Manuscript to *Genome*

Priming of a DNA metabarcoding approach for species identification and inventory in marine macrobenthic communities

Claudia Hollatz[1], Barbara R. Leite[1], Jorge Lobo[1,2], Hugo Froufe[3], Conceição Egas[3], Filipe O. Costa[1]

[1]CBMA- Centre of Molecular and Environmental Biology, Biology Department, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal, [2]MARE-Marine and Environmental Sciences Centre, Department of Environmental Sciences and Engineering, Faculty of Science and Technology, Lisbon New University, 2829-516 Monte de Caparica, Portugal, [3]Next Generation Sequencing Unit, UC-Biotech-CNC, Biocant Park, 3060-197, Cantanhede, Portugal.

Corresponding author:
Claudia Hollatz
Universidade do Minho
Departamento de Biologia
Campus de Gualtar
4710-057 Braga, Portugal
Phone: (+351) 253601527
e-mail: hollatz@bio.uminho.pt

ABSTRACT

In marine and estuarine benthic communities, the inventory and estimation of species richness are often hampered by the need for broad taxonomic expertise across several phyla. The use of DNA metabarcoding has emerged as a powerful tool for the fast assessment of species composition in a diversity of ecological communities. Here we tested the amplification success of five primer sets targeting different COI-5P regions by 454-pyrosequencing to maximize the recovery of two simulated macrobenthic communities containing 21 species (SimCom1 and 2). Species identification was first performed against a compiled reference library of macrobenthic species. Reads with similarity results to reference sequences between 70 to 97% were then submitted to GenBank and BOLD to attempt the identification of concealed species in the bulk sample. The combination of at least three primer sets was able to recover more species than any primer set alone, achieving 85% of represented species in SimCom1 and 76% in SimCom2. Our approach was successful to detect low-frequency specimens as well as concealed species in the bulk sample, indicating the potential for the application of this approach on marine bioassessment and inventory, including the detection of a “hidden” biodiversity that would hardly be possible based on morphology only.

Keywords: DNA barcoding, High-throughput sequencing, Bioassessment, Marine macrobenthos

INTRODUCTION

Benthic communities are important components of aquatic ecosystems and one of the key aquatic communities contemplated in the European Union's Water Framework Directive (WFD). Under the WFD, member states have to establish a national monitoring grid to guarantee good water quality in the target aquatic ecosystems. Macrobenthic communities are frequently used as indicators of water quality, as their differing sensitivities to various environmental stressors render them a valuable tool for aquatic biomonitoring programs. Despite the generalized use of benthic surveys in environmental monitoring, there are still technical weaknesses to amend, since ecological studies require species-level identifications of specimens, which typically rely on observable morphological characteristics (Tarbelet et al. 2012). Due to their broad taxonomic diversity, it is difficult to attain accurate species compositions in macrobenthic communities (Lobo et al. 2013). The difficulties include a shortage of taxonomic experts, the use of incomplete identification keys, and the collection of degraded or damaged specimens caused by sampling techniques (Knowlton 1993). In addition, taxonomic ambiguities and uncertainties are frequently generated by the presence of complex life stages and cryptic or hidden species (Knowlton 1993; Jarman and Elliott 2000; Bickford et al. 2007).

DNA-based identification, such as through DNA barcoding (Hebert et al. 2003), has been used as an integrative tool for identifying specimens to species. However, the traditional Sanger DNA-sequencing method is not adequate for processing complex environmental samples, especially for large-scale studies (Shokralla et al. 2012). The use of high-throughput sequencing (HTS) technologies coupled with DNA barcoding (i.e. DNA metabarcoding) can overcome these limitations, providing the opportunity to sequence and identify specimens from bulk DNA isolates of whole communities (Hajibabaei et al.

2011). Recently, there have been considerable efforts towards the design of short metabarcodes within the standardized barcode region that are suitable for use in HTS platforms and able to deliver a reliable taxonomic resolution (Leray et al. 2013; Gibson et al. 2014). Also, short barcodes may provide efficient recovery of sequence information from degraded or sheared specimens (Hajibabaei et al. 2006; Meusnier et al. 2008). Although primers targeting such regions can amplify DNA from single specimens relatively efficiently, the development of a PCR amplification assay for bulk samples, representing a mixture of a diverse range of taxa, may pose a challenging task. In this context, the use of multiple primer sets are one of the key recommendations to maximize species recovery from mixed DNA templates (Hajibabaei et al. 2012; Gibson et al. 2014).

Here we present an extended application of the DNA metabarcoding methodology for routine species-level identification and inventory of marine and estuarine macrobenthic communities. To this end, we first compiled a reference library of DNA barcodes of estuarine and coastal marine invertebrates from Portugal to be used as a central framework for DNA-based specimen identification. Second, we performed *in silico* analyses to evaluate the barcode quality for DNA-based species discrimination capacity. Finally, we investigated the ability of different primer sets to amplify and detect the diversity of species present in a macrobenthic assemblage of known species composition and abundance, through the use of manually contrived simulated communities. The standard cytochrome *c* oxidase subunit I barcode (COI-5P, Hebert et al. 2003) was selected as the target region, because it provides the required species-level identification (Tang et al. 2012), and it is the most broadly represented in public reference libraries (Leray et al. 2013).

MATERIALS AND METHODS

Preparation of the simulated macrobenthic communities

Specimens used for assembling the simulated macrobenthic communities were selected from the Molecular Ecology and Biodiversity research group collection, University of Minho, Portugal. A total of 21 species were selected, to embrace the broad phylogenetic diversity within the three major phyla typically present in macrobenthic communities. The distribution of species per phyla was respectively 5% Annelida, 33% Arthropoda, and 62% Mollusca. Annelida was less represented due to the lack of available specimens in the collection at the time the study was being conducted. Two simulated communities were assembled for genomic DNA extraction, each community containing a different number of specimens per species. SimCom1 had one specimen of each species, while SimCom2 had one to five specimens of each species, containing a total of 67 specimens. (Table S1).

DNA extraction

Whole specimens were pooled and homogenized in a grinder for each simulated macrobenthic community, and the resultant slurry was incubated at 56 °C to evaporate residual ethanol, for a minimum period of two hours. The dried mixture of each homogenized simulated community was divided into 10 microtubes of 1.5 mL (about 300 mg), and the total genomic DNA was extracted using E.Z.N.A. Mollusk DNA Kit (Omega Bio-tek), following manufacturer's instructions. After extractions, aliquots of genomic DNA were pooled in a single microtube of 1.5 mL, representing each simulated community (500 µL total volume).

PCR amplification of the full and partial COI-5P barcode fragments

A preliminary assessment was conducted to test the PCR amplification success of multiple primer pairs that have already been published in addition to a newly designed primer InvF1 (Table 1). The degenerate primer InvF1 (forward) was designed by hand based on available COI-5P sequences published by Lobo et al. (2013). Fifty-six sequences belonging to 7 phyla were selected to assist the design. It is 23 bp long and located in a conserved region between 146 -168 of the start (5' end) of COI-5P. Based on the results, five primer pair combinations, which amplify different fragments within the COI barcode region, were selected for the metabarcoding tests (Table 1B). Samples were prepared for 454 pyrosequencing by PCR amplification of the COI gene with fusion primers containing the Roche-454 A and B Titanium sequencing adapters, an eight-base barcode sequence in the fusion primer A, different for SimCom 1 and SimCom2, and the forward and the reverse primers. PCR reactions were conducted for each primer pair separately in 50 µL reactions with Advantage Taq (Clontech) using 0.2 mM of each primer, 0.2 mM dNTPs, 1X polymerase mix, 6% DMSO and 30 ng of DNA template. The PCR thermal cycling conditions for each primer pair are displayed in Table 1C. A negative control reaction (no DNA template) was included in all experiments. PCR success was checked by agarose gel electrophoresis.

High-throughput 454-sequencing protocol

The amplicons were quantified by fluorimetry with PicoGreen (Invitrogen, CA, USA) and pooled at equimolar concentration. The two simulated communities were sequenced in the A direction with the GS 454 FLX Titanium chemistry in the same sequencing run, following the amplicon sequencing protocol provided by the supplier (Roche, 454 Life Sciences, Branford, CT, USA).

Data processing and analysis

A reference DNA (COI-5P) barcode library of estuarine and coastal marine invertebrates from Portugal was compiled for taxonomic identification of pyrosequencing reads generated in both simulated communities. The reference library comprises 294 barcode sequences from 245 species mostly from the three main marine phyla (Annelida, Arthropoda and Mollusca), and include all the species used in the simulated communities with the exception of *Alvania mediolittoralis*. The sequences were retrieved from the BOLD database (Ratnasingham and Hebert 2007), as well as from projects managed by the Molecular Ecology and Biodiversity research group (dx.doi.org/10.5883/DS-3150). Each species in the reference library is represented by one to four sequences, which were selected among the longest and highest quality (absence of ambiguous bases) available, as well as among sequences displaying intraspecific distance above 2%. The sequences were aligned using the Clustal W method (Thompson et al. 1994) implemented in the program MEGA v.6.0 (Tamura et al. 2013). All sequences were checked for the presence of indels, stop codons, or unusual amino acid patterns, due to manual editing or sequencing errors.

Two *in silico* tests were carried out using our DNA barcode reference library in order to evaluate the performance of different COI fragment sizes on the species-level discrimination capacity. First, the full length of the barcode region was divided into multiple fragments starting at position 658-bp of the standard barcode region, with consecutive upstream cuts of 100 bp making up five shorter fragments (158, 258, 358, 458 and 558 bp). Second, fragments corresponding to amplicons generated for four tested primer pairs (310, 313, 418, 470 bp) were analysed. The Neighbor-Joining (NJ) method was used to construct phenograms (Saitou and Nei, 1987) in the program MEGA v.6.0, using the Kimura 2-parameter (K2P) substitution model (Kimura, 1980) and pairwise deletion of

missing data. Node support was assessed through 1000 bootstrap replicates. The monophyletic clades (NJ tree) were verified in two different phases: (1) percentage of monophyletic clades with internal divergence higher than 3%; (2) percentage of different species (considering Linnaean species names) that were grouped in the same clade with maximum divergence of 3%, in which case the genetic distance among species was verified using the p-distance metric, calculated using MEGA v.6.0 program. In addition, species delimitation analyses were reassessed for the most conflicting cases using the Poisson-Tree Processes (PTP) approach (Zhang et al. 2013). The species were inferred based on the analyses of the most-supported partition of the Bayesian tree using MrBayes v.3.1.2 (Ronquist et al. 2012). The run was conducted as follows: the maximum-likelihood model employed six substitution types (nst = 6); rate variation across sites was modeled using a gamma distribution, with a proportion of the sites being invariant (rate = invgamma); Markov chain Monte Carlo (MCMC) search was run for 20,000 generations, trees were sampled every 100 generations and the first 5000 trees were discarded as burnin.

The raw pyrosequencing reads (fastq files) were processed using an automatic pipeline implemented at the Next-Generation Sequencing Unit of UC-Biotech-CNC, University of Coimbra, Portugal. In a first step, sequencing reads were assigned to the appropriate samples based on the respective barcode (short unique sequences used to label DNA in multiplexed HTS experiments). Then, reads were quality filtered to minimize the effects of random sequencing errors, by trimming sequences with average phred score lower than 15 in a window of 7 bases and by elimination of sequence reads <100 bp and sequences that contained more than two undetermined nucleotides (N). The filtered reads obtained for each community were aligned against the reference library using the Usearch 6.1 software (Edgar, 2010). Finally, sequence similarity searches with a minimum p-distance of 97% were performed against the reference library to assign a primary taxonomic

identification. This cut-off value corresponds to the universal DNA-barcoding threshold proposed by Hebert and co-workers (2003) and used in other macroinvertebrate metabarcoding studies (Carew et al. 2013; Cowart et al. 2015; Leray and Knowlton 2015).

In order to possibly identify new taxa that had no representation in the reference library, a new similarity search was conducted for all sequences that displayed similarities against the reference library below 97% and above 70%. We used BOLD Identification System (IDS) and GenBank's BLASTn for this purpose. Only matches > 97% similarity were considered for taxon identification in this analysis.

RESULTS

In silico analysis of the impact of fragment size on species discrimination ability

In the reference library, the distribution of barcode sequences across the three main marine phyla was as follows: Annelida (17.5%), Arthropoda (62%), and Mollusca (17.5%). Other phyla with minor representations (<4%) in the library were: Chordata (1.70%), Cnidaria (0.30%), Echinodermata (1%) and Nemertea (0.32%) (Figure S1A). The vast majority of the COI-5P barcodes included in the reference library were identified to species (245), the remaining ones to genus (35), or family (14) (Figure S1B).

The NJ trees showed that regardless of the fragment size, nearly all species in the reference library were separated similarly in distinct clusters, although slight differences existed in clade distances depending on the different amplicons. In the full COI-5P barcode region, 272 out of the 280 taxa (either species or genus) could be discriminated in monophyletic groups in the NJ phenogram. Different taxa that could not be clearly discriminated (showing divergence < 3%) were reassessed through PTP analyses. The results showed three additional taxa (two specimens of *Littorina obtusata* clustered with

one specimen of *Littorina saxatilis*) that could not be discriminated , ending up with 269 out of the 280 taxa dicriminated in monophyletic groups (data available upon request). When smaller fragments within the full barcode region were used to recreate NJ trees, the same topology was observed with nearly no loss in the species discrimination, the exception being the NJ tree recreated for primer D (mlCOIintF/LoboR), where *L. obtusata* and *L. saxatilis* could not be discriminated. Representative NJ phenograms are shown in Figure S2.

DNA-based species identification through HTS

A total of 24,198 454-pyrosequencing reads were generated: 12,221 from SimCom1 and 11,977 from SimCom2. Following trimming, filtering, and quality checking, 7,709 (63%) sequences for SimCom1 and 7,084 (59%) sequences for SimCom2 were used for our analysis. A plot of sequence quality as a function of fragment length is shown in Figure S3. Of these sequences, 7,499 (97%) for SimCom1 and 6,282 (89%) for SimCom2 were assigned to single species, if the 454 read shared ≥97% of sequence similarity to a Sanger-generated COI-5P sequence in our reference library or to barcode sequences archived in the public databases BOLD and GenBank. The proportion of reads assigned to taxa in the reference library was 78% for SimCom1 and 74% for SimCom2, and after similarity search in the public databases the proportion of sequences with a match increased to 97% in SimCom1 and to 89% in SimCom2 (Figure 1), due to the detection of concealed species in the bulk sample. This increase in the number of sequences assigned to a species was observed in all five primer pairs for both simulated communities. The primer pair E presented the highest increase of assigned reads (1247 reads in SimCom1 and 784 in SimCom2), after conducting similarity searches against public databases. The primer pair D has the highest proportion of matching reads with

sequence similarity higher than 97% in both simulated communities, while primer pairs C (SimCom1) and B (SimCom2) presented the lowest proportion of assigned reads. The number of 454-pyrosequencing usable reads with sequence similarity higher than 97% per 21 selected species is presented in Figure 2.

Our results indicate that a few species dominated the sample with a high number of represented reads. The limpet *Patella aspera* was the most highly represented species with 6158 reads in total. The species *Patella vulgata* and *Phorcus lineatus* were the next species with most abundant reads. Contrariwise, a high number of species had fewer numbers of representative reads. Three species were detected by a single read: *Lekanesphaera rugicauda* (SimCom1), *Hediste diversicolor* (SimCom2), and *Cyathura carinata* (SimCom2).

Differential taxon detection among the primers and simulated macrobenthic communities

The results of taxon detection for each primer combination after pyrosequencing of simulated macrobenthic communities are displayed in Table 2. The effectiveness of a primer set for the detection of a taxon was confirmed if at least one representative read matched a reference sequence with similarity greater than or equal to 97%. By using a combination of at least three PCR-amplification primer pairs, 18 species were recovered from SimCom1 and 16 species from SimCom2, together recovering 19 out of 21 species. Two of the 21 species (*Alvania mediolittoralis* and *Nassarius incrassatus*) used in the pooled samples of each community were not recovered, but also did not amplify by PCR in the preliminary tests. Nevertheless, no single or combined primer set was able to recover 100% of species present in any of the simulated communities.

In spite of the lower number of specimens used (1 per species) in SimCom1, a higher number of species was recovered compared to SimCom2. Moreover, low-size or low-biomass specimens were also detected. Species such as *Echinogammarus marinus*, *Melita palmata, Lekanesphaera rugicauda* and *Cyathura carinata*, which had just one representative in SimCom1, were successfully amplified, although only for a single primer set. *Patella aspera* was the single species detected in both communities for all primer sets. In SimCom1, three species, *Apohyale prevostii*, *Patella vulgata*, and *Ocinebrina edwardsii*, were recovered for all five primer pairs, while in SimCom2, *Phorcus lineatus* was the only species detected in all primer combinations.

Taxon recovery rates among the simulated macrobenthic communities

We observed that at least three primer sets combined are required for the highest recovery results in each simulated community with 76% (16/21) of recovered species in SimCom1 and 86% (18/21) in SimCom2. In SimCom1, the most successful primer pairs were C (658 bp) and D (313 bp), with 61.9% of recovered species, while in SimCom2, primer pair A (418 bp) detected 57.1% of species. The least successful primers were E (310 bp), with 42.9% of species detected in SimCom1, and B (470 bp), that recovered only 26.6% of the species in SimCom2. Overall, the most successful combination was primers A, C, and D for SimCom1 and primers A, C, and E for SimCom2.(Figure 3), therefore, four primer pairs were required for the highest recovery considering both communities.

The success of species detection was different among the two simulated macrobenthic communities. For primer pair B, C, and D, success was higher in SimCom1. For example, primer set D had a detection success in SimCom1 of 61.9%, whereas in SimCom2 it detected fewer than five species, corresponding to a detection success of 38.1%. A

different trend was observed with the primer pair A, which showed a slight increase in detection success from SimCom1 (52.4%) to SimCom2 (57.1%), and primer pair E, which detected the same number of species, but different ones in the two simulated macrobenthic communities.

Detection of species not listed in the simulated communities

Notably, the sequence similarity search in GenBank and BOLD detected a total of 18 taxa identified to species or genus level (at ≥97% sequence similarity), which were not part of the listed species in the two simulated macrobenthic communities (Table 3). These taxa were distributed along different animal phyla, namely Annelida (1), Chordata (1), Mollusca (2), Arthropoda (3), and also two phyla of algae, Ochrophyta (5) and Rhodophyta (6). The concealed taxa were recovered mostly from SimCom2, with 14 species/genus detected, while SimCom1 recovered six unlisted taxa. The primer pair B detected more species in SimCom1, while in SimCom2 the primer pair A was able to detect more unrepresented taxa in the sample. The primer pair D and E had detected no unlisted taxa in SimCom1. The algae, *Myrionema strangulans*, detected in SimCom1 and the barnacle, *Chthamalus stellatus,* detected in SimCom2, were the unlisted taxa represented by the most reads (4 and 63, respectively). A total of 10 out of the 18 concealed species detected had a single read.

DISCUSSION

Application of HTS in the assessment of marine benthic diversity has been focused mostly on microbial eukaryotes or the meiofauna (Chariton et al. 2010; Fonseca et al. 2014; Guardiola et al. 2015; Lallias et al. 2015; Lejzerowicz et al. 2015) and plankton communities (Zhan et al. 2013; Brown et al. 2015; Chain et al. 2016). The approach to

target meiofauna (nematodes, protists, fungi, etc) differs in a number ways from the one required for the macrofauna, which is the focus of this study. Here we perform a bulk DNA extraction from whole mixed specimens of a communitty, rather then extracting environmental DNA (eDNA) from an environmental sample. We also target the standard barcode region (COI-5P) instead of 18S DNA (18SRNA gene), which is typically used in meiofaunal HTS-surveys. Using COI-5P enables us to obtain species-level identifications, which has been the desired level of taxonomic resolution in the long history of morphology-based macrobenthos assessments, and which cannot be assured with the target region used for meiofauna (Tang et al. 2012). Furthermore, a comprehensive reference library of marine macrobenthic organisms is being built for COI-5P in our research group, providing a solid backup for species identification accuracy. So far, only a few studies used the standard barcode region in metabarcoding of macrobenthic communities (Cowart et al. 2015; Leray and Knowlton 2015), but in neither case there was a prior assessment of the success rates of amplification in known and experimentally contrived communities (e.g. Hajibabaei et al. 2011), or the exploitation of multiplexed approaches (e.g. Gibson et al. 2014) to increase species recovery.

The first objective of our study was to assess the impact of the barcode length and the choice of target region on the ability to discriminate species. In this study, all phenograms constructed for the different COI-5P fragments displayed similar clustering patterns with high bootstrap values and almost no loss in the species discrimination ability compared to the full barcode These results indicate the suitability of smaller fragments of the COI-5P barcode region for species-level resolution using our dataset. Short barcodes were also reported to be effective for identification of moth and wasp museum specimens (Hajibabaei et al. 2006) and gut contents of coral reef fishes (Leray et al. 2013). Meusnier et al. (2008) also successfully used short barcodes across all major eukaryotic groups.

The second objective was to find the appropriate set of primers able to successfully amplify as completely as possible the widest range of species in a given community. Our results showed that the all five combined primer sets used in 454-pyrosequencing recovered up to 90% (19 species out of 21) represented in both simulated communities. We failed to detect two species, *Alvania mediolittoralis* and *Nassarius incrassatus*. Those specimens were stored at 4ºC in 95% ethanol for a long period. Galindo et al. 2014 reported that gastropods can easily retract into the shell resulting in poor penetration of the ethanol into the tissues and we suspect that the DNA was extensively degraded. Moreover, *Alvania mediolittoralis* was the only species that was not represented in our reference library. Nonetheless, we could not find any match above 70% identity with *Alvania* COI sequences when a new similarity search was conducted against GenBank (which includes four species of the genus).

This study presents new data regarding the detection success of target barcode regions in the recovery of species represented by a single small individual within a simulated community. In a study using artificially contrived communities, Pochon et al. (2013) demonstrated that individual species present at greater than 0.64% abundance could be detected. Other studies reported failures in DNA-based identification of species that were represented in low frequency and argued that bias associated with primer binding and the presence of competing COI sequence information could be the presumable causes (Hajibabaei et al. 2011; Hajibabaei et al. 2012). Indeed, the composition of samples seems to affect the species recovery, as we found that SimCom1, which was composed of only one specimen per species, had the best recovery results regarding small specimens (5 to 8 mm), when compared to SimCom2 containing higher number of specimens. Increasing sequencing depth by using other HTS platforms (e.g. Illumina) may help to overcome potential bias that originated from the dominance of amplicons from

certain species compared to others present in the mixture (Shendure and Ji, 2008; Shokralla et al. 2015).

Furthermore, we observed that the recovery of some species may be dependent on primer binding affinity, since species like *Acantochitona crinita, Cyathura carinata*, and *Lepidochitona cinerea* were not recovered with the same single primer set in both communities. Since the goal was to identify a wide range of species in the sample, the design and optimization of versatile primers are fundamental for an effective species recovery (Geller et al. 2013; Leray et al. 2013; Gibson et al. 2014). We have observed that only three primer sets were sufficient to recover the total number of species detected, although in different combinations for each simulated community. This approach is especially advantageous if one primer set is biased towards selective amplification of certain taxa. Several studies have shown that a multiplex amplification regime may increase the detection of species. A study conducted by Hajibabaei and collaborators (2011) showed that, using a multiplex PCR approach for NGS-based environmental barcoding, 100% detection was achieved for taxa represented with more than 1% individuals in the mixture. Using the 454 platform, Pochon and collaborators (2013) found that four distinct primer sets would be required to obtain species-level identification within the COI gene across five marine invasive groups. Similarly, Gibson et al. (2014) used 11 primer sets to investigate the diversity found in Malaise trap samples taken from tropical Costa Rica. Three gene regions (COI, 16S, and 18S) were analyzed, and they found a much higher recovery rate across taxa when all 11 primer sets were used compared to any primer set alone. The use of simulated communities with known composition allowed us to consistently assess the species biodiversity of the sample, including the identifications of singletons that, otherwise, could be considered as false positives. Despite the fact that some primer sets might not be adequate for a target species, we observed variations in the ability of the same primer set to recover a target species in both

of the simulated communities. One example seen in this study is the successful amplification of *Mytilus* by primer D in SimCom2 and its failure using the same primer in SimCom1, both communities containing one specimen. This could be the result of random sampling during PCR or sequencing. Biased PCR amplification has been reported in other studies (e.g. Bellemain et al. 2010; Zhang et al. 2014, Dowle et al. 2015). Nonetheless, this result indicates that some additional work is needed to test detection limit variations in samples containing diverse taxa at different abundances. Moreover, some adjustments in HTS sequencing protocols could be made in order to tune sequencing depth and coverage.

The proportion of reads did not correlate with specimen abundance, as SimCom1 has the same number of specimens per species but variance in read abundance, and the species abundance in SimCom2 does not correspond to read abundance (Figure 2). Hajibabaei et al. (2011) suggested that species with a higher affinity for their primer binding sites and/or species with higher abundance (i.e. more biomass in a bulk sample) can capture more primer molecules during the process of PCR annealing. The latter explanation does not corroborate our results, and the affinity of the primers used in this study appears to play a significant role in the observed number of reads and species detection. Nevertheless, to better investigate this issue it would be important to perform tests with other simulated communities containing different species and varying abundances to disentangle abundance from primer-binding effects.

Searching for species that could be possibly associated (e.g. species present in the gut contents, epibionts, etc) with others in our simulated community, an additional similarity search was conducted on BOLD and GenBank for sequences that originally generated matches between 70-97% against the reference library. Interestingly, a small number of 454-pyrosequencing reads (178 in total, minimum 1 and maximum 57 in different primer pairs and simulated communities) matched sequences at species or genus level that were

not originally represented as individuals in our experimental communities. Unrepresented species in bulk samples were also observed by Hajibabaei et al. (2011). We found 17 taxa identified to species or genus level concealed in our simulated communities. The polychaete *Eulalia viridis* was recovered in SimCom1. This worm often seeks protection during high tides underneath clumps of mussels (Emson, 1977) and could be associated with the *Mytilus* present in our sample. The arthropod *Mytilicola intestinalis* was detected in SimCom2 by a fair number of reads. This is a parasitic copepod living in the intestine of bivalves, such as oysters (Elsner et al. 2011) or cockles (Carballal et al. 2001), but they are most frequentely reported in mussels' intestine (Dethlefsen, 1985; Trotti et al. 1998). In addition, five species of Ochrophyta (brown alga) and six species of Rhodophyta (red alga) were detected, but most of them yielded a small number of reads. The species *Bangia atropurpurea*, *Corallina caespitosa*, *Myrionema strangulans*, *Porphyra umbilicalis*, and *Zonaria tournefortii* are found along the Portuguese coast (Guiry and Guiry 2015, http://www.algaebase.org/; Pereira and Neto, 2015). However, no database records were found in our coast for the remainder of the detected algae species. Various species in our simulated communities, including molluscan (e.g. *P. vulgata*, *P. aspera*, *Mytilus* sp.) and crustacean species (e.g. *E. marinus*, *C. carinata*) may feed on algae (Martins et al. 2010). Algae are also known to be able to live in epibiosis with groups of organisms such as crustaceans and mollusks. The detection of two barnacle species is very likely the result of their common occurrence in the shells of mussels, and, if so, this illustrates the exceptional detection ability of metabarcoding procedures compared to morphology-based assessments. The DNA of *P. depressa* found in our sample could have leaked to the ethanol used to preserve the unsorted specimens and was accidentally carried over with the specimens examined in the simulated communities. Hajibabaei and collaborators (2012) showed that DNA leakage to preserved ethanol can occur, and taxa can be detected through HTS of the preservative ethanol added to field collected organisms

(before sorting bulk benthic samples), demonstrating that ethanol can be an additional source of DNA providing useful information in metabarcoding studies.

Our study shows the feasibility of using COI for metabarcoding of macrobenthic communities, where a combination of new primer design and testing, together with multiplexing, can circumvent possible bias in the amplification success of the COI target region among different macrobenthic species in a bulk DNA template. This is possible because smaller and distinct target regions within the COI-5P barcode still allow species-level discrimination. Results suggest that multiplexing COI metabarcoding with only four primer combinations (e.g. A, C, D, and E) is enough to attain fairly high recovery rates in a phylogenetically diverse macrobenthic communties, even for taxa at low frequency and with comparatively minute biomass. Nevertheless, further improvement is still required in order to increase the recovery success rates. With the expected decreasing costs and throughput capability of HTS technologies, application of multiplexed approaches will be less costly, while improving detection success and overall quality of the assessments (Shokralla et al. 2012).

ACKNOWLEDGMENTS

This work was supported by FEDER through POFC-COMPETE by national funds from 'Fundação para a Ciência e a Tecnologia (FCT)' in the scope of the grant FCOMP-01-0124-FEDER-015429 and also by the strategic programme UID/BIA/04050/2013 (POCI-01-0145-FEDER-007569). This work was also funded by national funds through the FCT I.P. and by the ERDF through the COMPETE2020 - Programa Operacional Competitividade e Internacionalização (POCI). Jorge Lobo was supported by a PhD fellowship (SFRH/BD/69750/2010) from FCT and Claudia Hollatz by a CAPES Post-doctoral fellowship, under Science Without Borders Program (Ministry of Education, Brazil).

REFERENCES

Bickford, D., Lohman, D.J., Sodhi, N.S., Ng, P.K.L., Meier, R., Winker, K., Ingram, K.K., and Das, I. 2007. Cryptic species as a window on diversity and conservation. Trends Ecol. Evol. 22: 148-155.

Brown, E.A., Chain, F.J., Crease, T.J., MacIsaac, H.J., and Cristescu, M.E. 2015. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities?. Ecol. Evol. 5(11): 2234-2251.

Carballal, M.J., Iglesias, D., Santamarina, J., Ferro-Soto, B., and Villalba, A. 2001. Parasites and pathologic conditions of the cockle *Cerastoderma edule* populations of the coast of Galicia (NW Spain). J. Invertebr. Pathol. 78: 87-97.

Chain, F.J., Brown, E.A., MacIsaac, H.J., and Cristescu, M.E. 2016. Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. Diversity Distrib. 22: 493-504.

Dethlefsen, V. 1985.*Mytilicola intestinalis*, parasitism. Leaflet No. 24. ICES Identification Leaflets for Diseases and Parasites of Fish and Shellfish, ICES, Copenhagen, 4pp.

Dowle, E.J., Pochon, X., Banks, J.C., Shearer, K., and Wood, S.A. 2015. Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. Mol. Ecol. Resour. Available from http://dx.doi.org/10.1111/1755-0998.12488 [accessed 10 March 2016].

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinf. 26(19): 2460-2461.

Elsner, N.O., Jacobsen, S., Thieltges, D.W., and Reise, K. 2011. Alien parasitic copepods in mussels and oysters of the Wadden Sea.Helgol. Mar. Res. 65: 299-307.

Emson, R.H. 1977. The feeding and consequent role of *Eulalia viridis* (O. F. Muller) (Polychaeta) in intertidal communities. J. Mar. Biol. Ass. UK. 57: 93-96.

Fonseca, V.G., Carvalho, G.R., Nichols, B., Quince, C., Johnson, H.F., Neill, S.P., Lambshead, J.D., Thomas, W.K., Power, D.M., and Creer, S. 2014. Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. Glob. Ecol. Biogeogr. 23 (11):1293-1302

Galindo, L.A., Puillandre, N., Strong, E.E., and Bouchet, P. 2014. Using microwaves to prepare gastropods for DNA barcoding. Mol. Ecol. Resour 14:700-705.

Geller, J., Meyer, C., Parker, M., and Hawk, H. 2013. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Mol. Ecol. Resour. 13: 851-861.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W., and Hajibabaei, M. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc. Natl. Acad. Sci. USA. 111: 8007-8012.

Guardiola, M., Uriz, M.J., Taberlet, P., Coissac, E., Wangensteen, O.S., and Turon, X. 2015. Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from Sediments of Marine Canyons. PloS One 10: e0139633.

Guiry, M.D. and Guiry, G.M. 2015. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. Available from http://www.algaebase.org. [accessed October 17 2015].

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B., and Hebert, P.D.N. 2006. A minimalist barcode can identify a specimen whose DNA is degraded. Mol. Ecol. Notes 6: 959-964.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C., and Baird, D.J. 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One 6: e17497.

Hajibabaei, M., Spall, J.L., Shokralla, S., and van Konynenburg, S. 2012. Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC Ecology. 12: 28.

Hebert ,P.D.N., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. Biol. 270: 313-321.

Jarman, S.N. and Elliott, N.G. 2000. DNA evidence for morphological and cryptic Cenozoic speciations in the Anaspididae, 'living fossils' from the Triassic. J. Evol. Biol. 13: 624-633.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111-120.

Knowlton, N. 1993. Sibling species in the sea. Ann. Rev. Ecol.System. 24: 189-216.

Lallias, D., Hiddink, J.G., Fonseca, V.G., Gaspar, J.M., Sung, W., Neill, S.P., Barnes, N., Ferrero, T., Hall, N., Lambshead, P.J., and Packer, M. Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. ISME J. 9(5): 1208-21.

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T.A., Black, K.D., Pawlowski, J. 2015. High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. Sci. Rep. Available from http://dx.doi.org/10.1038/srep13932 [accessed 10 March 2016].

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T., and Machida, R.J. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Front. Zool. 10: 34.

Leray, M., and Knowlton, N. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proc. Natl. Acad. Sci. USA 112: 2076-2081.

Lobo, J., Costa, P.M., Teixeira, M.A.L., Ferreira, M.S.G., Costa, M.H., and Costa, F.O. 2013. Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. BMC Ecology 13: 34.

Martins, G.M., Faria, J., Furtado, M., and Neto, A.I. 2014. Shells of *Patella aspera* as 'islands' for epibionts. J. Mar. Biol. Assoc. U.K. 94: 1027-1032.

Martins, G.M., Thompson, R.C., Neto, A.I., Hawkins, S.J., and Jenkins, S.R. 2010. Exploitation of intertidal grazers as a driver of community divergence. J. Appl. Ecol. 47: 1282-1289.

Meusnier, I., Singer, G.A.C., Landry, J.F., Hickey, D.A., Hebert, P.D.N., and Hajibabaei, M. 2008. A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9: 214.

Pereira, L. and Neto, J.M. eds., 2014.Marine algae: biodiversity, taxonomy, environmental assessment, and biotechnology. CRC Press.

Pochon, X., Bott, N.J., Smith, K.F., and Wood, S.A. 2013. Evaluating Detection Limits of Next-Generation Sequencing for the Surveillance and Monitoring of International Marine Pests. PLoS One 8: e73935.

Ratnasingham, S., and Hebert, P.D.N. BOLD: The Barcode of Life Data System. 2007. Mol. Ecol. Resour. 7: 355-364.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. 2012. MrBayes 3.2: efficient Bayesian

phylogenetic inference and model choice across a large model space. System. Biol. 61 (3):539-42.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. Nat. Biotech. 26: 1135-1145.

Shokralla, S., Spall, J.L., Gibson, J.F., and Hajibabaei, M. 2012. Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. 2012. 21 (8): 1794-805.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., and Golding, G.B., and Hajibabaei, M. 2015. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Sci. Rep. 5: 9687.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21: 2045-2050.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30: 2725-2729.

Tang, C.Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T.G., and Fontaneto, D. 2012. The widely used small subunit 18S rDNA molecule greatly underestimates true

diversity in biodiversity surveys of the meiofauna. Proc. Natl. Acad. Sci. USA, 109(40): 16208-16212.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

Trotti, G.C., Baccarani, E.M., Giannetto, S., Giuffrida, A., and Paesanti, F. 1998. Prevalence of *Mytilicola intestinalis* (Copepoda: Mytilicolidae) and *Urastoma cyprinae* (Turbellaria: Hypotrichinidae) in marketable mussels *Mytilus galloprovincialis* in Italy. Dis. Aquat. Org. 32: 145-149.

Zhan, A., Hulak, M., Sylvester, F., Huang, X., Adebayo, A.A., Abbott, C.L., and MacIsaac, H.J. 2013. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. Meth. Ecol. Evol. 4(6): 558-565.

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinf. 29: 2869-2876.

Table 1. A - Schematic representation of the amplicons and their size, generated after PCR amplification. B - PCR primer combinations and respective thermal cycling conditions for the five primer pairs. C - Primers used for PCR amplification of COI-5P gene fragments from the two different simulated communities. The COI-5P barcode and the five primer pairs that were used in PCR amplification within the standard barcode are represented (5´-3').

A

B

Primer name	Sequence (5' → 3')	Reference
ArF2	GCICCIGAYATRGCITTYCCICG	Gibson et al., 2014
invF	ATRATYTTYTTYITIGTIATRCC	Lobo J, this study
jgLCO1490	TITCIACIAAYCAYAARGAYATTGG	Geller et al., 2013
mICOIintF	GGWACWGGWTGAACWGTWTAYCCYCC	Leray et al., 2013
LoboR	TAAAACYTCWGGRTGWCCRAARAAYCA	Lobo et al., 2013
jgHCO2198	TAIACYTCIGGRTGICCRAARAAYCA	Geller et al., 2013
ArR5	GTRATIGCICCIGCIARIACIGG	Gibson et al., 2014

C

Primer combinations	PCR conditions
ArF2/LoboR	94 °C 5’ 94 °C 30’’ | 46 °C 1’ | 68 °C 1’ 45x 68 °C 10’ | 4°C ∞
invF/LoboR	94 °C 5’ 94 °C 30’’ | 45 °C 90’’ | 68 °C 1’ 5x 94 °C 30’’ | 50 °C 90’’ | 68°C 1’ 40x 68 °C 10’ | 4°C ∞
jgLCO1490/ jgHCO2198	94 °C 5’ 94 °C 30’’ | 48 °C 30’’ | 68 °C 1’ 30x 68 °C 10’ | 4°C ∞
mlCOIintF/LoboR	94 °C 5’ 94 °C 30’’ | 62 °C (-1 per cycle) 30’’ | 68 °C 1’ 6x 94 °C 30’’ | 46 °C 30’’ | 68°C 1’ 25x 68 °C 10’ | 4°C ∞
ArF2/ArR5	94 °C 5’ 94 °C 30’’ | 46 °C 1’ | 68 °C 1’ 30x 68 °C 10’ | 4°C ∞

Table 2. Species detection (1) or failed detection (0) for each primer pair after HTS of SimCom1 and SimCom2. Dark gray: a species that was detected with the five primers in the two simulated communities; Light gray: the two species that were not detected with any of five primer pairs in the two simulated communities. A - primer pair ArF2/LoboR; B - primer pair invF/LoboR; C - primer pair jgLCO1490/jgHCO2198; D - primer pair mlCOIintF/LoboR; E - primer pair ArF2/ArR5.

Primers / Species	SimCom 1					SimCom 2				
	A	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
Hediste diversicolor	0	1	0	0	1	0	0	0	0	1
Apohyale prevostii	1	1	1	1	1	1	0	1	0	1
Corophium multisetosum	0	0	0	0	0	1	0	1	0	0
Echinogammarus marinus	1	0	0	0	0	1	1	0	0	1
Melita palmata	0	0	0	1	0	0	0	0	0	0
Cyathura carinata	0	0	1	0	0	0	0	1	0	0
Dynamene bidentata	1	0	1	0	1	1	0	1	0	1
Lekanesphaera rugicauda	0	0	0	1	0	0	0	0	0	0
***Mytilus* sp.**	0	1	1	0	0	0	1	1	1	0
Gibbula cineraria	1	0	1	1	1	1	0	0	0	1
Phorcus lineatus	1	1	1	1	0	1	1	1	1	1
Patella aspera	1	1	1	1	1	1	1	1	1	1
Patella vulgata	1	1	1	1	1	1	0	1	1	1
Alvania mediolittoralis	0	0	0	0	0	0	0	0	0	0
Nassarius incrassatus	0	0	0	0	0	0	0	0	0	0
Nassarius reticulatus	0	0	0	1	1	0	0	0	0	0
Nucella lapillus	1	0	1	1	1	1	0	0	0	0
Ocinebrina edwardsii	1	1	1	1	1	1	0	0	1	1
Siphonaria pectinata	0	1	1	1	0	0	0	1	1	0
Acanthochitona crinita	1	1	1	1	0	1	1	1	1	0
Lepidochitona cinerea	1	1	1	1	0	1	1	1	1	0

Table 3. Detected taxa that were not listed in the simulated communities after the sequence similarity search in public databases (at 97%). P: primer pair used: A - ArF2/LoboR; B - invF/LoboR; C - jgLCO1490/jgHCO2198; D - mlCOIintF/LoboR; E - ArF2/ArR5. R: number of reads generated by 454 pyrosequencing.

Phylum	Class	Order	Species	SimCom1		SimCom2	
				P	R(n)	P	R(n)
Annelida	Polychaeta	Phyllodocida	*Eulalia viridis*	**B**	**1**		**-**
Arthropoda	Maxillopoda	Poecilostomatoida	*Mytilicola intestinalis*			**A**	**21**
					-	**D**	**31**
						E	**6**
		Sessilia	*Chthamalus montagui*			**A**	**1**
						B	**1**
					-	**D**	**2**
						E	**5**
			Chthamalus stellatus			**A**	**26**
						B	**3**
					-	**C**	**10**
						D	**4**
						E	**20**
Chordata	Mammalia	Primates	*Homo sapiens*		-	**C**	**1**
Ochrophyta	Phaeophyceae	Dictyotales	*Zonaria tournefortii*		-	**B**	**1**
		Ectocarpales	*Chordariac* sp. 2GWS		-	**A**	**1**
			Ectocarpus sp. 1TAS		-	**A**	**1**
			Myrionema strangulans	**A**	**2**		**-**
				B	**2**		
			Streblonema sp. 2GWS	**A**	**1**		**-**
				B	**3**		
Mollusca	Gastropoda	Littorinimorpha	*Littorina saxatilis*		-	**D**	**1**

			Patella depressa	-	**A**	**3**
					E	**2**
Rhodophyta	Bangiophyceae	Bangiales	*Bangia atropurpurea*	**B** **1**	-	
			Bangia sp. 2LH	**C** **1**	**A**	**4**
					B	**9**
					C	**5**
			Porphyra umbilicalis	**B** **1**	**A**	**1**
	Florideophyceae	Corallinales	*Corallina caespitosa*	-	**B**	**1**
			Jania sp. 1MX	-	**D**	**2**
		Gigartinales	*Peyssonnelia* sp. 1WA	-	**B**	**4**

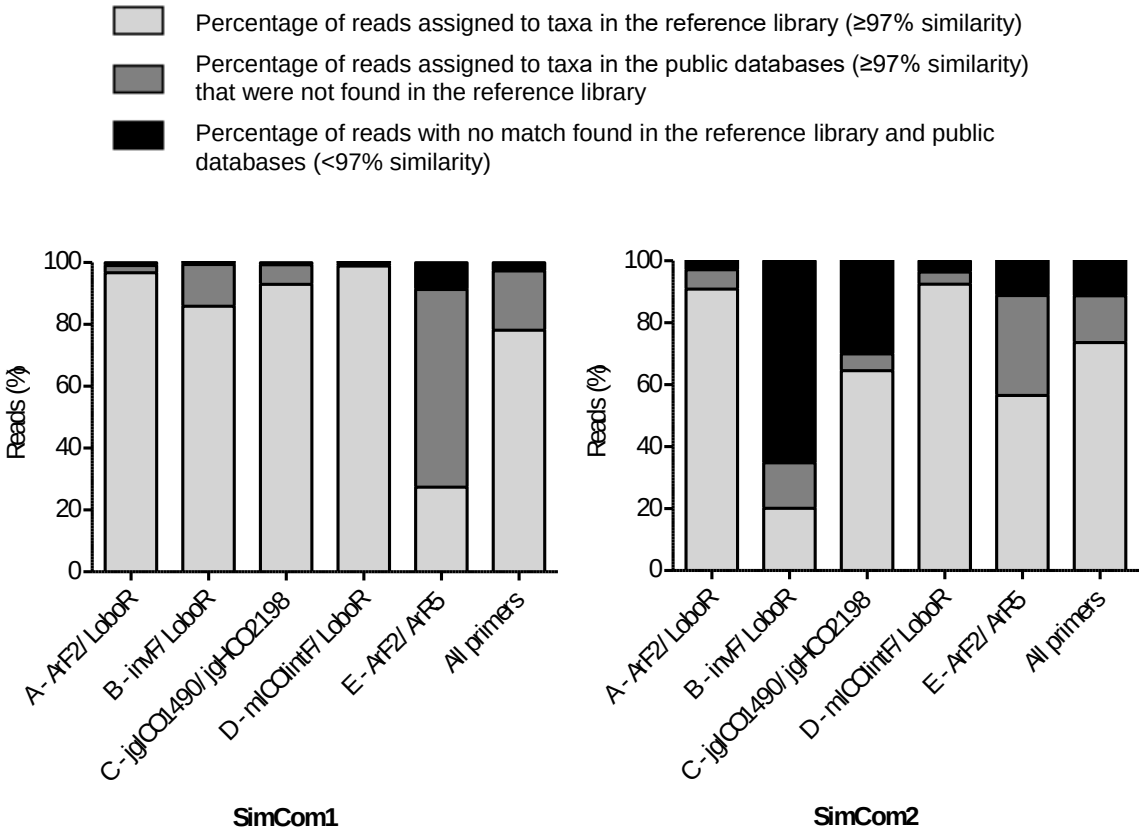
21

FIGURE LEGENDS

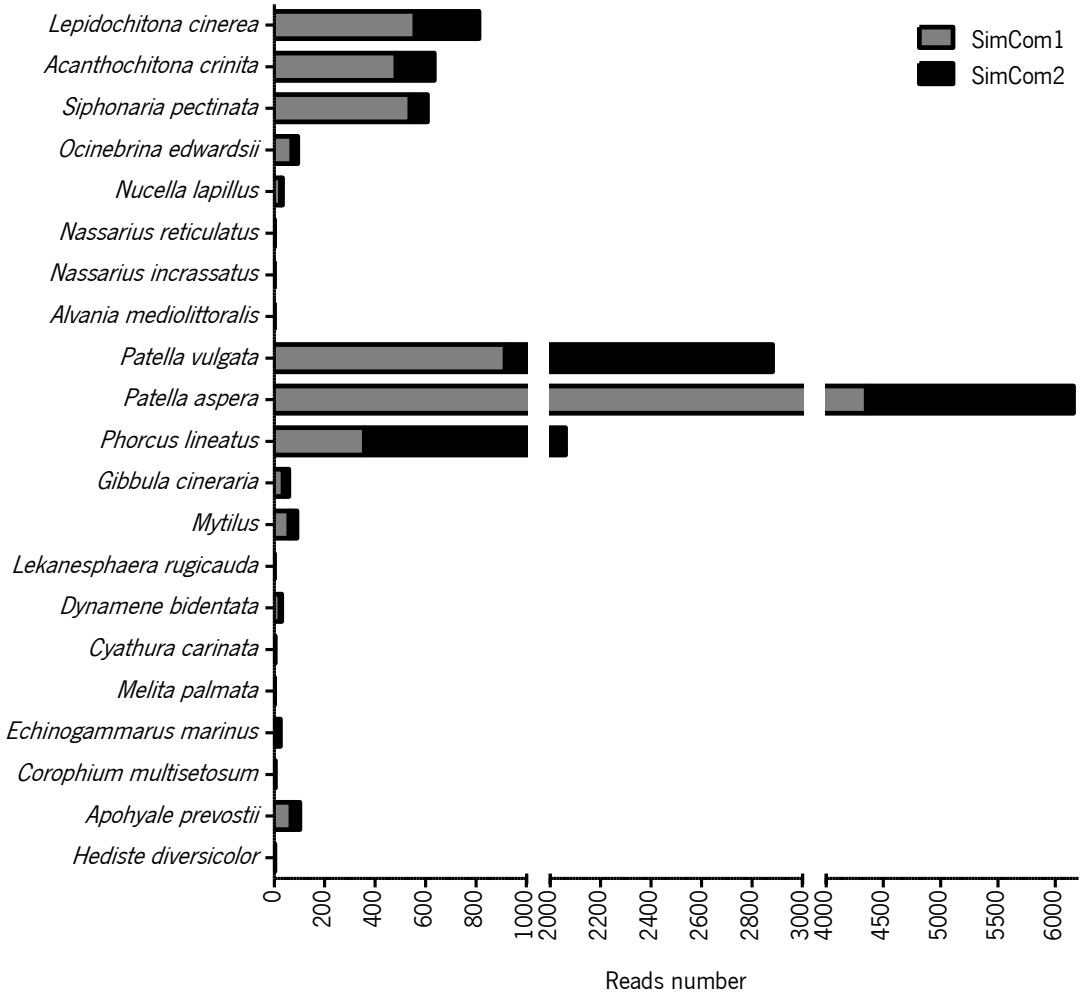
Figure 1. Percentage of quality-filtered 454 pyrosequencing reads assigned to taxa, generated from amplicons of different size and location within the COI barcode region. The reads' percentage is displayed separately for each and all combined five PCR primer pairs (A to E), and for each simulated macrobenthic community (SimCom1/SimCom2). Reads' assignments were performed through similarity search against a reference library or against public databases (GenBank, BOLD). Percentage of reads with no database match is also shown.

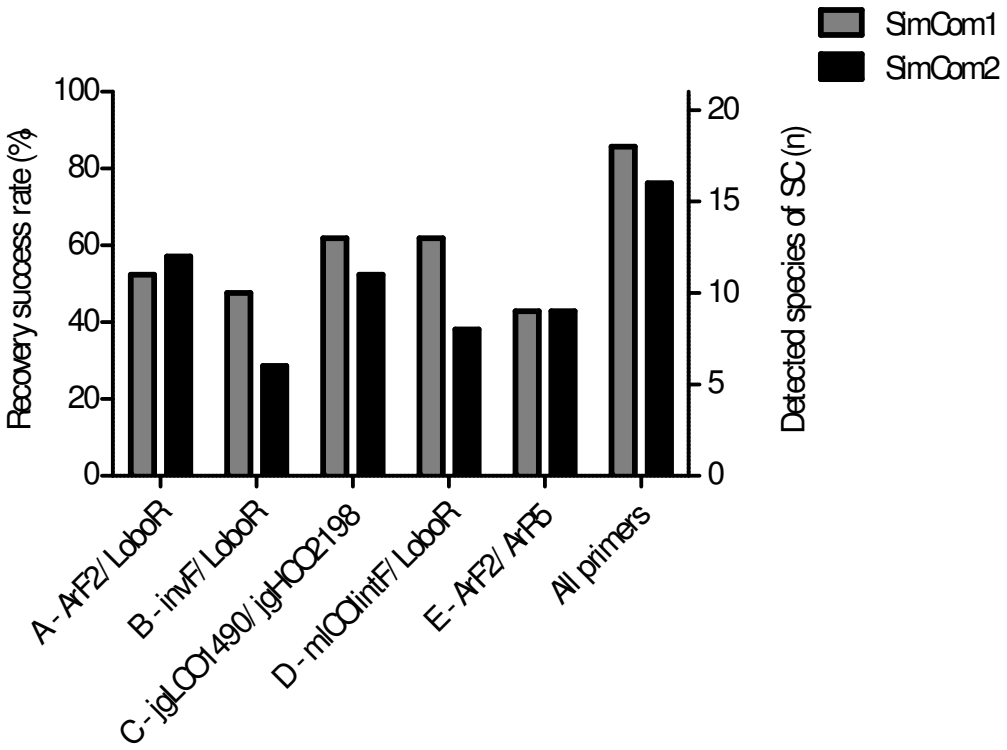
Figure 2. Number of reads generated by 454 pyrosequencing for each of the 21 species of three phyla in each of the two simulated macrobenthic communities.

Figure 3. Percentage of taxa recovered and number of species recovered for each and all combined five primer pairs by PCR 454-pyrosequencing of COI barcode region. Results are shown for each simulated macrobenthic community.


Percentage of reads assigned to taxa in the reference library (≥97% similarity)
Percentage of reads assigned to taxa in the public databases (≥97% similarity) that were not found in the reference library
Percentage of reads with no match found in the reference library and public databases (<97% similarity)
Reads (%)
100
80
60
40
20
0
A - ArF2/ LoboR
B - invF/ LoboR
C - jgLCO1490/ jgHCO2198
D - mlCOIintF/ LoboR
E - ArF2/ ArR5
All primers
SimCom1
SimCom2

1


Lepidochitona cinerea
Acanthochitona crinita
Siphonaria pectinata
Ocinebrina edwardsii
Nucella lapillus
Nassarius reticulatus
Nassarius incrassatus
Alvania mediolittoralis
Patella vulgata
Patella aspera
Phorcus lineatus
Gibbula cineraria
Mytilus
Lekanesphaera rugicauda
Dynamene bidentata
Cyathura carinata
Melita palmata
Echinogammarus marinus
Corophium multisetosum
Apohyale prevostii
Hediste diversicolor
SimCom1
SimCom2
0
200
400
600
800
1000
2000
2200
2400
2600
2800
3000
4000
4500
5000
5500
6000
Reads number


SimCom1
SimCom2
Recovery success rate (%)
Detected species of SC (n)
100
80
60
40
20
0
20
15
10
5
0
A - ArF2/ LoboR
B - invF/ LoboR
C - jgLCO1490/ jgHCO2198
D - mlCOIintF/ LoboR
E - ArF2/ ArR5
All primers