119 Jusino et al., 2019; Leray & Knowlton, 2017; Zhang, Chain, Abbott, & Cristescu, 2018). These efforts
120 increased the species detection rate but the quantitative bias was not solved completely (Piñol et al.,
121 2018).

122

123 Avoiding PCR enrichment will, by definition, solve the PCR bias issue (Porter & Hajibabaei, 2018).
124 Low (Linard, Crampton-Platt, Gillett, Vogler, & Timmermans, 2015) or high (Porter & Hajibabaei,
125 2018) coverage metagenome sequencing (i.e. sequencing a community DNA without any enrichment)
126 can be used to assemble entire organelle genomes. This approach provides an efficient way to recover
127 species richness and taxa relative biomass, although the proportion of organelle reads is extremely low
128 making metagenome sequencing much more expensive than PCR metabarcoding (Bista et al., 2018;
129 Gómez-Rodríguez, Crampton-Platt, Timmermans, Baselga, & Vogler, 2015; Zhou et al., 2013).
130 Furthermore, only a small part of an organelle genome is usable for taxonomic assignment as reference
131 databases mostly contain DNA barcode sequences (e.g. COI for metazoan, 16S for bacteria, ITS for
132 fungi, (Creer et al., 2016)). Although methods are being developed to reduce organelle genome
133 sequencing cost (Macher, Zizka, Weigand, & Leese, 2017), the construction of exhaustive organelle
134 genome reference databases will be a long-term and expensive process. Another PCR-free alternative is
135 capture enrichment where targeted sequences hybridize to baits and are retrieved by magnetism (Dowle,

136 Pochon, C. Banks, Shearer, & Wood, 2016). Contrary to metagenome sequencing, capture enrichment
137 increases the proportion of targeted reads reducing the sequencing cost (Jones & Good, 2016). Baits are
138 long oligonucleotides (more than 60 bp) which are designed from reference sequences. Capture
139 enrichment has already been extensively used in genomics and genetics where it is used to retrieve
140 thousands of loci (e.g. exons capture, (Hodges et al., 2007)) and SNPs arrays (Yang et al., 2009) of one
141 species in a single reaction prior to sequencing. It is also commonly used for ancient DNA to both
142 increase the amount of endogenous DNA and allow loci enrichment where DNA fragmentation tends to
143 complicate PCR reactions (Avila-Arcos et al., 2011; Horn, 2012). Capture enrichment has also recently
144 been used in phylogenetics where the goal is to target thousands of loci in many related species
145 (McCormack et al., 2013; Phuong & Mahardika, 2018). Conversely, capture enrichment is not
146 commonly used in community ecology (but see Dowle, Pochon, J, Shearer, & Wood, (2016); Shokralla
147 et al., (2016)) which opens new challenges as in this new context the diversity of organism is extremely
148 high but the number of targeted loci is rather small. Capture enrichment should be more robust to
149 identify taxa in a community than PCR enrichment because (i) thousands of different baits can be
150 designed and (ii) when sequences are unknown or species are polymorphic, few mismatches between
151 the baits and the targeted sequences should not bias DNA enrichment as they would do with PCR
152 primers (Li, Hofreiter, Straube, Corrigan, & Naylor, 2013; Paijmans et al., 2016; Portik, Smith, & Bi,
153 2016). For example, one bait designed for a species of the genus *Danio* permitted to detect others *Danio*
154 species in the Amazon basin where the reference species was absent (Mariac et al., 2018). In a single
155 study, capture enrichment was compared to PCR enrichment and was found to detect more taxa but was
156 unable to estimate relative abundances (Dowle et al., 2016). Liu et al. (2016) and Wilcox et al. (2018)
157 have shown that relative abundances can be recovered with capture enrichment if species-specific
158 corrections were to be applied to take into account variation in the number of mitochondria copy number
159 among species (Liu et al., 2016; Wilcox et al., 2018). Such corrections are unfortunately not suitable to
160 the complexity of field samples. Dowle et al. (2016) study was based on natural communities only
161 described at a coarse taxonomic level (family to genus level) and with indirect biomass measurements.

162 Thus, the capacity of capture enrichment to describe both community diversity, and relative biomass
163 without species-specific biomass corrections, requires further testing with controlled communities.

164

165 In this study, we first investigated the capacity of cytochrome oxidase subunit I (COI) gene capture
166 enrichment to detect taxa and retrieve initial biomass without species specific correction. Second, we
167 evaluated if DNA extracted from ethanol (etDNA) can be used as an alternative to organism picking
168 and homogenization (bulk DNA) by assessing species detection and initial biomass recovery of this
169 alternative template DNA with PCR and capture enrichment. Tests were carried out using two types of
170 freshwater mock communities (MC): (i) low diversity MC (10 species) with variable dry biomass across
171 taxa and (ii) high diversity MC (52 taxa) with homogeneous biomass across taxa from (Elbrecht &
172 Leese, 2015).

173

174 MATERIALS & METHODS

175 *1. Mock community design*

176 Freshwater macrozoobenthos specimens for the 10 species mock communities (MC) were sampled
177 from various streams in east of France during May 2017, except for two species, *Gammarus fossarum*
178 and *Chironomus riparius*, which came from Irstea livestock, France (ECOTOX team, RiverLy, France).
179 The 10 species were chosen to represent a wide taxonomic range and because they can be easily
180 identified to the species level by the naked eye (table 1). Two hundred milliliters of ethanol 96% (EtOH)
181 were dispense in bottles and individuals were placed alive in the ethanol. Eight samples with different
182 relative biomass for the 10 species were constructed (Figure 1, table 1). Samples were then stored for 6
183 months at 4°C until DNA extraction. DNA was extracted from the whole organisms (bulk DNA) and
184 from the preservative EtOH (etDNA).

185

186 The 52 taxa mock communities were designed by Elbrecht and Leese (2015). In a nutshell, for ten
187 different MC, a roughly similar dry biomass was collected from 52 taxa identified to the lowest
188 taxonomic level based on morphology, and the homogenized tissue pool was then extracted for each

189 sample using a salt extraction protocol (Figure 1, see Elbrecht & Leese (2015) for details). Because the
190 biomass was fairly homogeneous among taxa in these MC, we only used these MC to compare the
191 performance of capture and PCR enrichment in predicting species occurrences.

192

193 *2. DNA extraction of the 10 species mock communities*

194 *Bulk DNA.* For each MC, individuals were picked up and sorted by species in petri dishes. They were
195 let to dry overnight and dry biomass for each species was weighted. Mollusc shells and Trichoptera
196 cases were removed prior to the weighing. Individuals were pooled together and the entire community
197 was grounded with a bead mill MM200 (Retsch) during 4 min at 30 Hz. The whole homogenate was
198 extracted (mean \pm SD: 245.8 \pm 28.6 mg) with FastDNA® Spin Kit for Soil (MP Biomedicals, USA)
199 following manufacturer's protocol. The extracted DNA was purified with Agencourt AMPure XP
200 purification beads (Beckman Coulter, USA) to further remove solvents. A negative extraction control
201 was also made, following the same extraction protocol but without starting material.

202

203 *etDNA.* For each mock community, roughly 50 mL of preservative ethanol was collected. Glycogen
204 and sodium acetate were added to precipitate DNA and samples were placed at -80°C for at least 72
205 hours. After centrifugation, ethanol was removed and after total ethanol evaporation, the dried residue
206 was dissolved in the buffer solution of NucleoSpin Tissue® kit (Macherey-Nagel GmbH, Germany) and
207 DNA was extracted following manufacturer's instructions. As the amount of DNA was limiting,
208 extractions were repeated twice: a first extraction to compare PCR and capture enrichment methods
209 (extraction I); a second to compare two different primer pairs performance for this supposedly degraded
210 DNA template (extraction II). Four negative extraction controls were also made during extraction II by
211 filling tubes with 96% ethanol which were then extracted following the same protocol.

212

213 *3. COI mock community reference databases*

214 The 10 species reference database was built by sequencing the COI Folmer region of each species of
215 the 10 species mock communities (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994). One specimen per

216 species was extracted following the same protocol as for bulk DNA. PCR reactions were performed in
217 a total volume of 25 μ L with 1X of PCR standard buffer (including 3 mM MgCl₂, Eurobio, France), 0.05
218 U/ μ L of EurobioTaq DNA polymerase (Eurobio, France), 0.8 mM of each dNTP (Eurogentec,
219 Belgium), 0.1 mg/mL of BSA (New England BioLabs, USA), 0.4 μ M of each primer
220 (LCO1490/HCO2198, Table S1) and 0.5 μ L of template DNA. The amplification consisted in an initial
221 denaturation at 95°C for 3 min, followed by 40 cycles of denaturation at 95°C for 20 s, annealing at
222 51°C for 30 s and extension at 72°C for 45 s, with a final extension at 72°C for 5 min. Purified template
223 DNA was sequenced on both strands with the PCR primers using standard Sanger sequencing (Biofidal,
224 France). All sequences were manually checked for errors and cleaned up with Finch TV software 1.4.0.
225 This protocol did not work for one species (*Heptagenia sulphurea*) and its sequence was downloaded
226 from NCBI to complete the 10 species reference database (GenBank Accession Number HE651395.1).
227

228 The 52 taxa reference database was already available for the 52 taxa MC (Elbrecht & Leese, 2017).
229 It includes the haplotypes of all organisms used in their experiment, leading to a reference database of
230 212 COI sequences. For one taxon (Nematoda), Elbrecht and Leese (2017) could not get any
231 amplification. Since this taxon is missing in Elbrecht and Leese (2017) haplotype reference database, it
232 was not taken into account for downstream analysis.

233

234 *4. Amplicon sequencing and analysis*

235 *Library preparation and sequencing.* For the 10 species and 52 taxa MC bulk DNA and etDNA
236 (extraction II), a 421 bp fragment within the COI Folmer region was amplified with the BF2/BR2 primer
237 set with Illumina Nextera tails which is well suited for freshwater macrozoobenthos (Elbrecht & Leese,
238 2017), Table S1). For etDNA extractions I and II, a shorter region of 178 bp (Fwh1 primer set with
239 Illumina Nextera tails, (Vamos, Elbrecht, & Leese, 2017), Table S1) was also targeted. PCR reactions
240 were performed in triplicates in a total volume of 25 μ L with 1X of PCR buffer (including 3mM MgCl₂
241 and 400 μ M each dNTP, QIAGEN® Multiplex PCR kit, Qiagen, Germany), 0.5 μ M of each primer and
242 10 ng of template DNA for bulk DNA or 5 μ L of template DNA for etDNA. The amplification consisted

243 in an initial denaturation at 95°C for 5 min, followed by 25 (bulk DNA) or 35 (etDNA) cycles of
244 denaturation at 95°C for 30 s, annealing at 50°C (bulk DNA) or 52°C (etDNA) for 30 s and extension
245 at 72°C for 2 min, with a final extension at 72°C for 10 min. Triplicate PCR products were pooled and
246 purified with Agencourt AMPure XP purification beads (Beckman Coulter, USA) and quantified using
247 QuantiFluor® dsDNA System (Promega, USA). Ten ng of each purified PCR products were then used
248 in a second PCR to dual index each sample with a unique tag combination (Table S3 for tag
249 combinations details) and to add Illumina adapters. PCR reactions were performed in a total volume of
250 25 µL with 1X of PCR buffer (BIOAmp® Blend Mix, Biofidal, France), 2.5 nM of MgCL₂, 200 µM of
251 each dNTP, 0.25 µM of Illumina primer (Table S1) and 0.02 U/µL of HOT BioAmp Taq (Biofidal,
252 France). PCR products were purified with Agencourt AMPure XP purification beads (Beckman Coulter,
253 USA), quantified using QuantiFluor® dsDNA System (Promega, USA) and pooled at the same
254 concentration (2 nM) for sequencing. The PCR amplicon libraries were sequenced using a 2*250 paired-
255 end V3 MiSeq sequencing kit (Biofidal, France). Negative controls were included to every PCR during
256 the library preparation.

257

258 *Bioinformatic analysis.* Reads were delivered demultiplexed and adapter trimmed. The reads of the
259 52 taxa mock communities, bulk DNA and etDNA of the 10 species mock communities were processed
260 independently with the same bioinformatic pipeline. First, forward and reverse reads were merged with
261 Vsearch 2.8.4 with a minimum overlapping of 10 nucleotides and a maximum difference of the
262 overlapping region of 5 nucleotides (Rognes, Quince, Nichols, Flouri, & Mahé, 2016). Then, the primer
263 regions were removed from the merged reads with cutadapt 1.9.1 (M. Martin, 2014) and the reads were
264 quality filtered (maximum expected error (ee) of 1, minimum length of 200, no Ns allowed) and
265 dereplicated with Vsearch. Sequences observed less than twice (< 2 reads) were removed. Then,
266 chimeras were *de novo* detected and removed with Vsearch. Finally, sequences were clustered with a
267 similarity cutoff of 97% identity into MOTUs (Molecular Operational Taxonomic Units). MOTUs were
268 assigned to species using the 10 species reference database or the 52 taxa reference database with the
269 blastn algorithm (Camacho et al., 2009). Only alignments with an e-value under 1E-10, a query cover

270 over 200 bases for bulk DNA and over 90 bases for etDNA and an identity over 97% were conserved
271 as good alignments for further analysis. MOTUs were also compared to a complete COI protein
272 reference database made from 8 taxa (*Asellus aquaticus* ADA69754.1, *Daphnia pulex* AAD33231.1,
273 *Dinocras cephalotes* AGZ03516.1, *Gammarus fossarum* YP_009379680.1, *Physella acuta*
274 YP_008994230.1, *Radix balthica* HQ330989.1, *Sericostoma personatum* AJR19241.1, *Thremma*
275 *gallicum* AJR19254.1) present in one or both MCs using diamond 0.9.22 (blastx, more sensitive option,
276 e-value threshold of 1E-10 (Buchfink, Xie, & Huson, 2014)). This allowed us to detect COI sequences
277 even if they did not belong to the species used in the MC. We also estimated the rate of contamination
278 by comparing the quality filtered reads to a protein database containing the proteomes of 100 eukaryotic
279 species from Ensembl (Zerbino et al., 2017) and Ensembl Metazoa (Kersey et al., 2017) as well as the
280 proteomes of 837 prokaryotic species retrieved from the Microbial Genome Database for Comparative
281 Analysis (Uchiyama, Mihara, Nishide, & Chiba, 2015) selecting one species per genus. Sequences were
282 assigned to coarse taxonomic groups (archaea, eubacteria, fungi, plant, protist, protostomia– here
283 Arthropoda, Annelida, Brachiopoda, Mollusca, Nematoda and Platyhelminthes – and other metazoa –
284 corresponding to Deuterostomia) using the best diamond hit (blastx, more sensitive option, e-value
285 threshold of 1E-10).

286

287 5. Capture sequencing and analysis

288 *Bait design.* While one of the primary goals of the present study is to assess capture versus PCR
289 enrichment efficiency using mock communities, we chose to develop a larger set of baits that could be
290 used in future assessment of French freshwater macrozoobenthos diversity. Up to now and aside from
291 non-Chironomidae dipteran, 3,245 macrozoobenthos species belonging to 22 orders have been identified
292 in small and medium streams of France (Aukema & Rieger, 2013; Coppa, 2019; D'Hondt & Ben
293 Ahmed, 2009; Dusoulier, 2008; Gargominy et al, 2011; Grand & Boudot, 2007; Henry & Magniez,
294 1983; Le Doaré & Vinçon, 2019; Pattée & Goubault, 1981; Piscart & Bollache, 2012; Queney, 2011;
295 Serra et al, 2015; Souty-Grosset et al, 2006; Thomas, 2019; Tillier, 2019; Vallenduuk, 2004). One
296 hundred twenty two species of six orders can be considered as marginal in streams and were thus

297 discarded. All the available COI sequences were downloaded for the 3,123 remaining species from
298 GenBank and BOLD with PrimerMiner 0.18 in December 2016 (Elbrecht & Leese, 2016). Four orders
299 had less than half of their known taxa with available COI barcode and were discarded. Within the 12
300 remaining orders, the species without COI sequences were barcoded whenever possible, i.e. when
301 tissues could be obtained. Organisms were extracted using Chelex (BioRad, USA) and the PCR and
302 sequencing conditions were identical to those of section 3 of the methods. Sequences were deposited on
303 GenBank (97 species, GenBank Accession Numbers MK584300:MK584515). At the end, 1,525 species
304 out of 1,689 known species had a DNA barcode available, representing more than 90% of the targeted
305 species (Table S2). For bait development, sequences were processed by order. First, COI sequences
306 were aligned with blastx 2.7.1 to a reference *Drosophila yakubai* COI sequence (Accession Number:
307 NC_001322.1) and identical sequences were collapsed using a perl script
308 (http://github.com/TristanLefebure/collapse_to_uniq_seq). Then, 120 bp baits were constructed *in silico*
309 using BaitFisher with a tiling of 60 bp and a cluster threshold of 5%, leading to a total of 15,038 baits
310 generated from 100,367 unique sequences (Mayer et al., 2016). The COI *in silico* baits were then sent
311 for RNA bait synthesis to Arbor Biosciences (USA).

312

313 *Library preparation, hybridization and sequencing.* Starting quantity of DNA was 1 µg for bulk
314 DNA (10 and 52 taxa MC) and between 14 and 57 ng for etDNA (extraction I). DNA of each sample
315 was sheared into approximately 600 bp nucleotide fragments by ultrasound sonication with a Qsonica
316 Q800R (Qsonica, USA). Library preparation was conducted using NEBNext® Ultra™ II DNA Library
317 Prep Kit for Illumina® (New England BioLabs, USA) following manufacturer's instructions. Briefly,
318 after sheared DNA end repair, the 5' ends were phosphorylated and the 3' ends were A-tailed. Then,
319 Illumina Nextera tails (Table S1) were ligated to the DNA fragments followed by a clean-up and a size
320 selection of 500-700 nucleotides long fragments with Agencourt AMPure XP purification beads
321 (Beckman Coulter, USA). Finally, DNA fragments were amplified to dual index the libraries (Table S3
322 for tag combinations details) and to add Illumina adapters (Table S1). COI capture enrichment was
323 conducted using myBaits® Custom kit following manufacturer's instructions (Arbor Biosciences, USA).

324 One hundred ng of library DNA was used for capture enrichment for bulk DNA and 230 ng for etDNA
325 following manufacturer's instructions for degraded DNA. Baits were diluted 10 times and hybridization
326 lasted 24h for both DNA templates with the exception that hybridization for etDNA was done at 55°C
327 instead of 65°C following manufacturer's instructions for degraded DNA. The final library amplification
328 was performed in a total volume of 50 µL per reaction with the KAPA HiFi DNA Polymerase (Kapa
329 Biosystems, USA) with P5 and P7 Illumina primers (Table S1) using the following conditions: initial
330 denaturation at 98°C for 2 min, followed by 21 cycles of denaturation at 98°C for 20 s, annealing at
331 60°C for 30 s and extension at 72°C for 1 min, with a final extension at 72°C for 5 min. Capture library
332 concentrations were determined by qPCR with a KAPA qPCR kit (KAPA Library Quant Kit, Kapa
333 Biosystems, USA) and pooled at the same concentration for sequencing. The capture libraries were
334 sequenced using a 2*300 paired-end V3 MiSeq sequencing kit (Biofidal, France).

335

336 *Bioinformatic analysis.* Reads were delivered demultiplexed and adapter trimmed. As for PCR
337 analysis, the reads of the 52 taxa MC, bulk DNA and etDNA of the 10 species MC were processed
338 independently with the same bioinformatic pipeline. First, forward and reverse reads were merged with
339 Vsearch with a minimum overlapping of 10 nucleotides and a maximum difference of the overlapping
340 region of 5 nucleotides (Rognes et al., 2016). Because shearing could lead to fragment longer than 600
341 nucleotides, merged and non-merged sequences were conserved for downstream bioinformatic steps.
342 Reads were then quality filtered (maximum ee of 1, minimum length of 150, 50 Ns allowed) with
343 Vsearch. Because the bait set contained baits for other genes for other projects (i.e. 16S, NAD1, NAD4,
344 NAD5, CYTB, and ATP6) and because their presence in the bait set can alter downstream results, the
345 reads corresponding to these genes were removed from the quality filtered reads in all downstream
346 analysis. They were recovered with a blastn on a reference database (blastn-short, e-value threshold of
347 1E-10) containing the complete sequence of NAD1, NAD4, NAD5, CYTB and ATP6 genes of 8 species
348 (*Asellus aquaticus* ADA69754.1, *Daphnia pulex* AAD33231.1, *Dinocras cephalotes* AGZ03516.1,
349 *Gammarus fossarum* YP_009379680.1, *Physella acuta* YP_008994230.1, *Radix balthica* HQ330989.1,
350 *Sericostoma personatum* AJR19241.1, *Thremma gallicum* AJR19254.1), the 16S of the 10 species MC

351 (Table S1) and the 16S corresponding to the 52 taxa MC (downloaded from NCBI). The remaining reads
352 were assigned to species using the 10 species reference database or the 52 taxa reference database using
353 BLAST algorithm (Camacho et al., 2009). Only alignments with an e-value below 1E-10, a query cover
354 over 250 and an identity over 97 were used for taxonomic assignment. As for amplicon, filtered reads
355 were also compared to a complete COI protein reference database to estimate the total number of COI
356 reads and to a protein database to estimate the rate and origin of contaminations.

357

358 *6. Capture efficacy*

359 Capture efficacy (sensus (Cha & Thilly, 1993)) was evaluated by measuring the percentage of
360 targeted reads for each capture library (i.e. capture specificity) and the X-fold enrichment (i.e. capture
361 efficiency) as in Maggia et al. (2017). The percentage of targeted reads is the ratio of the number of
362 reads assigned with diamond on the COI protein database or with blastn to a 16S, NAD1, NAD4, NAD5,
363 CYTB and ATP6 nucleotide reference database to the number of quality filtered reads. To estimate the
364 X-fold enrichment, four samples from the 10 species MC were sequenced without any enrichment. The
365 X-fold enrichment was calculated using the ratio of the percentage of targeted reads from the capture
366 library to the percentage of targeted reads from the enrichment-free library.

367

368 *7. Species detection and initial biomass recovery*

369 To compensate for sequencing effort as well as total biomass variation among samples, read count
370 assigned to species and biomass (mg) were transformed in proportion of reads (i.e. ratio of the number
371 of COI reads assigned to the species to the total number of COI reads assigned to the COI mock
372 community reference database) and proportion of biomass (i.e. ratio of the biomass of the species to the
373 total biomass of the MC), respectively.

374

375 *Species detection.* A species was considered present in a sample when represented by at least one
376 read for the capture sequencing pipeline and one MOTU for the amplicon sequencing pipeline. We

377 calculated the detection sensitivity (S) which measures the number of detection success to the total
378 number of trials ($S = \frac{\text{totalnumberofspeciesdetected}}{\text{numberofsamples*numberoftaxa}}$).

379

380 *Biomass recovery.* For the bulk DNA and etDNA of the 10 species MC, logistic models were built
381 to investigate the relationship between read proportion and biomass proportion. For each DNA template
382 and enrichment method, we compared sets of mixed effects models (i.e. no fixed effect and biomass as
383 fixed effect with no random effect, random intercept, random intercept and slope by species or samples)
384 using Aikake's Information Criterion (AIC, (Burnham & Anderson, 2002)). We also summarized to
385 what extent each method was able to predict species biomass by comparing the observed read
386 proportions to the expected read proportions where read proportion perfectly predict biomass proportion
387 (i.e. a $y=x$ relationship). To this aim, we calculated the mean absolute error (MAE) for each method as
388 follow: $\frac{1}{n} \sum |observed - expected|$ (Willmott & Matsuura, 2005). The lower the MAE is, the closer
389 the observed read proportions are to the expected read proportions.

390

391 All statistical analyses were conducted with R (R Core team, 2018). All logistic models relating
392 biomass proportion to read proportion were fitted using lme4 package (glmer functions, with a Logit
393 link and a Binomial family, (Bates, Mächler, Bolker, & Walker, 2014)).

394

395 RESULTS

396 *1. Sequencing results*

397 For bulk DNA, amplicon sequencing produced 60,337 to 365,331 raw reads and capture sequencing
398 165,089 to 857,040 raw reads per sample. For etDNA, amplicon sequencing produced 209,425 to
399 365,331 raw reads for Fwh1 primer pair of extraction I, 219,124 to 314,441 raw reads for Fwh1 primer
400 pair of extraction II, 172,036 to 280,065 raw reads for BF2/BR2 primer pair of extraction II and capture
401 sequencing 217,240 to 911,484 raw reads (Table 2). In both approaches, the capture sequencing effort

402 was higher but a higher proportion of reads was discarded through the bioinformatic pipeline,
403 particularly for etDNA (Table 2).

404

405 2. Capture efficacy

406 The percentage of targeted reads after capture enrichment (i.e. capture specificity) ranged from 16.24
407 to 86.66% (mean \pm se: 63.82% \pm 24.83) for bulk DNA and from 19.52 to 40.79% (mean \pm se: 27.18 \pm
408 8.39) for etDNA (Table 2, Table S5). For bulk DNA, the 52 taxa MC showed higher and more
409 homogeneous capture specificity (mean \pm se: 80.21 \pm 5.09) in comparison with bulk DNA of the 10
410 species MC (mean 43.34 \pm 24.52). The average X-fold enrichment (i.e. capture efficiency) was 974
411 (range: 214-2470) meaning that on average 974-fold more targeted reads were sequenced with capture
412 enrichment than without any enrichment.

413 The percentage of COI reads that were assigned to a MC species was very heterogeneous among
414 DNA template and enrichment methods. Amplicons on bulk DNA gave the best results (mean: 96.45%
415 COI assignment), followed by capture on bulk (62.33 %), amplicon on etDNA (46.36 %) and finally
416 capture on etDNA (33.71 %). Reads that do not match to the reference MC COI database can have
417 multiple origins including contamination from other organisms. The majority of reads were assigned to
418 protostomians for all but one experiment – capture enrichment on etDNA – where half of the reads
419 belonged to eubacteria and the other half to protostomians (Table 2, Figure S2).

420

421 3. PCR and capture enrichment for bulk DNA

422 *Species detection.* Capture enrichment detected systematically more species ($S=0.96$) than PCR
423 ($S=0.68$) among the 10 species MC (Figure 2). In PCR libraries, three species belonging to Gastropoda
424 (*Ancylus fluviatilis* and *Physella acuta*) and Amphipoda (*Gammarus fossarum*) were never detected.
425 With capture enrichment, every species was detected in 5 out of 8 samples (Figure 2). In the 52 taxa
426 MC, we found the same pattern, with capture enrichment ($S=0.96$) detecting more species than PCR
427 enrichment ($S=0.93$) (Figure 2). Each enrichment method eventually failed to detect a small but different
428 set of taxa.

429

430 *Biomass recovery.* Whatever the enrichment method, following the AIC, the best mixed model to
431 predict initial species biomass using species read proportion was the model combining biomass as a
432 fixed effect and species as a random effect on the intercept and slope (Table 3). This suggests that the
433 mean read proportion between species is different independently of their biomass and that the
434 relationship between biomass proportion and read proportion also differs among species (Figure 3).
435 Using this type of mixed models, a significant relationship between biomass and read proportion was
436 found for both enrichment methods (PCR: p-value=0.024; capture: p-value=0.001, Figure 3). The mean
437 absolute errors between observed and expected read proportion if biomass proportions were to be
438 perfectly translated into read proportions (i.e. $y=x$ relationship) were higher for PCR (MAE=0.11) than
439 for capture (MAE=0.056). Under a scenario where there is no relationship between biomass and read
440 proportions and where each species contribute to the same read proportion independently of its biomass
441 (i.e. 1/10 read proportion), the MAE would be of 0.07, again highlighting that the PCR enrichment step
442 wiped out most of the biomass signal. While absolute biomass variations may be lost, we also tested if
443 biomass ranks could be recovered using Spearman's rank correlation coefficient. Again PCR enrichment
444 performed poorly compared to capture enrichment (average Spearman rho 0.53 and 0.67 for PCR and
445 capture enrichment, respectively).

446

447 4. *etDNA performance*

448 As for bulk DNA, capture enrichment detected more species ($S=0.97$) than PCR ($S=0.85$) (Figure 4,
449 extraction I) on the etDNA template. Concerning PCR, the primer pair Fwh1 performs better ($S=0.86$)
450 than the primer pair BF2/BR2 ($S=0.7$) on etDNA (Figure 4, extraction II). This was expected as the long
451 BF2/BR2 fragment may be difficult to amplify on degraded DNA. Nevertheless the two primers pairs
452 also failed to detect different species probably indicating a primer bias rather than a degradation
453 problem. To confirm this, we tested if the sample composition was mostly driven by DNA template and
454 enrichment method using a Principal Coordinate Analysis (PCoA). Samples clustered by enrichment
455 method and PCR primers but not by DNA template (Figure 5). Therefore, in this experiment, the DNA

456 template had little to no impact compared to the enrichment method and primers. Concerning initial
457 biomass recovery, the same models as for bulk DNA were selected for etDNA using the AIC (Table 3).
458 However, contrary to bulk DNA, for both enrichment methods, no significant relationship between
459 biomass proportion and read proportion was found (PCR: p-value=0.37; capture: p-value=0.58). The
460 difficulty to infer biomass from read proportion using etDNA was also supported by high MAE values
461 (0.102 and 0.104 for PCR and capture enrichment, respectively) and low rank correlations (average
462 Spearman rho= 0.40 and 0.28 for PCR and capture enrichment, respectively).

463

464 DISCUSSION

465 *Capture versus PCR enrichment for metabarcoding*

466 Capture enrichment consistently led to higher species detection rates compared to PCR enrichment
467 whatever the DNA template or sample taxonomic complexity. Using capture in the 10 species MC, only
468 one species was missed in one third of the samples where it was at low abundance (dry biomass < 1
469 mg). In the 52 taxa MC, capture enrichment also globally performed better than PCR enrichment in
470 terms of species detection (20 against 35 failed detections, respectively). Both methods also missed
471 different sets of taxa. In PCR enrichment, the commonly missed taxa systematically presented 1 or 2
472 mismatches with the primer pair (Figure S1) reinforcing the view that primer mismatches are the main
473 cause for species non-detection (Elbrecht et al., 2017). Regarding capture enrichment, 8 taxa had no bait
474 designed for them (i.e. Ceratopogonidae, Blephariceridae, Dicranota, Simuliidae, Tipulidae,
475 Trombidiformes, Dugesia and *Daphnia pulex*) but only two taxa (*Daphnia pulex* and Dugesia) were
476 almost systematically missed. These two taxa have COI haplotypes that are at a divergence of at least
477 15% to any bait whereas the 6 detected taxa have haplotypes that are closer to baits originally designed
478 for other taxa (Figure S1). The divergence to the oligonucleotide has already been described as an
479 important parameter for species detection (Liu et al, 2015; Portik et al, 2016; Vallender et al, 2011; van
480 der Valk et al, 2017). In particular, Liu et al (2015) using a pool of 49 taxa found a lower enrichment
481 efficiency when the divergence to the bait was higher than 20%. Portik et al (2016), this time working
482 on a multi-loci capture data set, found a negative linear relationship between the probability of

483 sequencing a locus and the divergence to the bait. This translated into a sharp decline from 60% to 20%
484 chance of sequencing when the divergence increased from 10% to 20%. Altogether, for both enrichment
485 methods, the divergence to the oligonucleotide (i.e. primer or bait) appears to be a determinant factor
486 for species detection. Nevertheless, capture enrichment is much more robust to this divergence issue as
487 it combines three characteristics that PCR enrichment misses. First, capture enrichment is less sensitive
488 to mismatches (18 mismatches or 15% divergence) than PCR (2 mismatches or 5% divergence). Second,
489 because thousands of different baits are designed for a given 120-base region, the probability to
490 encounter mismatches is reduced with capture. Finally, several 120-base regions (5 in this study leading
491 to a total cover of 600 bases) are targeted, increasing the probability to design at least one functional
492 bait for a given species even if this species was not in the database used to design the baits. Alternatively
493 species detection can be improved in traditional PCR-based metabarcoding by targeting multiple DNA
494 fragment (e.g. (Drummond et al., 2015; Zhang, Chain, Abbott, & Cristescu, 2018)). This is also likely
495 to be true but easier to implement with capture enrichment where baits for different loci can be used
496 simultaneously (e.g. Liu et al, 2016). The ability to detect more robustly known and unknown
497 biodiversity using a single enrichment step is a pivotal step towards a better understanding of ecosystem
498 function and structure.

499

500 Capture enrichment also yields read proportions that are better predictors of species relative
501 biomasses than PCR enrichment for bulk DNA. Both enrichment methods presented a positive
502 relationship between relative biomass proportion and read proportion but this relationship was closer to
503 a linear $y=x$ relationship for capture enrichment. That said, a random species effect still remained in the
504 capture enrichment model. Its origin could be attributed to a bait bias similar to the primer bias
505 encountered in PCR enrichment but it can also be explained by other factors. Indeed, using a
506 mitochondrial marker to reconstruct an entire dry biomass community is eventually doomed to fail.
507 First, the amount of mitochondrial DNA belonging to a species in a sample will vary as a function of
508 the average number of mitochondrial genomes per cell which is likely to vary extensively among
509 species, life stages and tissues (Rooney et al, 2015). Wilcox et al. (2018) found that total initial species

510 DNA abundances could be recovered using capture enrichment if a correction was applied to normalize
511 for variation in the number of mitochondrial genome per cell. Second, dry biomass proportions are a
512 very rough estimator of cell numbers. Thus, variability in cell number per biomass and in mitochondria
513 per cell combines to blur the biomass/read proportion relationship. A similar mock community
514 experiment but with precise knowledge of the number of mitochondrial genome per species in a sample
515 is needed to better test for the existence of bait biases. One can argue that even if capture enrichment
516 yields good estimates of each species mitochondrial genome proportion in a sample, this estimate will
517 still be a poor biomass or abundance proxy. Interestingly, here we found that read proportions from gene
518 capture enrichment already provides biomass rank estimates which in many ecological contexts, for
519 instance in bioassessment (Elbrecht et al., 2017), would already be informative and may be sufficient.

520

521 *Ethanol DNA as a fast alternative to bulk DNA for species detection*

522 In agreement with Martins et al. (2019), our results demonstrate the potential of etDNA to replace
523 bulk DNA for macrozoobenthos samples if only species occurrence data is required. Species detection
524 was not determined by DNA template but by the enrichment method and the primer pair used. Hence,
525 species detection with etDNA was equivalent to species detection with bulk DNA. Nonetheless, no
526 relationship between read proportion and biomass proportion was highlighted for etDNA. For biomass
527 recovery to work with etDNA, each species should release a quantity of DNA in the ethanol
528 proportionally to its biomass. However, the release of DNA in the ethanol might differ among
529 development stages or species. When put alive in ethanol, some individuals regurgitate and release a lot
530 of their DNA (Anderson et al., 2013). For example, in the class Gastropoda, one sub-class has an
531 operculum which isolates them from the environment. When put in ethanol, these gastropods close their
532 shell and the quantity of DNA released will be lower than for other gastropods without operculum. Also,
533 after ethanol fixation, depending on the presence of a shell or a thick exoskeleton, the amount of released
534 DNA during the soaking period may also be extremely variable across taxa. In conclusion, while DNA
535 from ethanol offers a fast and non-destructive way to identify the species in a sample, differences in the
536 way species release their DNA in ethanol seem to prevent the use of ethanol for quantitative estimates.

537 Compared to bulk DNA, DNA extracted from preservative ethanol is more fragmented and much less
538 concentrated. In addition to environmental contaminants such as the bacterial contaminants observed in
539 this study, this template may also be more susceptible to reagent contamination and cross-contamination.
540 The adoption of this alternative template may therefore need stricter sampling, sample handling and
541 DNA extraction protocols than the one regularly used in community ecology laboratories.

542

543 *Optimising capture enrichment for metabarcoding*

544 Despite capture enrichment already delivers better results than PCR enrichment in terms of species
545 detection or biomass prediction, the efficiency and specificity of this emerging approach can still be
546 optimized for metabarcoding purposes. Although capture enrichment was efficient to enrich the COI
547 sequences of the macrozoobenthos present in our samples with an enrichment of at least 214-fold
548 compared to non-enriched libraries, the percentage of targeted reads was highly heterogeneous and low
549 for some samples. The specificity of the capture enrichment was particularly low when the diversity of
550 the sample was low (43% of targeted reads for the 10 species MC compared to 80% for the 52 taxa MC).
551 How the taxonomic diversity of a sample influences the efficiency of capture enrichment remains
552 unexplained and warrants further experiments where the impact of a gradually increasing sample
553 diversity is tested while controlling for other factors such as the extraction protocol or the overall
554 phylogenetic diversity.

555 The specificity of capture enrichment estimated by the number of COI reads that matched a taxa
556 included in the MC was systematically lower than with PCR enrichment (about 40% of the COI reads
557 did not match a MC taxa). When looking at the coarse taxonomic composition of the COI reads
558 sequenced after capture enrichment, we found that for bulk DNA, most of the reads (> 95%) were
559 assigned to protostomians ruling out the possibility that bacterial or plant COI contaminations alone are
560 reducing capture specificity. DNA from other protostomians that interact with the taxa included in the
561 MC (e.g. preys or parasites) may contaminate the samples and reduce capture specificity. Nevertheless,
562 we would expect the same to happen with PCR enrichment which is not the case. Alternatively, Li,
563 Schroeder, Ko, and Stoneking (2012) found that the robustness of capture to mismatches is also a

564 drawback when targeting mitochondrial genes: in addition to capture the targeted mitochondrial gene,
565 it also captures nuclear copies of this gene (NUMTs). The number of NUMT loci per genome is highly
566 variable across taxa but can easily reach hundreds to thousands of loci (Hazkani-Covo, Zeller, & Martin,
567 2010). In some species there might therefore be more NUMTs in a cell than there are mitochondrial
568 genomes. Given that baits can capture divergent loci and that there is many different baits, NUMTs may
569 represent a significant burden in mitochondrial capture enrichment. Finally, capture enrichment also
570 increase the sequencing coverage of the flanking regions of the targeted sequences (Portik et al, 2016).
571 With the pipeline used in this study, the reads containing more than 41% of a flanking region will not
572 be classified as a read coming from a targeted region. The lower specificity of capture enrichment
573 observed in our samples, but not observed with PCR enrichment, might therefore be explained by a
574 combined effect of all these factors.

575 For capture enrichment on etDNA, we estimated that almost half of the reads belonged to eubacteria
576 and the other half to protostomians. This is likely to be linked to the regurgitation of gastric microbiota
577 and the release of epidermic microbiota by some species. By this process, bacterial DNA may actually
578 dominate the ethanol DNA. In this experiment, a lower hybridization temperature was used for etDNA
579 (55°C) than for bulk DNA (65°C) in the hope of capturing more fragmented DNA. Retrospectively, this
580 lower temperature is likely to have decreased the specificity of the capture enrichment and increased the
581 representation of bacterial DNA. Indeed, the specificity of capture enrichment is higher for degraded
582 DNA and samples with a lot of contaminants when a higher hybridization temperature is used (Paijmans
583 et al., 2016). Noteworthy, 95% of the bacterial reads did not share any similarity with the mitochondrial
584 COI, again reinforcing the idea that the large amount of bacterial reads in etDNA samples is probably
585 the result of the combination of non-stringent hybridization conditions and bacterial DNA dominance
586 in etDNA.

587 As found in this study and others (e.g. Liu et al., 2016; Mariac et al., 2018; Portik et al, 2015), the
588 divergence between the targeted DNA and the baits is a critical factor in capture enrichment. As such,
589 to recover all taxa in a community, the bait design must be optimized to reduce this divergence. The
590 priority is thus to design baits using a reference database that is exhaustive. While reference DNA

591 barcodes are available for most species for some groups, this is far from being the case for many groups
592 (e.g. (Sonet et al., 2013)). Knowing the divergence threshold of 15% (this study) over which species are
593 not captured and detected, alternative strategies could be deployed to barcode key missing taxa in the
594 reference databases. Indeed, a very limited set of baits is sufficient to represent a whole family or a
595 genus if the intra-group genetic diversity is lower than the divergence threshold (Mariac et al., 2018).
596 This diversity is heterogeneous between taxa group, for example the maximum intra-family COI
597 divergence in Gammaridea (Amphipoda) is 30% but is only 17% in Chloroperlidae (Plecoptera).
598 Therefore, prior investigations are needed to establish the diversity of each group and the number of
599 taxa that have to be described before a robust set of baits can be designed. Another alternative would be
600 to design baits from a set of representative COI sequences and mutate them according to a given
601 divergence to obtain a set of baits that can hybridize to most sequences. Such baits would permit to
602 capture non-barcoded or even new species but could also reduce capture specificity by capturing
603 untargeted sequences such as bacterial genes or NUMTs.

604 In addition to the bait design strategy, other capture enrichment parameters may need further
605 optimization for a metabarcoding application. The hybridization temperature is known to be a major
606 factor controlling capture efficacy (Li et al., 2013). However, Paijmans et al. (2016) compared different
607 hybridization temperatures but unexpectedly did not find any interaction between hybridization
608 temperature and distance to the bait. But Paijmans et al. (2016) used a pool of 21 feline species, with all
609 the probes designed using a single reference species. Therefore their results probably need to be tested
610 on a data-set that is more similar to a metabarcoding data-set, with more divergent taxa and probes
611 designed from a large database, before the interaction between hybridization temperature and distance
612 to the bait can be completely dismissed. The number of pooled libraries in a single hybridization reaction
613 is another potentially important parameter. While pooling many samples in a single capture reaction
614 significantly cuts the hybridization budget, Portik et al. (2016) showed that it significantly reduces the
615 complexity of the sequencing libraries. In a metabarcoding context, this may result in a lower species
616 discovery rate in disfavor of the rare species. In the same vein, the number of capture rounds is a
617 substantial parameter increasing capture efficacy but also the hybridization budget (Li et al., 2013;

618 Mariac et al., 2018; Templeton et al., 2013; van der Valk et al., 2017). With a second round of capture,
619 Mariac et al. (2018) increased capture specificity from 0.57% to 70.5% and capture efficiency by 40
620 times but with the penalty of doubling hybridization costs. After sequencing capture enriched
621 metabarcoding libraries, another challenge is the bioinformatic treatment of the reads. Indeed,
622 processing capture data targeting one or few DNA barcodes for many species is considerably different
623 than assembling multiple targeted genes for a single organism. To our knowledge, while there is several
624 assembly pipelines that were developed for the latter case (co-assembly of different loci from a single
625 species, (e.g. (Allio et al., 2019))), there is no available bioinformatic pipelines for the former.
626 Assembling metabarcoding capture reads is challenging as it requires the independent assembly of
627 orthologous sequences belonging to different species which may be more or less divergent and more or
628 less polymorphic. An alternative to assembly is to directly map the capture reads on a reference database
629 as done in this study. While functional, this solution is sub-optimal because capture reads do not start
630 and end at the same position as in amplicon sequencing. This complicates the read clustering and
631 taxonomic assignment, with many capture reads probably lost as they only partially map to the reference
632 library. While the sequencing of the flanking regions of targeted region would probably be highly
633 valuable for taxonomic assignment, as long as there is no dedicated bioinformatic solution, it
634 paradoxically complicates the processing and analysis of capture reads. Finally, a comparison of these
635 mock community results with field samples in which the diversity is much higher and where the template
636 is also much more complex is warranted to confirm the potential of capture enrichment in community
637 ecology.

638

639 CONCLUSION

640 Capture enrichment is a promising alternative to PCR enrichment for metabarcoding. Its main
641 advantage is to provide better species detection thanks to its robustness to mismatches. At this point,
642 while performing much better than PCR enrichment, absolute biomass reconstruction is not applicable
643 without species mitogenome copy number correction and is therefore not tractable for community
644 studies where hundreds of species can be encountered. Yet, biomass rank ordination appears to be robust

645 and could be used for ecological purpose. Albeit more bacterial contamination, the use of etDNA
646 coupled to capture enrichment presents interesting compromises. It saves a lot of time by ignoring the
647 organism picking step and permits to save organisms for further analyses. However, etDNA should be
648 used for studies where quantitative information is not required. Finally, these promising results call for
649 additional testing of the bait design strategy and hybridization reaction parameters, as well as testing
650 capture enrichment on complex field samples.

651

652 ACKNOWLEDGEMENTS

653 We thank Maxence Forcellini, Bertrand Launay and Guillaume Le Goff for morphological
654 macrozoobenthos identification and fieldwork assistance. We thank Clémentine François and Florian
655 Leese for advices and assistance in bioinformatic analysis and bait design. We thank the three
656 anonymous reviewers for their helpful and relevant comments on the manuscript. This study was
657 supported by the French Biodiversity Agency (AFB) “Grant 26: headwater biodiversity dynamics: a
658 molecular perspective”, and the CNRS Mission pour les Initiatives Transverses et Interdisciplinaires
659 (project XLIFE CAPTAS).

660

661 REFERENCES

- 662 Abdelfattah, A., Malacrinò, A., Wisniewski, M., Cacciola, S. O., & Schena, L. (2018). Metabarcoding:
663 A powerful tool to investigate microbial communities and shape future plant protection
664 strategies. *Biological Control*, 120, 1-10. doi:10.1016/j.biocontrol.2017.07.009
- 665 Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., . . . Gnirke, A. (2011).
666 Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome*
667 *Biology*, 12, R18. doi:10.1186/gb-2011-12-2-r18
- 668 Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2019).
669 doi:10.1101/685412

- 670 Anderson, J. T., Zilli, F. L., Montalto, L., Marchese, M. R., McKinney, M., & Park, Y.-L. (2013).
671 Sampling and Processing Aquatic and Terrestrial Invertebrates in Wetlands. In *Wetland*
672 *Techniques* (Vol. 1, pp. 143-195).
- 673 Arif, I. A., Khan, H. A., Al Sadoon, M., & Shobrak, M. (2011). Limited efficiency of universal mini-
674 barcode primers for DNA amplification from desert reptiles, birds and mammals. *Genetics and*
675 *Molecular Research*, *10*, 3559-3564. doi:10.4238/2011.October.31.3
- 676 Aukema, B. & Rieger C. (1995) Catalogue of the Heteroptera of the Palaearctic Region. Vol. 1 :
677 Enicocephalomorpha, Dipsocoromorpha, Nepomorpha, Gerromorpha and Leptopodomorpha.
678 *Netherlands Entomological Society, Wageningen, The Netherlands*. 222 p.
- 679 Aukema, B., Rieger, C. & Rabitsch, W. (2013) Catalogue of the Heteroptera of the Palaearctic Region.
680 Vol. 6 : Supplement. *Netherlands Entomological Society, Wageningen, The Netherlands*. 629
681 p.
- 682 Avila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V.,
683 Rasmussen, M., . . . Gilbert, M. T. (2011). Application and comparison of large-scale solution-
684 based DNA capture-enrichment methods on ancient DNA. *Sci Rep*, *1*, 74.
685 doi:10.1038/srep00074
- 686 Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment
687 made possible by next-generation DNA sequencing. *Molecular Ecology*, *21*, 2039-2044.
688 doi:10.1111/j.1365-294X.2012.05519.x
- 689 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using
690 lme4. doi:10.18637/jss.v067.i01
- 691 Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., & Kausrud, H. (2010). ITS as an
692 environmental DNA barcode for fungi: An in silico approach reveals potential PCR biases.
693 *BMC Microbiology*, *10*, 1-9. doi:10.1186/1471-2180-10-189
- 694 Bista, I., Walsh, K., Zhou, X., Seymour, M., Christmas, M., Bradley, D., . . . Creer, S. (2018).
695 Performance of amplicon and shotgun sequencing for accurate biomass estimation in

- 696 invertebrate community samples. *Molecular Ecology Resources*, 18, 1020-1034.
697 doi:10.1111/1755-0998.12888
- 698 Bortolus, A. (2008). Error Cascades in the Biological Sciences: The Unwanted Consequences of Using
699 Bad Taxonomy in Ecology. *AMBIO: A Journal of the Human Environment*, 37, 114-118.
700 doi:10.1579/0044-7447(2008)37[114:ecitbs]2.0.co;2
- 701 Bringloe, T. T., Cottenie, K., Martin, G. K., & Adamowicz, S. J. (2016). The importance of taxonomic
702 resolution for additive beta diversity as revealed through DNA barcoding. *Genome*, 1140, 1130-
703 1140.
- 704 Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND.
705 *Nature Methods*, 12, 59-60. doi:10.1038/nmeth.3176
- 706 Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical
707 Information-Theoretic Approach (2nd ed). *Ecological Modelling*, 172, 488.
708 doi:10.1016/j.ecolmodel.2003.11.004
- 709 Calvignac-Spencer, S., Merkel, K., Kutzner, N., Kuhl, H., Boesch, C., Kappeler, P. M., . . . Leendertz,
710 F. H. (2013). Carrion fly-derived DNA as a tool for comprehensive and cost-effective
711 assessment of mammalian biodiversity. *Mol Ecol*, 22(4), 915-924. doi:10.1111/mec.12183
- 712 Camacho, C., Madden, T. L., Ma, N., Coulouris, G., Avagyan, V., Bealer, K., & Papadopoulos, J.
713 (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.
714 doi:10.1186/1471-2105-10-421
- 715 Cha, R. S., & Thilly, W. G. (1993). Specificity, efficiency, and fidelity of PCR. *Genome Res.*, 3, 18-29.
- 716 Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, K., . . . Bik, H. (2016). The
717 ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and*
718 *Evolution*, 56, 68-74. doi:10.1111/2041-210X.12574
- 719 Deiner, K., Bik, H. M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., . . .
720 Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey
721 animal and plant communities. *Mol Ecol*, 26(21), 5872-5895. doi:10.1111/mec.14350

- 722 de Moraes-Barros, N., Silva, J. A. B., & Morgante, J. S. (2011). Morphology, molecular phylogeny, and
723 taxonomic inconsistencies in the study of *Bradypus* sloths (Pilosa: Bradypodidae) *Journal of*
724 *Mammalogy*, *92*, 86-100. doi:10.1644/10-mamm-a-086.1
- 725 D'Hondt, J. L., & BEN A., R. (2009). Catalogue et clés tabulaires de détermination des Hirudinées d'eau
726 douce de la faune Française. *Bulletin de la Société zoologique de France*, *134*, 263-298.
- 727 Dowle, E. J., Pochon, X., C. Banks, J., Shearer, K., & Wood, S. A. (2016). Targeted gene enrichment
728 and high-throughput sequencing for environmental biomonitoring: a case study using
729 freshwater macroinvertebrates. *Molecular Ecology Resources*, *16*, 1240-1254.
730 doi:10.1111/1755-0998.12488
- 731 Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C. M., . . . Nelson,
732 N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment.
733 *GigaScience*, *4*. doi:10.1186/s13742-015-0086-1
- 734 Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance?
735 Testing primer bias and biomass-sequence relationships with an innovative metabarcoding
736 protocol. *PLoS ONE*, *10*, 1-16. doi:10.1371/journal.pone.0130324
- 737 Elbrecht, V., & Leese, F. (2016). Development and validation of DNA metabarcoding COI primers for
738 aquatic invertebrates using the R package "PrimerMiner". *PeerJ*, 1-23.
739 doi:10.7287/peerj.preprints.2044v1
- 740 Elbrecht, V., & Leese, F. (2017). Validation and Development of COI Metabarcoding Primers for
741 Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, *5*, 1-11.
742 doi:10.3389/fenvs.2017.00011
- 743 Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., . . . Leese, F.
744 (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects.
745 *PeerJ*, *4*, e1966. doi:10.7717/peerj.1966
- 746 Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and
747 weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream
748 monitoring. *Methods in Ecology and Evolution*, *8*, 1265-1275. doi:10.1111/2041-210X.12789

- 749 Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of
750 mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular*
751 *Marine Biology and Biotechnology*, 3(5), 294-299.
- 752 Forshaw, J. M. (2010). *Parrots of the World* (P. U. Press Ed. Vol. 70).
- 753 Frank, J. A., Wilson, B. A., Weisbaum, J. S., Sharma, S., Reich, C. I., & Olsen, G. J. (2008). Critical
754 Evaluation of Two Primers Commonly Used for Amplification of Bacterial 16S rRNA Genes.
755 *Applied and Environmental Microbiology*, 74, 2461-2470. doi:10.1128/aem.02272-07
- 756 Forster, D. W., Bull, J. K., Lenz, D., Autenrieth, M., Paijmans, J. L. A., Kraus, R. H. S., . . . Fickel, J.
757 (2018). Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel
758 species. *Mol Ecol Resour*, 18(6), 1356-1373. doi:10.1111/1755-0998.12924
- 759 Gargominy, O., Prié, V., Bichain, J. M., Cucherat, X., & Fontaine, B. (2011). Liste de référence annotée
760 des mollusques continentaux de France. *MalaCo*, 7, 307-382.
- 761 Gibson, J. F., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., & Hajibabaei, M. (2015).
762 Large-scale biomonitoring of remote and threatened ecosystems via high-throughput
763 sequencing. *PLoS ONE*, 10, e0138432. doi:10.1371/journal.pone.0138432
- 764 Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M. J. T. N., Baselga, A., & Vogler, A. P.
765 (2015). Validating the power of mitochondrial metagenomics for community ecology and
766 phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, 6, 883-894.
767 doi:10.1111/2041-210X.12376
- 768 Grand, D., & Boudot, J. P. (2007). Les libellules de France, Belgique et Luxembourg. *Biotope*
- 769 Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R., & Golding, G. B. (2016). A new way to
770 contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science
771 and biomonitoring. *Philosophical Transactions of the Royal Society of London B: Biological*
772 *Sciences*, 371, 20150330. doi:10.1098/rstb.2015.0330
- 773 Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a
774 freshwater benthic macroinvertebrate community through non-destructive environmental

- 775 barcoding of DNA from preservative ethanol. *BMC Ecology*, 12, 28. doi:10.1186/1472-6785-
776 12-28
- 777 Harper, L. R., Buxton, A. S., Rees, H. C., Bruce, K., Brys, R., Halfmaerten, D., . . . Hänfling, B. (2018).
778 Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds.
779 *Hydrobiologia*, 826(1), 25-41. doi:10.1007/s10750-018-3750-5
- 780 Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular poltergeists: mitochondrial DNA
781 copies (numts) in sequenced nuclear genomes. *PLoS Genet*, 6(2), e1000834.
782 doi:10.1371/journal.pgen.1000834
- 783 Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2006). Barcoding animal life: cytochrome c
784 oxidase subunit 1 divergences among closely related species *Proceedings of the Royal Society*
785 *of London. Series B: Biological Sciences*, 270, 96-99. doi:10.1098/rsbl.2003.0025
- 786 Henry, J. P., & Magniez, G. (1983). Introduction pratique à la systématique des organismes des eaux
787 continentales françaises-4. Crustacés Isopodes (principalement Asellotes). *Publications de la*
788 *Société Linnéenne de Lyon*. 52,, 319-357.
- 789 Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., . . . McCombie, W. R. (2007).
790 Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, 39, 1522.
791 doi:10.1038/ng.2007.42
- 792 Horn, S. (2012). Target enrichment via DNA hybridization capture. *Methods Mol Biol*, 840, 177-188.
793 doi:10.1007/978-1-61779-516-9_21
- 794 Hubert, N., Delrieu-Trottin, E., Irisson, J. O., Meyer, C., & Planes, S. (2010). Identifying coral reef fish
795 larvae through DNA barcoding: A test case with the families Acanthuridae and Holocentridae.
796 *Molecular Phylogenetics and Evolution*, 55, 1195-1203. doi:10.1016/j.ympev.2010.02.023
- 797 Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., . . . Yu, D. W. (2013).
798 Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*,
799 16, 1245-1257. doi:10.1111/ele.12162
- 800 Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Mol*
801 *Ecol*, 25(1), 185-202. doi:10.1111/mec.13304

- 802 Jusino, M. A., Banik, M. T., Palmer, J. M., Wray, A. K., Xiao, L., Pelton, E., . . . Lindner, D. L. (2019).
803 An improved method for utilizing high-throughput amplicon sequencing to determine the diets
804 of insectivorous animals. *Molecular Ecology Resources*, *19*, 176-190. doi:10.1111/1755-
805 0998.12951
- 806 Kersey, P. J., Stein, J., Zadissia, A., Yates, A., Paulini, M., Urban, M., . . . Bolser, D. M. (2017). Ensembl
807 Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids*
808 *Research*, *46*, D802-D808. doi:10.1093/nar/gkx1011
- 809 Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., . . . Zimmermann, J. (2016).
810 DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic
811 ecosystems in Europe. *Research Ideas and Outcomes*, *2*, e11321. doi:10.3897/rio.2.e11321
- 812 Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., . . . Weigand, A. M. (2018).
813 Why We Need Sustainable Networks Bridging Countries, Disciplines, Cultures and Generations
814 for Aquatic Biomonitoring 2.0: A Perspective Derived From the DNAqua-Net COST Action.
815 *Advances in Ecological Research*, *58*, 63-99. doi:10.1016/bs.aecr.2018.01.001
- 816 Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic
817 OTUs in Illumina COI metabarcoding. *PeerJ*, *5*, e3006. doi:10.7717/peerj.3006
- 818 Leys, M., Keller, I., Räsänen, K., Gattolliat, J. L., & Robinson, C. T. (2016). Distribution and population
819 genetic variation of cryptic species of the Alpine mayfly *Baetis alpinus* (Ephemeroptera:
820 Baetidae) in the Central Alps. *BMC Evolutionary Biology*, *16*, 1-15. doi:10.1186/s12862-016-
821 0643-y
- 822 Li, M., Schroeder, R., Ko, A., & Stoneking, M. (2012). Fidelity of capture-enrichment for mtDNA
823 genome sequencing: Influence of NUMTs. *Nucleic Acids Research*, *40*. doi:10.1093/nar/gks499
- 824 Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. (2013). Capturing protein-coding genes
825 across highly divergent species. *Biotechniques*, *54*(6), 321-326. doi:10.2144/000114039
- 826

- 827 Linard, B., Crampton-Platt, A., Gillett, C. P. D. T., Vogler, A. P., & Timmermans, M. J. T. N. (2015).
828 Metagenome Skimming of Insect Specimen Pools: Potential for Comparative Genomics.
829 *Genome Biology and Evolution*, 7, 1474-1489. doi:10.1093/gbe/evv086
- 830 Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., . . . Zhou, X. (2016). Mitochondrial capture enriches
831 mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular*
832 *Ecology Resources*, 16, 470-479. doi:10.1111/1755-0998.12472
- 833 Macher, J. N., Zizka, V. M. A., Weigand, A. M., & Leese, F. (2017). A simple centrifugation protocol
834 for metagenomic studies increases mitochondrial DNA yield by two orders of magnitude.
835 *Methods in Ecology and Evolution*, 9(4), 1070-1074. doi:10.1111/2041-210X.12937
- 836 Maggia, M. E., Vigouroux, Y., Renno, J. F., Duponchelle, F., Desmarais, E., & Nunez, J. (2017). DNA
837 Metabarcoding of Amazonian Ichthyoplankton Swarms. 1-14.
838 doi:10.1371/journal.pone.0170009
- 839 Mariac, C., Vigouroux, Y., Duponchelle, F., García-Dávila, C., Nunez, J., Desmarais, E., & Renno, J.
840 F. (2018). Metabarcoding by capture using a single COI probe (MCSP) to identify and quantify
841 fish species in ichthyoplankton swarms. *PLoS ONE*, 13, 1-15.
842 doi:10.1371/journal.pone.0202976
- 843 Martin, G. K., Adamowicz, S. J., & Cottenie, K. (2016). Taxonomic resolution based on DNA barcoding
844 affects environmental signal in metacommunity structure. *Freshwater Science*, 35, 701-711.
845 doi:10.1086/686260.
- 846 Martin, M. (2014). Cutadapt removes adapter sequences from high-throughput sequencing reads.
847 *EMBnet.journal*, 17, 10. doi:10.14806/ej.17.1.200
- 848 Martins, F. M. S., Galhardo, M., Filipe, A. F., Teixeira, A., Pinheiro, P., Pauperio, J., . . . Beja, P. (2019).
849 Have the cake and eat it: Optimizing nondestructive DNA metabarcoding of macroinvertebrate
850 samples for freshwater biomonitoring. *Mol Ecol Resour*, 19(4), 863-876. doi:10.1111/1755-
851 0998.13012

- 852 Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., . . . Niehuis, O. (2016).
853 BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design.
854 *Molecular biology and evolution*, *33*, 1875-1886. doi:10.1093/molbev/msw056
- 855 McCormack, J. E., Harvey, M. G., Faircloth, B. C., Crawford, N. G., Glenn, T. C., & Brumfield, R. T.
856 (2013). A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-
857 throughput sequencing. *PLoS One*, *8*(1), e54848. doi:10.1371/journal.pone.0054848
- 858 Oksanen, J., Blanchet G.F., Friendly M., Kindt R., Legendre P., McGlenn D., Minchin P.R., O'Hara R.
859 B., Simpson G. L., Solymos P., Henry M., Stevens H., Szoecs E. and Wagner H. (2019). vegan:
860 Community Ecology Package. R package version 2.5-5. [https://CRAN.R-](https://CRAN.R-project.org/package=vegan)
861 [project.org/package=vegan](https://CRAN.R-project.org/package=vegan)
- 862 Paijmans, J. L., Fickel, J., Courtiol, A., Hofreiter, M., & Forster, D. W. (2016). Impact of enrichment
863 conditions on cross-species capture of fresh and degraded DNA. *Mol Ecol Resour*, *16*(1), 42-
864 55. doi:10.1111/1755-0998.12420
- 865 Pattée, E., & Gourbault, N. (1981). Introduction pratique à la systématique des organismes des eaux
866 continentales françaises. 1 Turbellariés Tricladés Paludicoles (planaires d'eau douce).
867 *Publications de la Société Linnéenne de Lyon*, *50*, 279-304.
- 868 Phuong, M. A., & Mahardika, G. N. (2018). Targeted Sequencing of Venom Genes from Cone Snail
869 Genomes Improves Understanding of Conotoxin Molecular Evolution. *Mol Biol Evol*, *35*(5),
870 1210-1224. doi:10.1093/molbev/msy034
- 871 Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches
872 limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of
873 arthropods. *Molecular Ecology Resources*, *15*, 819-830. doi:10.1111/1755-0998.12355
- 874 Piñol, J., Senar, M. A., & Symondson, W. O. C. (2018). The choice of universal primers and the
875 characteristics of the species mixture determine when DNA metabarcoding can be quantitative.
876 *Molecular Ecology*. doi:10.1111/mec.14776
- 877 Pinto, A. J., & Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in
878 pyrosequencing datasets. *PLoS ONE*, *7*. doi:10.1371/journal.pone.0043093

- 879 Piscart, C., & Bollache, L. (2012). Crustacés amphipodes de surface: gammares d'eau douce.
880 *Association française de limnologie*. p122.
- 881 Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for
882 biodiversity analysis. *Molecular Ecology*, 27, 313-338. doi:10.1111/mec.14478
- 883 Prié, V., Puillandre, N., & Bouchet, P. (2013). Bad taxonomy can kill: molecular reevaluation of *Unio*
884 *mancus* Lamarck, 1819 (Bivalvia: Unionidae) and its accepted subspecies. *Knowledge and*
885 *Management of Aquatic Ecosystems*, 11. doi:10.1051/kmae/2013071
- 886 Queney, P. (2004). Liste taxonomique des Coléoptères 'aquatiques' de la faune de France (avec leur
887 répartition sommaire). *Le Coléoptériste*, 7, 3-27
- 888 R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for
889 Statistical Computing, Vienna.
- 890 Rognes, T., Quince, C., Nichols, B., Flouri, T., & Mahé, F. (2016). VSEARCH: a versatile open source
891 tool for metagenomics. *PeerJ*, 4, e2584. doi:10.7717/peerj.2584
- 892 Schnell, I. B., Thomsen, P. F., Wilkinson, N., Rasmussen, M., Jensen, L. R., Willerslev, E., . . . Gilbert,
893 M. T. (2012). Screening mammal biodiversity using DNA from leeches. *Curr Biol*, 22(8), R262-
894 263. doi:10.1016/j.cub.2012.02.058
- 895 Shokralla, S., Gibson, J., King, I., Baird, D., Janzen, D., Hallwachs, W., & Hajibabaei, M.
896 (2016). Environmental DNA Barcode Sequence Capture: Targeted, PCR-free Sequence Capture
897 for Biodiversity Analysis from Bulk Environmental Samples. doi:10.1101/087437
- 898 Sonet, G., Jordaens, K., Braet, Y., Bourguignon, L., Dupont, E., Backeljau, T., . . . Desmyter, S. (2013).
899 Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of
900 forensically important Diptera from Belgium and France. *ZooKeys*, 365, 307-328.
901 doi:10.3897/zookeys.365.6027
- 902 Souty-Grosset, C., Holdich, D., Noel, P., Reynolds, J. D., & Haffner, P. (2006). Atlas of crayfish in
903 Europe. *Muséum national d'Histoire naturelle*. p188
- 904 Stein, E. D., White, B. P., Mazor, R. D., Miller, P. E., & Pilgrim, E. M. (2013). Evaluating Ethanol-
905 based Sample Preservation to Facilitate Use of DNA Barcoding in Routine Freshwater

- 906 Biomonitoring Programs Using Benthic Macroinvertebrates. *PLoS ONE*, 8, 1-7.
907 doi:10.1371/journal.pone.0051273
- 908 Sweeney, B. W., Battle, J. M., Jackson, J. K., & Dapkey, T. (2011). Can DNA barcodes of stream
909 macroinvertebrates improve descriptions of community structure and water quality? *Journal of*
910 *the North American Benthological Society*, 30, 195-216. doi:10.1899/10-016.1
- 911 Tachet, H., Richoux, P., Bournard, M., & Usseglio-Polatera, P. (2010). Invertébrés d'eau douce:
912 systématique, biologie, écologie. *Paris: CNRS éditions*.
- 913 Templeton, J. E. L., P.M., B., Llamas, B., J., S., Haak, W., Cooper, A., & Austin, J. J. (2013). DNA
914 capture and next-generation sequencing can recover whole mitochondrial genomes from highly
915 degraded samples for human identification. *Investigative Genetics*, 5(26).
- 916 Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for
917 monitoring past and present biodiversity. *Biological Conservation*, 183, 4-18.
918 doi:10.1016/j.biocon.2014.11.019
- 919 Uchiyama, I., Mihara, M., Nishide, H., & Chiba, H. (2015). MBGD update 2015: Microbial genome
920 database for flexible ortholog analysis utilizing a diverse set of genomic data. *Nucleic Acids*
921 *Research*, 43, D270-D276. doi:10.1093/nar/gku1152
- 922 Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., . . . Dejean, T. (2015).
923 Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding.
924 *Molecular Ecology*, n/a-n/a. doi:10.1111/mec.13428
- 925 Vallenduuk, H. J., & Cuppen, M. J. (2004). The aquatic living caterpillars (Lepidoptera: Pyraloidea:
926 Crambidae) of Central Europe. A key to the larvae and autecology. *Lauterbornia*, 49, 1-17.
- 927 Vamos, E., Elbrecht, V., & Leese, F. (2017). Short COI markers for freshwater macroinvertebrate
928 metabarcoding. *Metabarcoding and Metagenomics*, 1, e14625. doi:10.3897/mbmg.1.14625
- 929 Van Bortel, W., Harbach, R. E., Trung, H. D., Roelants, P., Backeljau, T., & Coosemans, M. (2001).
930 Confirmation of *Anopheles varuna* in Vietnam, previously misidentified and mistargeted as the
931 malaria vector *Anopheles minimus*. *American Journal of Tropical Medicine and Hygiene*, 65,
932 729-732. doi:10.4269/ajtmh.2001.65.729

- 933 van der Valk, T., Lona Durazo, F., Dalen, L., & Guschanski, K. (2017). Whole mitochondrial genome
934 capture from faecal samples and museum-preserved specimens. *Mol Ecol Resour*, 17(6), e111-
935 e121. doi:10.1111/1755-0998.12699
- 936 Wilcox, T. M., Piggott, M. P., Young, M. K., McKelvey, K. S., Schwartz, M. K., & Zarn, K. E. (2018).
937 Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular*
938 *Ecology Resources*, 18, 1392-1401. doi:10.1111/1755-0998.12928
- 939 Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root
940 mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79-
941 82. doi:10.3354/cr00799
- 942 Yang, H., Ding, Y., Hutchins, L. N., Szatkiewicz, J., Bell, T. A., Paigen, B. J., . . . Churchill, G. A.
943 (2009). A customized and versatile high-density genotyping array for the mouse. *Nature*
944 *Methods*, 6, 663. doi:10.1038/nmeth.1359
- 945 Zerbino, D. R., Flicek, P., Juettemann, T., Zadissa, A., Lavidas, I., Achuthan, P., . . . Loveland, J. E.
946 (2017). Ensembl 2018. *Nucleic Acids Research*, 46, D754-D761. doi:10.1093/nar/gkx1098
- 947 Zhang, G. K., Chain, F. J. J., Abbott, C. L., & Cristescu, M. E. (2018). Metabarcoding using multiplexed
948 markers increases species detection in complex zooplankton communities. *Evolutionary*
949 *Applications*, 11, 1901-1914. doi:10.1111/eva.12694
- 950 Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., . . . Huang, Q. (2013). Ultra-deep sequencing
951 enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR
952 amplification. *GigaScience*, 2, 4. doi:10.1186/2047-217X-2-4
- 953 Zizka, V. M. A., Leese, F., Peinert, B., & Geiger, M. F. (2018). DNA metabarcoding from sample
954 fixative as a quick and voucher-preserving biodiversity assessment method. *Genome*, 1-41.
955 doi:10.1139/gen-2018-0048

956

957 DATA ACCESSIBILITY

958 Additional reference sequences developed for bait design are available on Genbank (Accession
959 Numbers MK584300:MK584515). COI and 16S sequences barcoded for the 10 species MC assignment

960 are available on GenBank (Accession Numbers MK584516:MK584524 for COI and
961 MK584525:MK584534 for 16S). The bait set designed for this study is available on Zenodo (Zenodo
962 (<https://doi.org/10.5281/zenodo.2581410>)). FASTQ files from PCR and capture enrichment are available
963 on NCBI (Accession Number SRA SRP188737). Scripts for bioinformatic analysis are available on
964 GitHub (<https://github.com/mailysgauthier/bioinf-cap-PCR>).

965

966 AUTHOR CONTRIBUTIONS

967 Samples of the 10 species MC were collected by MG. Methodology design was conceived by MG,
968 TL, CD, VE and TD. Laboratory work was conducted by MG, LKD, AN. Data analysis was conducted
969 by MG, TL and VE. MG and TL led the writing of the manuscript. All authors contributed to write the
970 manuscript.

971

972 TABLE AND FIGURE CAPTIONS

973 Table 1. Dry biomass in mg and number of individuals (*italic*) of freshwater invertebrate species in
974 the 10 species mock communities.

975

976 Table 2. Assessment of PCR and capture enrichment specificity on bulk and ethanol DNA. "% of
977 targeted reads": for capture enrichment only, percentage of reads that align to the COI, 16S, NAD1,
978 NAD4, NAD5, CYTB, and ATP6 reference databases, "% of COI reads": percentage of reads that align
979 to the COI protein reference database, "% of COI reads assigned to a MC species": percentage of the
980 number of reads that were successfully assigned to a species using the COI nucleotide reference
981 databases to the COI assigned reads and "% of non-protostomian reads": percentage of reads that align
982 to non-protostomian groups. Indicated values are mean number and standard deviation. For sample
983 values, see Table S4 (PCR enrichment) and Table S5 (Capture enrichment). 10 sps: 10 species MC; 52
984 taxa:52 taxa MC; Ø means that the measure cannot be calculated.

985

986

987 Table 3. Testing the link between read proportion (dependant variable) and initial biomass
988 proportion. Only relative ΔAIC to the best model are shown (i.e. $AIC_{model} - AIC_{best\ model}$). Models with
989 and without fixed effect (biomass proportion) and two random effects (species and sample) were built.
990 For each random effect, three models were built: no random effect, random intercept and random
991 intercept and slope. Best models correspond to the models with relative ΔAIC of 0. Ø means that the
992 model could not be tested.

993

994 Figure 1: Overview of the experimental design of the study. Two different mock communities (MCs)
995 were used: 10 species MCs and 52 taxa MCs. Ten species MCs were used to assess initial biomass
996 recovery and etDNA performances with PCR and capture enrichment. Both MC were used to assess
997 taxa detection. etDNA: ethanol DNA.

998

999 Figure 2 Taxa recovery performance assessed using read proportion for PCR (BF2/BR2 primer pair,
1000 left) and capture (right) enrichment and for bulk DNA of the 10 species mock communities (MCs) (top)
1001 and bulk DNA of the 52 taxa MC (bottom). Read proportions are shown for each taxa (in rows) and
1002 mock community (in column). A crossed out cell indicates no assigned read. For detailed read counts,
1003 see Tables S6 (10 species MCs) and S7 (52 taxa MCs).
1004

1005 Figure 3: Relationship between biomass proportion and read proportion for PCR (left) and capture
1006 (right) enrichment method. Logistic mixed model predictions with random intercept and slope by
1007 species are shown (solid line). The dashed red line corresponds to the expected linear relationship where
1008 one unit of read proportion equals to one unit of biomass proportion.
1009

1010 Figure 4: Taxa recovery performance assessed using read proportion for etDNA of the 10 species
1011 mock communities. Differences between PCR (top left) and capture (top right) enrichment were
1012 assessed using a first ethanol extraction (extraction I). Differences between Fwh1 (bottom left) and
1013 BF2/BR2 (bottom right) primer pairs were assessed using a second ethanol extraction (extraction II).
1014 Sequence abundances are shown for each taxa (in rows) and mock community (in column). A crossed
1015 out cell indicates no assigned read. For detailed read counts, see Tables S6.
1016

1017 Figure 5: Principal Coordinate Analysis based on dissimilarity (Jaccard distances) in read proportions
1018 of each pair of the 10 species mock community samples. Samples are classified following their DNA
1019 template (etDNA and bulk DNA) and enrichment methods (BF2/BR2 amplicon, Fwh1 amplicon and
1020 capture). etDNA1 and etDNA2 correspond to the first and second extraction of DNA from ethanol,
1021 respectively. Percentage of variance explained by each axis is shown in bracket. Ellipses are drawn with
1022 a confidence limit of 0.95. Samples are grouped by enrichment method rather than by DNA template.

1 Table 1. Dry biomass in mg and number of individuals (*italic*) of freshwater invertebrate species in
 2 the 10 species mock communities.

3

| Species | <i>Mock community</i> | | | | | | | |
|--|-----------------------|--------------------|--------------------|--------------------|---------------------|---------------------|---------------------|---------------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| <i>Chironomus riparius</i> Meigen, 1804 | 1.84 (<i>5</i>) | 1.83 (<i>5</i>) | 2.58 (<i>5</i>) | 2.42 (<i>5</i>) | 3.66 (<i>10</i>) | 4.62 (<i>11</i>) | 3.89 (<i>10</i>) | 4.44 (<i>10</i>) |
| <i>Epeorus assimilis</i> Eaton, 1885 | 102.32 (<i>5</i>) | 55.53 (<i>6</i>) | 77.89 (<i>5</i>) | 96.47 (<i>5</i>) | 62.94 (<i>4</i>) | 110.73 (<i>4</i>) | 114.80 (<i>4</i>) | 74.23 (<i>4</i>) |
| <i>Heptagenia sulphurea</i> (O.F. Müller, 1776) | 15.73 (<i>5</i>) | 20.79 (<i>5</i>) | 21.08 (<i>5</i>) | 18.02 (<i>5</i>) | 38.31 (<i>10</i>) | 28.19 (<i>10</i>) | 32.58 (<i>10</i>) | 38.37 (<i>10</i>) |
| <i>Isoperla rivulorum</i> (Pictet, 1841) | 37.97 (<i>5</i>) | 31.88 (<i>5</i>) | 36.28 (<i>5</i>) | 30.89 (<i>5</i>) | 35.11 (<i>6</i>) | 33.39 (<i>6</i>) | 36.22 (<i>6</i>) | 39.47 (<i>6</i>) |
| <i>Nemurella picteti</i> Klapálek, 1900 | 2.83 (<i>5</i>) | 2.49 (<i>5</i>) | 2.34 (<i>5</i>) | 3.60 (<i>5</i>) | 0.50 (<i>1</i>) | 0.97 (<i>1</i>) | 0.77 (<i>1</i>) | 0.56 (<i>1</i>) |
| <i>Hydropsyche siltalai</i> Doehler, 1963 | 68.42 (<i>5</i>) | 53.15 (<i>5</i>) | 53.63 (<i>5</i>) | 53.32 (<i>5</i>) | 80.36 (<i>8</i>) | 66.27 (<i>8</i>) | 86.71 (<i>8</i>) | 79.86 (<i>8</i>) |
| <i>Athripsodes aterrimus</i> (Stephens, 1836) | 6.10 (<i>6</i>) | 6.68 (<i>5</i>) | 7.87 (<i>5</i>) | 6.23 (<i>5</i>) | 9.10 (<i>6</i>) | 7.96 (<i>6</i>) | 5.39 (<i>6</i>) | 7.01 (<i>6</i>) |
| <i>Physella acuta</i> (Draparnaud, 1805) | 23.30 (<i>5</i>) | 24.85 (<i>5</i>) | 23.03 (<i>5</i>) | 31.30 (<i>5</i>) | 16.81 (<i>4</i>) | 15.22 (<i>4</i>) | 13.25 (<i>4</i>) | 20.68 (<i>4</i>) |
| <i>Ancylus fluviatilis</i> O.F. Müller, 1774 | 13.42 (<i>5</i>) | 18.16 (<i>5</i>) | 18.40 (<i>5</i>) | 15.42 (<i>5</i>) | 20.79 (<i>8</i>) | 20.47 (<i>8</i>) | 23.21 (<i>8</i>) | 15.98 (<i>8</i>) |
| <i>Gammarus fossarum</i> Koch, 1836 | 23.07 (<i>5</i>) | 20.28 (<i>5</i>) | 16.89 (<i>5</i>) | 16.72 (<i>5</i>) | 4.4 (<i>1</i>) | 4.02 (<i>1</i>) | 5.17 (<i>1</i>) | 3.76 (<i>1</i>) |

4
 5
 6

7 Table 2. Assessment of PCR and capture enrichment specificity on bulk and ethanol DNA. "% of
8 targeted reads": for capture enrichment only, percentage of reads that align to the COI, 16S, NAD1,
9 NAD4, NAD5, CYTB, and ATP6 reference databases, "% of COI reads": percentage of reads that align
10 to the COI protein reference database, "% of COI reads assigned to a MC species": percentage of the
11 number of reads that were successfully assigned to a species using the COI nucleotide reference
12 databases to the COI assigned reads and "% of non-protostomian reads": percentage of reads that align
13 to non-protostomian groups. Indicated values are mean number and standard deviation. For sample
14 values, see Table S4 (PCR enrichment) and Table S5 (Capture enrichment). 10 sps: 10 species MC; 52
15 taxa:52 taxa MC; Ø means that the measure cannot be calculated.
16
17

| bulk DNA | <i>Mock community</i> | <i>Raw reads</i> | <i>Quality filtered reads</i> | <i>% targeted reads</i> | <i>% COI reads</i> | <i>% of COI reads assigned to a MC species</i> | <i>% of non-protostomian reads (contaminants)</i> |
|---|-----------------------|-------------------|-------------------------------|-------------------------|--------------------|--|---|
| <i>PCR enrichment</i> | 10 sps | 170,026 ± 40,272 | 114,701 ± 31,160 | Ø | 49.54 ± 4.83 | 96.01 ± 3.24 | 2.77 ± 1.30 |
| | 52 taxa | 135,168 ± 40,871 | 91,935 ± 32,720 | Ø | 42.33 ± 4.61 | 96.55 ± 3.80 | 2.37 ± 1.36 |
| <i>Capture enrichment</i> | 10 sps | 493,967 ± 288,240 | 243,483 ± 86,526 | 43.34 ± 24.52 | 25.16 ± 18.81 | 61.17 ± 5.55 | 10.73 ± 3.78 |
| | 52 taxa | 553,825 ± 165,868 | 264,694 ± 75,864 | 80.21 ± 5.09 | 69.30 ± 6.51 | 63.27 ± 1.96 | 5.31 ± 1.70 |
| <i>Enrichment-free</i> | 10 sps | 921,863 ± 68,162 | 574,443 ± 43,547 | 1.19 ± 0.31 | 0.023 ± 0.028 | 26.14 ± 9.21 | 31.66 ± 1.96 |
| etDNA | <i>Mock community</i> | <i>Raw reads</i> | <i>Quality filtered reads</i> | <i>% targeted reads</i> | <i>% COI reads</i> | <i>% of COI reads assigned to a MC species</i> | <i>% of non-protostomian reads (contaminants)</i> |
| <i>PCR enrichment - Fwh1 extraction I</i> | 10 sps | 283,644 ± 52,944 | 233,856 ± 43,256 | Ø | 62.97 ± 4.47 | 46.36 ± 11.47 | 18.61 ± 12.78 |
| <i>PCR enrichment - Fwh1 extraction II</i> | 10 sps | 205,229 ± 81,671 | 169,772 ± 69,282 | Ø | 66.34 ± 5.88 | 47.72 ± 14.65 | 10.13 ± 4.12 |
| <i>PCR enrichment - BF2/BR2 extraction II</i> | 10 sps | 203,406 ± 43,493 | 123,148 ± 47,936 | Ø. | 48.97 ± 9.81 | 60.95 ± 18.88 | 28.21 ± 19.29 |
| <i>Capture Enrichment – extraction I</i> | 10 sps | 517,202 ± 234,942 | 309,637 ± 147,304 | 27.18 ± 8.39 | 19.09 ± 5.82 | 33.71 ± 10.96 | 56.81 ± 10.17 |

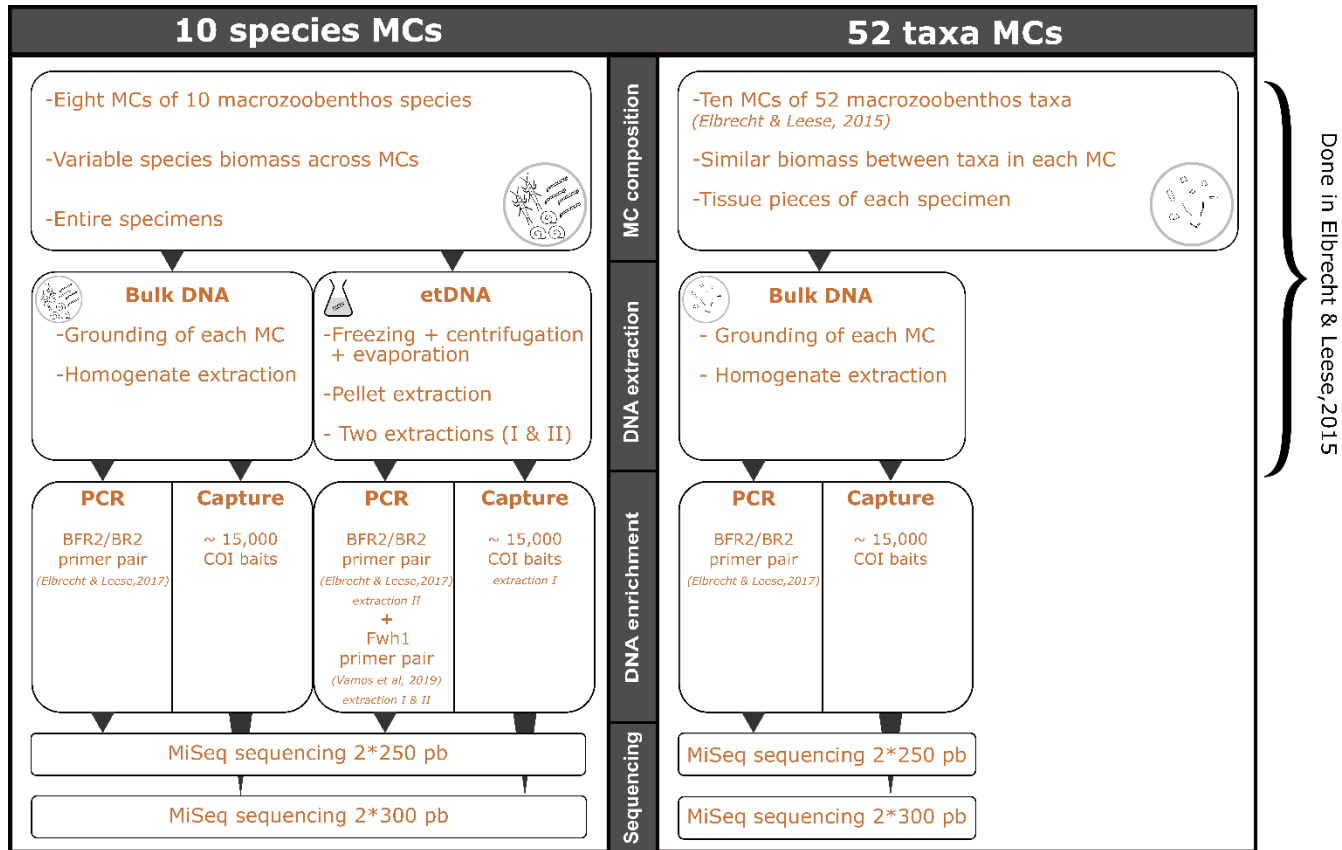
18

19 Table 3. Testing the link between read proportion (dependant variable) and initial biomass
 20 proportion. Only relative ΔAIC to the best model are shown (i.e. $AIC_{\text{model}} - AIC_{\text{best model}}$). Models with
 21 and without fixed effect (biomass proportion) and two random effects (species and sample) were built.
 22 For each random effect, three models were built: no random effect, random intercept and random
 23 intercept and slope. Best models correspond to the models with relative ΔAIC of 0. \emptyset means that the
 24 model could not be tested.
 25
 26

| bulk DNA | <i>Fixed effect</i> | <i>Random effect</i> | Random effect | | |
|---------------------------|---------------------|----------------------|----------------------|------------------|--------------------------|
| | | | <i>None</i> | <i>Intercept</i> | <i>Intercept + slope</i> |
| <i>PCR enrichment</i> | 1 | Species | 2,012,800 | 31,216 | \emptyset |
| | | Sample | | 2,012,802 | |
| | biomass proportion | Species | 537,148 | 6,674 | 0 |
| | | Sample | | 496,237 | 382,656 |
| <i>Capture enrichment</i> | 1 | Species | 320,984 | 9,223 | \emptyset |
| | | Sample | | 320,986 | |
| | biomass proportion | Species | 185,777 | 5,154 | 0 |
| | | Sample | | 185,322 | 182,666 |
| etDNA | 1 | Species | 2,602,013 | 121,099 | \emptyset |
| | | Sample | | 2,602,015 | |
| | biomass proportion | Species | 1,897,689 | 97,590 | 0 |
| | | Sample | | 1,894,974 | 1,748,461 |
| <i>Capture enrichment</i> | 1 | Species | 195,670 | 43,136 | \emptyset |
| | | Sample | | 195,672 | |
| | biomass proportion | Species | 166,761 | 26,717 | 0 |
| | | Sample | | 166,760 | 102,699 |

27
 28

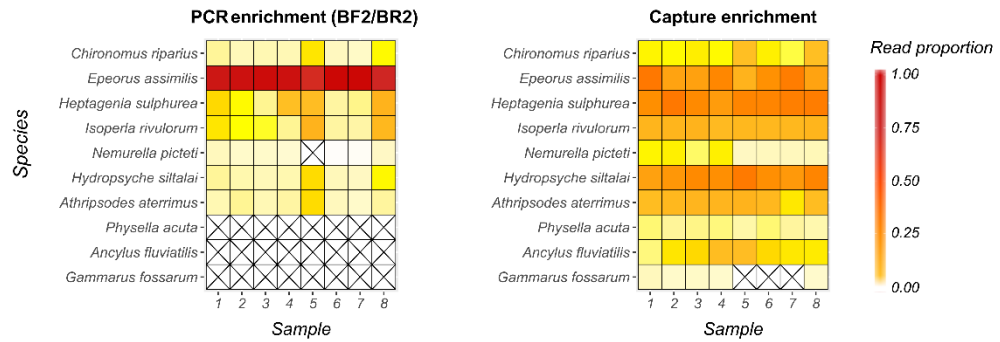
29 Figure 1: Overview of the experimental design of the study. Two different mock communities (MCs)
 30 were used: 10 species MCs and 52 taxa MCs. Ten species MCs were used to assess initial biomass
 31 recovery and eDNA performances with PCR and capture enrichment. Both MC were used to assess
 32 taxa detection. eDNA: ethanol DNA.
 33
 34



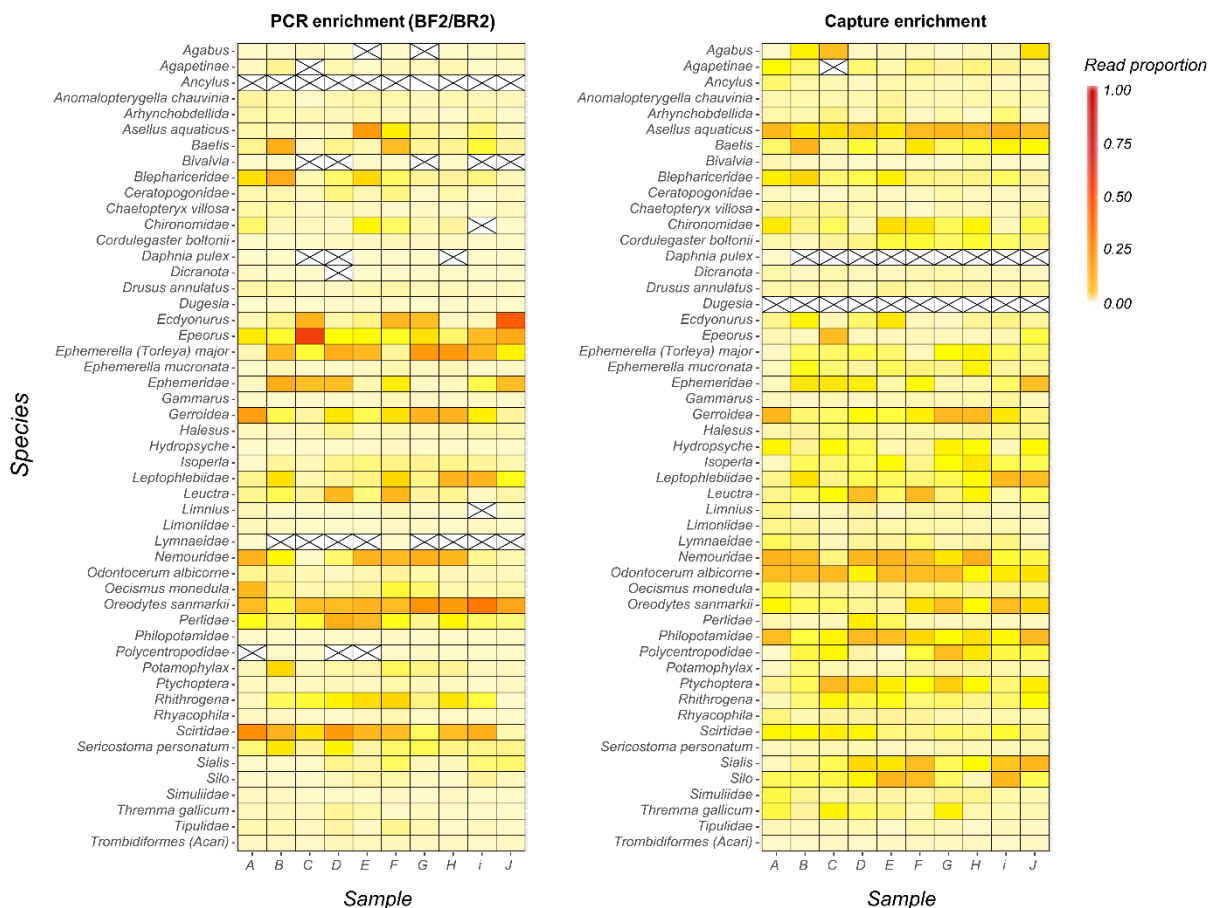
35

36 Figure 2 Taxa recovery performance assessed using read proportion for PCR (BF2/BR2 primer pair,
 37 left) and capture (right) enrichment and for bulk DNA of the 10 species mock communities (MCs) (top)
 38 and bulk DNA of the 52 taxa MC (bottom). Read proportions are shown for each taxa (in rows) and
 39 mock community (in column). A crossed out cell indicates no assigned read. For detailed read counts,
 40 see Tables S6 (10 species MCs) and S7 (52 taxa MCs).
 41
 42

10 species mock communities

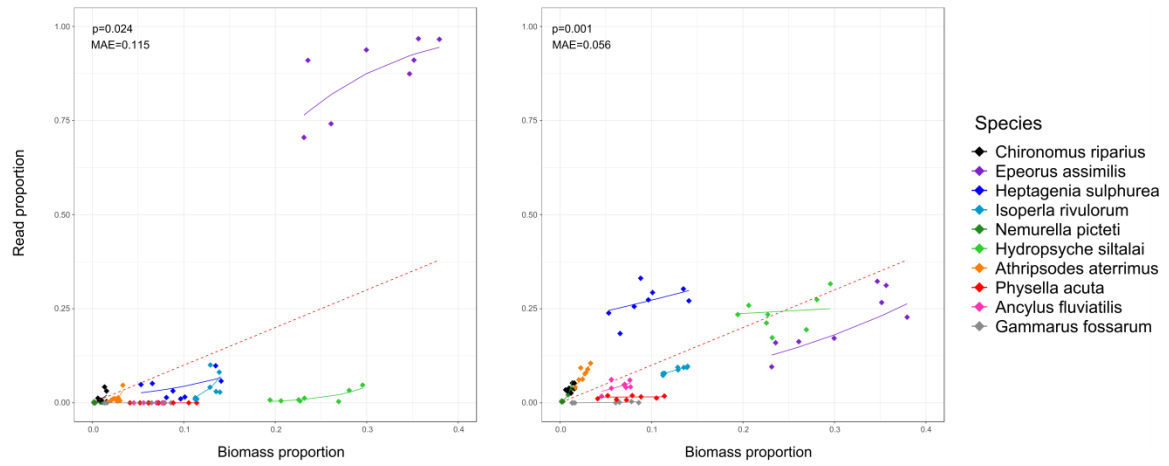


52 taxa mock communities



43

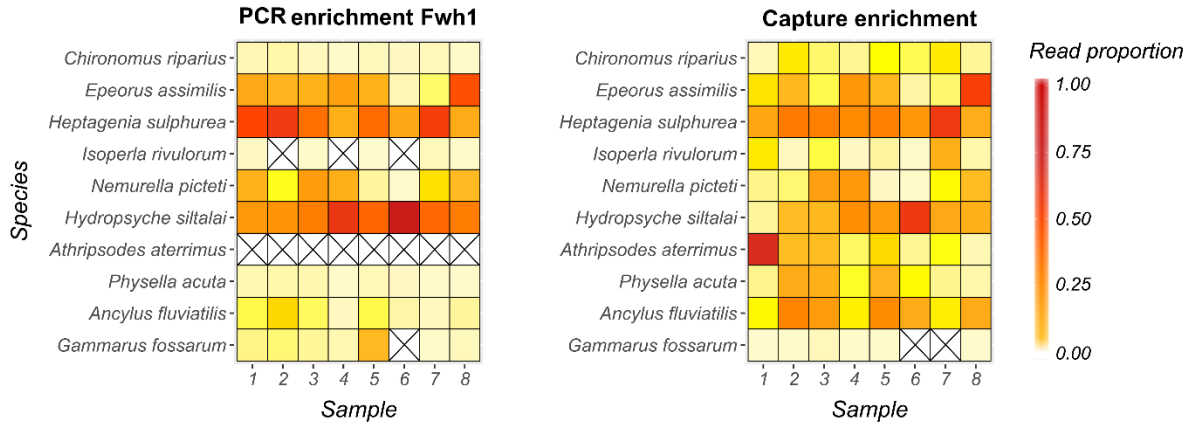
44 Figure 3: Relationship between biomass proportion and read proportion for PCR (left) and capture
45 (right) enrichment method. Logistic mixed model predictions with random intercept and slope by
46 species are shown (solid line). The dashed red line corresponds to the expected linear relationship where
47 one unit of read proportion equals to one unit of biomass proportion.
48
49



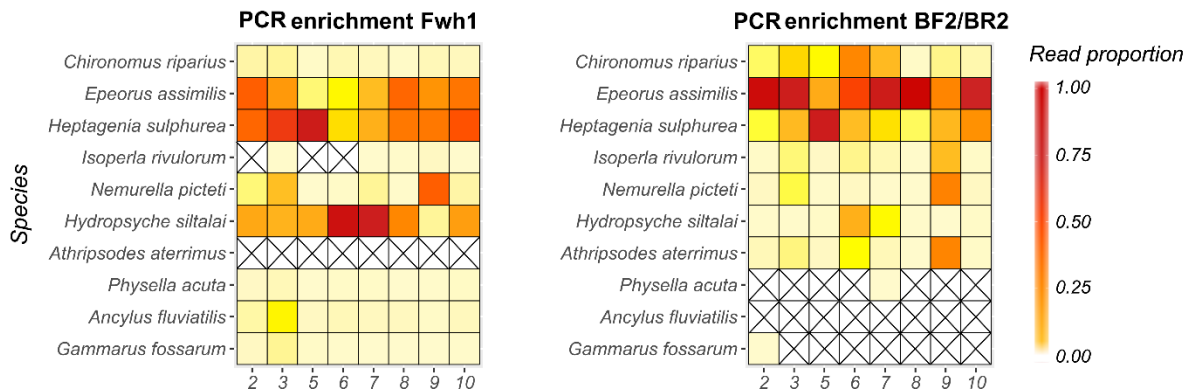
50

51 Figure 4: Taxa recovery performance assessed using read proportion for etDNA of the 10 species
 52 mock communities. Differences between PCR (top left) and capture (top right) enrichment were
 53 assessed using a first ethanol extraction (extraction I). Differences between Fwh1 (bottom left) and
 54 BF2/BR2 (bottom right) primer pairs were assessed using a second ethanol extraction (extraction II).
 55 Sequence abundances are shown for each taxa (in rows) and mock community (in column). A crossed
 56 out cell indicates no assigned read. For detailed read counts, see Tables S6.
 57

extraction I



extraction II



58

59 Figure 5: Principal Coordinate Analysis based on dissimilarity (Jaccard distances) in read proportions
60 of each pair of the 10 species mock community samples. Samples are classified following their DNA
61 template (etDNA and bulk DNA) and enrichment methods (BF2/BR2 amplicon, Fwh1 amplicon and
62 capture). etDNA1 and etDNA2 correspond to the first and second extraction of DNA from ethanol,
63 respectively. Percentage of variance explained by each axis is shown in bracket. Ellipses are drawn with
64 a confidence limit of 0.95. Samples are grouped by enrichment method rather than by DNA template.
65
66

