Universidade do Minho
Escola de Ciências

Barbara Daniela Rocha Leite

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

October 2015

Barbara Daniela Rocha Leite
Priming of a DNA metabarcoding approach for species identification and inventory in marine macrobenthic communities

UMinho | 2015

Universidade do Minho
Escola de Ciências

Barbara Daniela Rocha Leite

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

Master Thesis
Master in Molecular Genetics

Supervised by
Professor Doctor Filipe Oliveira Costa
Doctor Claudia Hollatz

October 2015

DECLARAÇÃO

Nome: Barbara Daniela Rocha Leite

Endereço eletrónico: barbaradrl.bio@gmail.com Telefone: 916537001

Cartão do Cidadão: 14164251

Título da dissertação: Priming of a DNA metabarcoding approach for species identification and inventory in marine macrobenthic communities

Orientadores:

Professor Doutor Filipe Oliveira Costa

Doutora Claudia Hollatz

Ano de conclusão: 2015

Mestrado em Genética Molecular

Universidade do Minho, _____/_____/_____

Assinatura:_____

(Barbara Daniela Rocha Leite)

ACKNOWLEDGMENTS

I would like to address a few words to the ones who helped me through my academic journey and directly or indirectly contributed to this work.

Firstly, I would like to express my gratitude to my adviser, Professor Filipe Costa, for he instilled in me the greed to work in this area, and therefore made me challenge myself. A big thanks for your orientation, piece of advice, all the knowledge transmitted, patience, availability and promptness. Your honesty and encouragement as well as your support were the key to the development of my critical thinking and essential to my success. To you a huge thanks.

Secondly, to Claudia Hollatz for all the knowledge transmitted, all the patience, availability and dedication to help me when I need. Thank you for your sympathy, friendship and cheeriness. Since day one you helped me in the lab and your guidance was remarkable. It was a pleasure working with you!

To Dr. Conceição Egas, Dr. Hugo Froufe and Jorge Lobo for the collaboration in the project. Thank you for the availability, technical support and sharing of ideas.

To my lab colleagues for the companionship and the good moments we spent.

To Susana, Jéssica, Benedita and Marina for being there when I need. Thank you for the support during this journey and for the awesome moments we have spent. I never forgot my "*Milho companheiras*" who managed to put a smile on my face when times were tough. Your friendship has been important and I hope it prevails for the rest of our lives.

To my mother, father and sister. Though the distance and sacrifices were hard, they were important on my development as a person. You made me who I am. The pride you always shown, permanently encouraging me to pursue my goals were of the utmost importance. A special thanks to my little nephew for all the laughter, cries and our special *cu-cu* moments. For my grandfathers: I promise "to behave nicely". To them and to all my family in general an eternal thank you.

To José, who always believe in me. Thank you for your dedication, piece of advice and specially for the patience at the end of this journey. Thank you for your friendship and all the happiness that you bring to my days. Thanks for all your love. Thanks for everything!

The present study was financed by FEDER through POFC-COMPETE, in the scope the project FCOMP-01-0124-FEDER-015429 funded by "Fundação para a Ciência e a Tecnologia" (FCT), Portugal. Work at CBMA was supported by FCT I.P. through the strategic funding UID/BIA/04050/2013.

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

ABSTRACT

In marine and estuarine benthic communities, the inventory and estimation of species richness are often hampered by the need of broad taxonomic expertise across several phyla. The use of DNA metabarcoding has emerged as a powerful tool on the fast assessment of species composition from whole environmental communities. Yet, specifically designed methodologies for marine and estuarine macrobenthic communities are still lacking. Here we tested the amplification success of five primer sets targeting different COI-5P regions with fragments ranging from 310 to 658 bp. To this end, we used two simulated macrobenthic communities (SimCom1 and 2), each community containing the same number of species (21), but different number of specimens (SimCom1: 21; SimCom2: 67). Sequences were generated using high-throughput sequencing on 454 platform and species identification were first performed against a compiled reference library of macrobenthic species. In order to achieve new identifications at species level, which had no representation in the reference library, two public databases, BLASTn and BOLD-IDS, were used to rerun those sequences with similarity between 70-97%. Interestingly, amplicons of 313 and 658 bp were equally successful on the detection of species in SimCom1 (≈62%), while for SimCom2 the highest success rate were obtained using a 418 bp fragment. However, the combination of the five primer sets was able to detect more sequences than any primer set alone, achieving 85% of represented species in SimCom1 and 76% in SimCom2, across all analysed marine phyla (Annelida, Arthropoda and Mollusca). Unrepresented species were also detected in these communities, such as algae and the mussel parasitic copepod *Mytilicola intestinalis*. We demonstrated that the application of combined primer sets coupled with high-throughput technologies has a great potential to overcome the challenges on marine bioassessment, and inventory, including the detection of a "hidden" biodiversity that could not possibly be identified based on morphology.

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

Desenvolvimento da técnica de DNA metabarcoding para identificação e inventário de espécies em comunidades macrobentónicas marinhas

RESUMO

O inventário e a estimativa da riqueza de espécies, em comunidades bentónicas marinhas e estuarinas, são frequentemente dificultados pela necessidade de um amplo e especializado conhecimento taxonómico dos diversos filos. A utilização de *DNA metabarcoding* surgiu como uma ferramenta poderosa para uma rápida avaliação da composição das espécies constituintes das comunidades ambientais. No entanto, ainda falta conceber metodologias especificamente desenhadas para comunidades macrobentónicas marinhas e estuarinas. No presente estudo, testou-se o sucesso de amplificação de cinco pares de *primers* referentes a diferentes regiões do gene COI-5P com fragmentos que variam entre 310 a 658 pb. Com esta finalidade, usou-se duas comunidades macrobentónicas simuladas (SimCom1 e 2), cada comunidade contendo o mesmo número de espécies (21), mas um diferente número de espécimes (SimCom1: 21; SimCom2: 67). Usou-se a sequenciação de alto débito na plataforma 454 para gerar sequências e a identificação de espécies foi primeiramente realizada contra uma biblioteca de referência compilada de espécies macrobentónicas. De modo a obter-se novas identificações ao nível da espécie, que não tinham representação na biblioteca de referência, foram usadas duas bases de dados públicas, BLASTn e BOLD-IDS, para executar novamente as sequências com similaridades entre 70-97%. Curiosamente, os amplicões de 313 e 658 pb foram igualmente bem-sucedidos na deteção de espécies na SimCom1 (≈62%), enquanto que para SimCom2 obteve-se a maior taxa de sucesso utilizando o fragmento de 418 pb. No entanto, a combinação dos cinco pares de *primers* foi capaz de detetar mais sequências do que qualquer par de *primers* por si só, obtendo-se 85% das espécies representadas em SimCom1 e 76% em SimCom2, em todos os filos marinhos analisados (Annelida, Arthropoda e Mollusca). Nas comunidades simuladas também foram detetadas espécies que não estavam representadas, como algas e o copépode parasita de mexilhões *Mytilicola intestinalis*. Este estudo demonstrou que através da aplicação de combinações de pares *primers* juntamente com tecnologias de alto débito há um grande potencial

para ultrapassar os desafios da avaliação e do inventário da biodiversidade marinha, incluindo a deteção de biodiversidade "escondida", a qual não seria possível identificar através da morfologia.

Palavras-Chave: *DNA barcoding*, Sequenciação de alto débito, Avaliação da biodiversidade, Macrobentos marinhos

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

µL	Microliter
AMBI	AZTI's Marine Biotic Index
ATP	Adenosine triphosphate
BIN	Barcode Index Number System
BLAST	Basic Local Alignment Search Tool
BOLD	Barcode of Life Database
BOLD-IDS	Barcode of Life Database Identification System
BOLI	Barcode of Life Initiative
bp	Base pairs
CBOL	Consortium for the Barcoding of Life
COI-5P	5' end of the Cytochrome *c* Oxidase subunit I gene
DDBJ	DNA Data Bank of Japan
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleotide
dNTPs	Deoxyribonucleotide Triphosphates
EMBL	European Molecular Biology Laboratory
g	Gramme
HTS	High Throughput Sequencing
iBOL	International Barcode of Life
ITS	Internal Transcribed Spacer
K2P	Kimura 2-Pameter
matK	Maturase K
mL	Mililiter
NCBI	Nacional Center for Biotechnology Information
NJ	Neighbor-Joining
ºC	Degree Celsius

OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
PPI	Pyrophosphate
rbcl	Ribulose-biphosphate Carboxylase
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
SimCom	Simulated Communities
WFD	Water Framework Directive
WoRMS	World Register of Marine Species

1. INTRODUCTION

The estimation of species richness and the recognition of the interactions with ecosystem functioning are essential to understand global biodiversity. Likewise, the impact of environmental change and anthropogenic disturbances need to be identified and mitigated to maintain a healthy environment and sustainable economy. However, the questions concerning to the historical genetic structure and identification of species remain a mystery (Bik *et al.*, 2012).

The total number of extant species in the world is approximately 100 million (Chapman, 2009). However fewer than two million have been formally known (Fonseca *et al.*, 2010) and despite being an estimate, there is a redundancy in descriptions of many species names (Paterson *et al.*, 2010). There is still a huge gap in our knowledge of biodiversity, the task of cataloguing all biological diversity faces primary problems, such as lack of resources, expertise and novel approaches to identify new taxa. This problem has been commonly referred by the scientific community as "taxonomic impediment" (Rodman and Cody, 2003). Additional conceptual and operational challenges to understand the complexity of biological diversity emerged from the inability of the researchers to find universal criteria for species recognition (Costa and Antunes, 2012).

The taxonomists work have been underestimated and the maintenance and development of infrastructures are needed (Bouchet, 2006). Technological developments and global initiatives are on demand to perform a profound change on taxonomy, increasing their influence in society (Wheeler, 2008).

1.1 DNA barcodes in taxonomic identification of species

Taxonomy is a scientific discipline responsible for identification, description and classification of biodiversity to define groups of species based on their common characteristics (Costa and Antunes, 2012; Padial *et al.*, 2010). Over the years, taxonomists were capable to describe and to catalogue species. The first method implemented for species identification was based on easily observable morphological characteristics (Taberlet *et al.*, 2012). Expert taxonomists employed optical techniques and this may led to incorrect identifications due to

phenotypic plasticity and genetic variability of the species (Hebert *et al.*, 2003a). Furthermore, the study of cryptic species (morphologically indistinguishable species) (Costa and Antunes, 2012), early developmental stages (eggs and larvae), parts of specimen bodies (e.g. one leg) or semi-digested samples (e.g. gut contents) (Lindeque *et al.*, 2013) were limited due to the necessity of high level of expertise and the limitations of morphological keys, which were often effective only for a particular life stage (Hebert *et al.*, 2003a).

The taxonomic challenge posed by cryptic species has been recognized for nearly 300 years (Bickford et al., 2007), similar morphology presented between species may lead to wrong species identification. For example, Herbert and collaborators (2004) revealed that previously considered single species with a large distribution range were indeed several species with seemingly morphologically identical adults but different juveniles with preference for different resources. Also, a study developed with the polychaete *Eurythoe complanata* demonstrated that this species was previously considered like a cosmopolitan single species, presenting a great morphological similarity with a wide geographic distribution, and through molecular analysis demonstrated the existence of ambiguities and high levels of genetic divergence, after being categorized as cryptic species (Barroso *et al.*, 2010). Thus, the morphological taxonomic techniques may not reveal this "hidden" biodiversity and a significant proportion of diversity can be underestimated (Costa and Antunes, 2012). Therefore, the use of morphological approaches for routine species identification are complicated, it demands time and expertise across different phyla (Corell and Rodríguez-Ezpeleta, 2014). The advent of molecular techniques has given biologists a new tool for detecting biodiversity to overcome this operational constraints.

Over the past years several approaches has been developed to utilize DNA-based species identification (Lindeque *et al.*, 2013). In 1980, methods were proposed for species identification based on DNA hybridization (Southern Blots and RFLP). Few years later, studies using DNA-based species identification significantly expanded by PCR-based amplification of DNA and the design of primers (Taberlet *et al.*, 2012). DNA-based approaches revealed to be a source of information that allows access to biodiversity beyond morphology. This approach also demonstrated to be a tool which enables species identification to non-taxonomic experts. However these proposed methods also have disadvantages as expensive, time consuming and fail in the detection of taxa present in low abundance (Costa and Antunes, 2012).

In 1977 the Sanger sequencing emerged and enabled to recover sequence data from single specimen at a time (Sanger *et al.*, 1977). The advent of Sanger DNA sequencing technology allowed the application of genomic approaches to taxon diagnosis using DNA sequences to identify organisms. Furthermore, Sanger sequencing led to large-scale, broad-scope biosystematics projects with a wide range of applications (Shokralla *et al.*, 2012).

More recently, in 2003, developments in DNA technology led to a complement of taxonomy through the use of new genomic approaches for taxon diagnosis to identify species (Blaxter *et al.*, 2005; Costa and Antunes, 2012). Paul Hebert and colleagues developed DNA barcoding approach. They used a relatively short sequence (i.e. approximately 650 bp) of a eukaryotic genome standardized zone (e.g. COI), named as DNA barcodes, as a molecular tag to generate vast DNA libraries for species identification in many taxa. In DNA barcoding approach, after DNA extraction is necessary to perform DNA amplification with barcoding primers and then proceed to sequencing. Finally, a sequence-based taxonomic identification via standard reference databases of known organisms is performed (Hebert *et al.*, 2003a). The primers designed to DNA barcoding are versatile primers that are used in PCR amplification based on a single barcode within a short variable DNA region, target the same locus, and applied to different taxa found universally across diverse phyla (Lobo *et al.*, 2013; Taberlet *et al.*, 2012). The use of these primers is huge importance to barcoding success in species identification, in order to have high resolution of taxonomic discrimination to improve the efficiency of taxon detection (Leray *et al.*, 2013). Therefore, this method intended to facilitate and increase the biodiversity discovery, in order to transform our ability of species identification in a practical and objective approach (Costa and Antunes, 2012).

Thereby, DNA barcoding approach shows to be universal (the same pattern to all organisms), rapid, rigorous, objective and practical. This molecular tool can improve conventional approaches limitations by allowing species identification in any stage of the life cycle and in analysis of gut contents and excreta (Hebert *et al.*, 2003a). The emergence of these technology also help the resolution of the taxonomic impediment with the ability to faster a practical catalogue and describe biological diversity (Costa and Antunes, 2012; Teletchea, 2010). Moreover, further examination of divergent taxa can now allow the detection of morphological, ecological and behavioral differences (Lobo *et al.*, 2015), going beyond the taxonomy. DNA barcoding has a broad scientific applications, such as in conservation biology, which can catalyze many studies with an interconnection between

different groups of taxonomists, in wide target taxa (Stoeckle, 2003). Research projects on birds (Hebert *et al.*, 2004), fish (Costa *et al.*, 2012), algae (Le Gall and Saunders, 2010), benthic macroinvertebrates (Costa *et al.*, 2007), macrofauna (Knox *et al.*, 2012), meiofauna (Fonseca *et al.*, 2010) and others taxonomic groups has been performed.

The application of DNA barcoding approach should take into account certain criteria in order to improve the limitations of morphologic identification. In DNA extraction the resistance to DNAse digestion can be a problem. In environmental samples, the extracellular DNA is adsorbed contrary to free DNA leading to the exchange of cell lysis step by a saturated phosphate buffer (Taberlet *et al.*, 2012). Furthermore, the presence of additional taxa or decaying organic matter in sample can inhibit PCR and sequence reactions (Creer *et al.*, 2010). The species characteristics are also important factor for achieving an efficient DNA extraction. For example, the molluscs are an important group of organisms which are challenging to perform DNA extraction due to the high amount of mucopolysaccharides in their tissues that inhibit polymerase activity (Barco *et al.*, 2015).

The efficiency of the PCR amplification protocol is a critical step for barcoding successfully studies because they can introduce biases during amplification. The formation of PCR-induced chimeras is one of the most commonly source of sequence artifacts. Chimeras are produced when incomplete extension occurs during PCR amplification and the resulting amplicon fragments acts as a primer for a different sequence, leading to occurrence of false diversity estimates (Bik *et al.*, 2012; Corell and Rodríguez-Ezpeleta, 2014; Creer *et al.*, 2010; Fonseca *et al.*, 2012). These negative effects can be minimized through PCR optimization and bioinformatics software developments (e.g. Perseus, Quince *et al.*, 2011) (Fonseca *et al.*, 2012). When PCR reaches the plateau phase, drive by the use of PCR cycles with a fast ramping rate, heteroduplex formation can occur which give artificial gene diversity (Kurata *et al.*, 2004). The coamplification of divergent heteroplasmic copies of mitochondrial DNA can overestimate the number of unique species, introducing biases (Song *et al.*, 2008). The annealing temperature, by reducing at lower temperatures (Ishii and Fukui, 2001) and the number of replication cycles, by keeping low the number of cycles (Qiu *et al.*, 2001) are important parameters to reduce bias of primer binding. Also, the use of high template concentrations, intelligent primer selection and mixed replicate reaction preparations can be reduce the PCR-induced biases (Shokralla *et al.*, 2012).

The sequenced gene region should be identical between specimens but different between species. Furthermore, the ideal gene target must be sufficiently conserved to be amplified with broad-range primers (Stoeckle, 2003). Many different nuclear and organellar DNA regions can be targeted for DNA amplification and sequencing (Taberlet *et al.*, 2012). The genetic markers that can be used are the nuclear ribosomal RNA gene (12S, 16S and 18S), nuclear gene ITS (internal transcriber spacer), chloroplast genes *matK* (maturase K) and *rbcl* (ribulose-biphosphate carboxylase), and the mitochondrial gene COI (cytochrome c oxidase subunit I) (Stoeckle, 2003). The 16S is commonly used in studies of bacteria identification (e.g. Sogin *et al.*, 2006) (Shokralla *et al.*, 2012). Fungi contain introns in mitochondrial gene, however applying reverse transcription in conjunction with PCR, ITS can be used for identification of fungi species (e.g. Nilsson *et al.*, 2008; Seifert, 2009) (Begerow *et al.*, 2010). In plants, *matK*, *rbcl* and ITS can used to target for barcoding, due to the low sequence variation in mitochondrial DNA of plants (e.g. CBOL Plant Working Group, 2009) (Stoeckle, 2003). The COI and 18S are widely applicable in animal barcoding (e.g. Folmer *et al.*, 1994; Fonseca *et al.*, 2014) (Corell and Rodríguez-Ezpeleta, 2014). Besides the fact that is important to have consensus for universal barcodes, sometimes flexibility is needed in the marker choice. In nematodes, studies recognized that COI is inappropriate due to sequence diversity (Deagle *et al.*, 2014). Also, there are similar problems for plant barcodes, due to the low level of variability and low variation in phylogenetic markers (e.g. Cho *et al.*, 2004) (Chase *et al.*, 2005).

The analysis of DNA barcode sequences involves three important steps. The first step is the sequence alignment to compare corresponding loci and the second is the construction of phylogenetic trees, using clustering methods such as Neighbor Joining (NJ) method (Saitou and Nei, 1987), to evaluate genetic distances among species (La Rosa *et al.*, 2013). The last step is processing data generated by DNA sequencing approaches to make different analyses. The barcode-based identifications of unknown organisms relies on the ability to match a given sequence to a library of reference barcodes based on known species (Hajibabaei *et al.*, 2011). Recently diverged species or the appearance of new species, through hybridization, difficult sequence-based species identification due to the intraspecific and interspecific genetic variation, which differ between groups of species (Stoeckle, 2003). The ability to quantify the absolute abundance of individuals based on sequence read counts is sometimes a problem. The variation in the number of target gene copies between species, the number of target organelles per individual and the

variation in tissue cell density makes impossible species identification from sequence read data (Aylagas *et al.*, 2014; Bik *et al.*, 2012). Adopting bioinformatics approaches, by using recovering sequences to operational taxonomic units (OTU), can reduce the barcoding inefficient caused by the large magnitude of taxonomic coverage (Creer *et al.*, 2010; Deagle *et al.*, 2014). Therefore, the analysis of molecular data is only based on the presence/absence of taxa.

DNA barcoding approach has some disadvantages. As referred above, amplification of nuclear copies of DNA mitochondrial and chloroplastidial fragments (Song *et al.*, 2008), chimeras (Fonseca *et al.*, 2012) or heteroduplex formation (Kurata *et al.*, 2004) are examples of limitations that can lead to misidentification and, consequently, statistical problems. Furthermore, the use of single-locus for preliminary barcode-based species delineation can lead to complications, such as incomplete lineage sorting. In these cases, the analysis of single-locus data, should be considered as OTU (Kekkonen and Hebert, 2014). OTU are clusters of species which allows in taxa identification through sequence identity (Bik *et al.*, 2012; Blaxter *et al.*, 2005). The Barcode Index Number System (BIN) is an analytical method that apply clustering algorithms creating a structured registry for OTU recognition, and sequences are automatically assigned to a BIN on the BOLD Workbench (http://www.boldsystems.org/). Considering that each specimen has one assigned name, creating a global exclusivity of names, the objectivity of DNA barcoding studies increase (Ratnasingham and Hebert, 2013).

Technological advances in taxonomy are not the solution to species identification problems. Contrariwise, the complementation of conventional approaches with DNA barcoding can have impact on the scientific community and enhance the species discovery (Costa and Antunes, 2012).

Table 1 Applications of DNA (meta)barcoding approach for various loci and a broad range of organisms, compiling different studies.

Key-applications	Description	Reference
	18S	
Marine metazoan communities; HTS	Analysis of links between ecosystem structure and function and phyletic diversity of meiofuanal communities	Fonseca *et al.*, 2010
Zooplankton; HTS	Study of diversity and species richness of zooplankton communities	Lindeque *et al.*, 2013
Meiofauna; HTS	Macroecology studies of meiofaunal communities and evaluation of diversity levels	Fonseca *et al.*, 2014
Marine metazoan communities; Biomonitoring; HTS	Evaluation of the quality of marine benthic ecosystems by comparing morphological and eDNA/RNA-based inventories	Lejzerowicz *et al.*, 2015
	COI	
Invertebrate phyla; "Universal" primers design	"Universal" primers design to amplify COI gene from metazoan invertebrates	Folmer *et al.*, 1994
DNA barcoding approach	Development of DNA barcoding approach, based on COI gene, for species-level assessment and identification	Hebert *et al.*, 2003a
Birds	Identification of birds species and determination of intra- and interspecific differences	Hebert *et al.*, 2004
Lepidoptera; Cryptic-species	Identification of *Astraptes fulgetor* butterfly, with the combination of morphological and molecular tools	Hebert *et al.*, 2004
Moth; Wasp; Mini-barcodes sequences	Identification of moth and wasp museum species using short barcode sequences	Hajibabaei *et al.*, 2006
Ciliate protozoa	Species identification and variability studies of *Tetrahymena thermophila* species	Lynn and Strüder-Kypke, 2006

Crustacea	Identification of Crustacea at order- and species-level	Costa *et al.*, 2007
Holozooplankton; Biomonitoring;	Identification and recognition of holozooplankton species	Bucklin *et al.*, 2010
Benthic macroinvertebrate communities; biomonitoring;	Biomonitoring of freshwater benthic macroinvertebrate taxa	Hajibabaei *et al.*, 2011
Benthic macroinvertebrates; Non-destructive source of DNA; Multiplex PCR strategy; HTS	Evaluation the ability of non-destructive DNA access and a multiplex PCR approach for biodiversity analysis of benthic macroinvertebrates	Hajibabaei *et al.*, 2012
Macrofauna of deep-sea; MOTUs	Quantification and comparison diversity of macrofauna of deep-sea habitats	Knox *et al.*, 2012
Soil extracellular DNA;	New sampling and extraction protocols for DNA metabarcoding analyses of soil extracellular DNA	Taberlet *et al.*, 2012
Arthropods; Biomonitoring; Biodiversity assessment; HTS	Detection of arthropod taxa and estimation of diversity metrics	Yu *et al.*, 2012
Top-shells (gastropods)	Identification of gastropods (*Gibbula* sp.) providing a consistent data set of COI sequences	Barco *et al.*, 2013
Marine invertebrates; Newly primers design	Redesign of PCR Folmer primers: jgLCO1490/jgHCO2198 for amplification of COI gene of marine invertebrates	Geller *et al.*, 2013
Marine metazoan; Newly primer design	mlCOIintF primer design and combination with jgHCO2198 for amplification of COI gene of marine metazoan diversity	Leray *et al.*, 2013
Marine metazoan communities; Newly primers design	LoboR/F primers design for amplification of COI-5P gene of marine metazoan species	Lobo *et al.*, 2013
Copepods	Identification of marine copepods and reliability and resolution analysis of statistical approaches	Blanco-Bercial *et al.*, 2014
Lepidoptera; HTS	Application of HTS technologies for parallel acquisition of DNA barcodes from 190 specimens simultaneously	Shokralla *et al.*, 2014

Polychaeta	Evaluation of the performance of DNA barcodes in discrimination of polychaete	Lobo *et al.*, 2015
	Multi loci	
Marine invertebrates; Gut contents	Study of macrophagous and microphagous diet. Amplification of COI for analysis of animals ingested and 18S for analysis of lesser eukaryotes ingested	Blankenship and Yayanos, 2005
Nematode communities; HTS	Identification and diversity assessment of nematode species, amplifying small and large subunit of rRNA	Porazinska *et al.*, 2009
Aquatic macroinvertebrate communities; Biomonitoring; HTS	Species-level identification, based on COI and Cytochrome B of mtDNA, to diagnostic biomotoring of aquatic ecosystem	Carew *et al.*, 2013
Marine macroinvertebrates; Biomonitoring	Presence/Absence species evaluation using genetics based AMBI to amplify COI and 18S gene	Ayalagas *et al.*, 2014
Zooplankton; DNA extraction	Alternative DNA extraction protocol for metabarcoding analysis, based on 18S and COI, on zooplankton communities	Corell and Rodríguez-Ezpeleta, 2014
Arthropod macrobiome; microbiome; HTS	Utilization of 16S, 18S and COI to test detection capacity of arthropods and microbiome from bulk sample	Gibson *et al.*, 2014
Seagrass communities; Invertebrate communities; Biomonitoring	Identification and diversity estimation of invertebrate taxa associated with seagrass communities by comparing morphological and molecular inventories (based on COI and 18S)	Cowart *et al.*, 2015

The DNA barcode impact on life cataloging emerge global project focus on a wide range of species. The Barcode of Life Initiative (BOLI) began with the proposal of the DNA barcoding approach (2003). DNA barcodes are used to access biodiversity information, and consequently to build a new system for species identification – an open access database of reference barcodes (Costa and Antunes, 2012; Costa and Carvalho, 2010).

The Consortium for the Barcoding of Life database (CBOL - http://www.barcodeoflife.org/) implement DNA barcoding to promote a global scale genomic project, such as Marine Barcode of Life (MarBOL - http://www.marinebarcoding.org/), collaborating with a variety of institutions (Costa and Antunes, 2012). At present, CBOL involve 200 Member Organizations from 50 countries, which promotes barcoding through research groups, networks, workshops, conferences and training. The CBOL aims explore and develop DNA barcoding potential to species identification through the link of CBOL's taxonomic data to publicly accessible sequences and the development of barcoding to make it more cheaper, faster and portable (Deagle *et al.*, 2014). Actually, the public access to DNA barcoding data are possible on Barcode of Life Database (BOLD - http://www.boldsystems.org/), GenBank of National Center for Biotechnology Information (NCBI - http://www.ncbi.nlm.nih.gov/genbank/), European Molecular Biology Laboratory (EMBL - http://www.embl.org/) and DNA Data Bank of Japan (DDBJ - http://www.ddbj.nig.ac.jp/). BOLD database has allowed an improvement in taxonomic identification through providing barcode sequences and their association to other taxonomic data (e.g. geolocation data). To avoid the conflicting and dispersal data among databases, informatics tools allow databases collaboration, such as World Register of Marine Species (WoRMS, Worms Editorial Board, http://www.marinespecies.org/).

In 2010 was launched the International Barcode of Life project (iBOL - http://www.ibolproject.org/). This is a global project that use DNA barcodes as a tool for identifying known species and discover new ones in order to apply in such areas: forensics, conservation, diseases control and ecosystem monitory (Taberlet *et al.*, 2012). The aim of the project is barcode a five million specimens, in order to construct a parameterized DNA barcode reference library for 500 000 eukaryotic species until 2015.

The contribution of DNA barcoding to technological, organizational and conceptual developments lead to improved taxonomy and to discover new species, without need of

morphological descriptions, increasing the capacity of efficiently manage ecosystems and monitory and recognize biodiversity (Costa and Antunes, 2012). Furthermore, the genetic techniques generated are cheaper, faster and more accurate taxonomic identification (Corell and Rodríguez-Ezpeleta, 2014).

1.2 DNA metabarcoding

Identification of multiple species, in a single experiment, from a single complex environmental sample is an extension of the barcoding concept and has been referred as DNA metabarcoding (Taberlet *et al.*, 2012; Taberlet *et al.*, 2012). DNA metabarcoding overcomes standardized DNA barcoding difficulties: identification of single specimens, DNA needs to be more or less intact and requires the isolation of specimens, which is time consuming and difficult. Also, the products obtained from DNA barcoding are generally sequenced by Sanger method, while in metabarcoding the mixed products are sequenced by high throughput sequencing technologies (Corell and Rodríguez-Ezpeleta, 2014). Therefore, the goal of DNA metabarcoding is identify taxa at species level, using a large number of samples (Taberlet *et al.*, 2012).

Using high-throughput sequencing (HTS) in metabarcoding studies, a single bulk sample containing the entire organisms of an environmental community can be analyzed (Taberlet *et al.*, 2012). Furthermore, this sample can also include degraded DNA (such as soil, water, faeces or originates from cell lysis) (Taberlet *et al.*, 2012). Comparing microbiota in healthy and disease individuals (Chen *et al.*, 2014), inferring ecosystem healthy (Hajibabaei *et al.*, 2011), study ancient DNA (Sønstebø *et al.*, 2010) or analyze diets from DNA fragments (Deagle *et al.*, 2009) are some examples of HTS applications.

Since 2005, the appearance of HTS has been improvements in sequencing output, decreasing the costs and time consuming and reducing sources of PCR bias (Mardis, 2008; Shendure and Ji, 2008), enabling the utilization of HTS in a variety of applications. Access to massive amounts of sequencing data and improvements in read length leading to a better representation of sample diversity (Shokralla *et al.*, 2012). For example, Sogin and colleagues (2006) using 16S as specific gene marker and applying HTS approach were able to analyze DNA sequence data from marine microbial community.

The available HTS technologies can be classified into two categories: PCR-based technologies and single-molecule sequencing (Shokralla *et al.*, 2012). The commonly used HTS platforms for PCR-based technologies as, for example, Roche 454 Genome Sequencer (Roche Diagnostics Corp., Branford, CT, USA) or HiSeq 2000 (Illumina Inc., San Diego, CA, USA). The HeliScope (Helicos BioSciences Corp-, Cambridge, MA, USA) or PacBio RS SMRT system (Pacific Biosciences, Menlo Park, CA, USA) are systems used for single-molecule sequencing (Shendure and Ji, 2008). The rapid progress on HTS technologies led to the emergence of various sequencing systems. Due this, depending on the ecological research platforms should be appropriate (Shokralla *et al.*, 2012).

The 454 Genome Sequences (www.454.com) was the first HTS technology which allowed sequencing 400-600 million bp per run with 400-500 bp sequence lengths in a single experiment by using real-time sequencing-by-synthesis pyrosequencing technology, increasing the sequencing capacity (Costa and Antunes, 2012). This is more five orders of magnitude than in traditional Sanger sequencing (Taberlet *et al.*, 2012). In this technique (Figure 1), after DNA amplification from environmental samples, the DNA fragments are bound to beads, one fragment per bead (1:1 proportion), the beads are isolated individually and occurs emulsion PCR (oil micro-reactors that contain PCR components). The emulsion is broken, the DNA strands are denaturated and beads are individually deposited into well of fiber-optic slide. Beads carrying immobilized enzymes are sequenced and deposited into each well. The complementary strand is synthetized enzymatically to detect which base is added at each step. One of the four dNTPs (deoxynucleotides) is added to DNA and DNA polymerase incorporates the complementary to template. This incorporation releases PPI stoichiometrically. Then, ATP sulfurylase converts PPI (pyrophosphate) to ATP (adenosine triphosphate) acting as fuel to mediate the conversion of luciferin to oxyluciferin. This reaction generate visible light which is detected by a camera and analyzed in a program. Finally the reaction can start again with another nucleotide and ends when DNA sequence of the single stranded template is determined (Rothberg and Leamon, 2008).

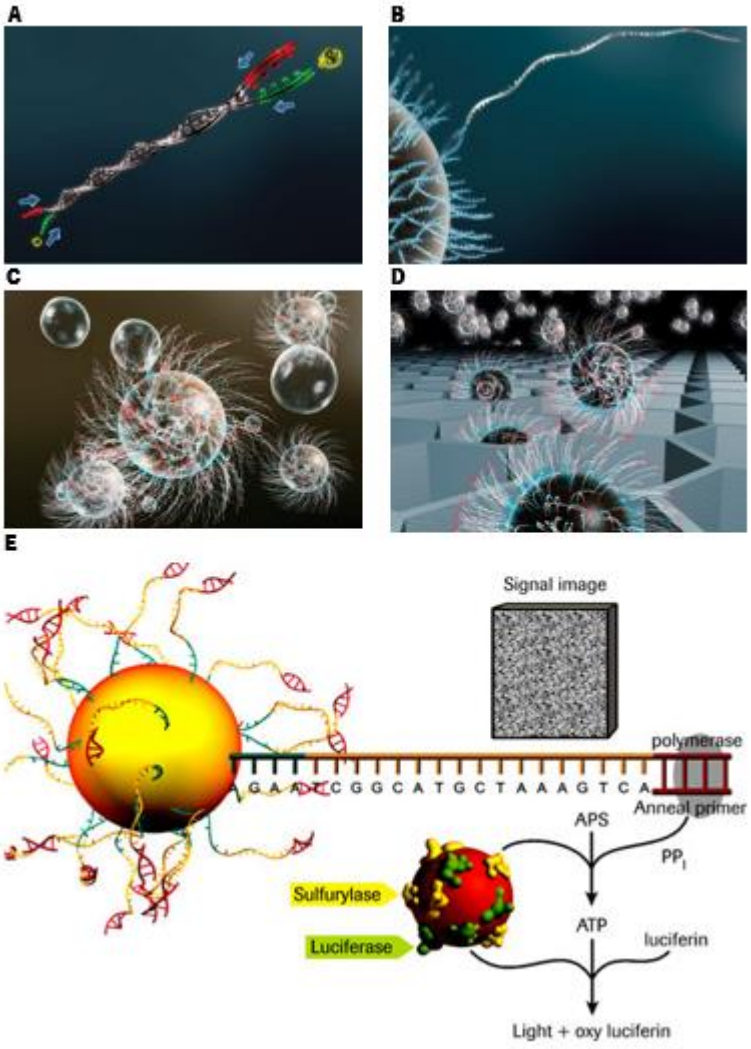

Figure 1 Overview of the 454 sequencing technology. A – Library preparation. B – Fragments bound to beads (1:1). C – Emulsion PCR amplification. D – Load the beads onto the PicoTiterPlate device (1:1). E – Pyrosequencing reaction of 454 Sequencing Systems. Adapted by http://454.com/.

The main advantages on the use of HTS-metabarcoding approaches are the long read length produced in a relatively short time, capability to apply bioinformatic tools and the low chances of premature chain termination and non-simultaneous extension (Hajibabaei *et al.*, 2011). Furthermore, due to determination of taxon detection and identification efficiency, the success of this approach relies on the primer sets used and the target loci (Leray *et al.*, 2013).

The genetic markers that can be used to DNA metabarcoding studies are the same as used in barcoding (referred in 1.1 section). The past taxonomic analysis is focused on nuclear genes, especially in 18S. Developed studies demonstrated that this nuclear region have a prevalence of insertions, which can introduce bias during PCR amplification, deletions, that can complicate

sequence alignments, and reported problems associated to recombination (Hebert *et al.*, 2003a; Stoeckle, 2003). Furthermore, morphology-based identification and DNA metabarcoding approach rely on 18S gene, using meiofaunal taxa, showed an underestimation of species diversity relative to COI (Tang *et al.*, 2012). Contrariwise, some of these limitations are not present in mitochondrial genome (Figure 2).

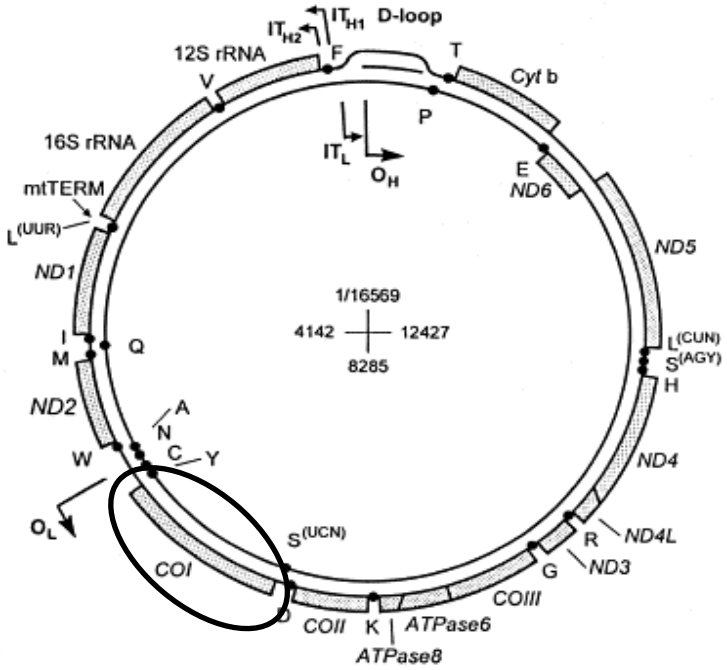

Figure 2 Map of the human mitochondrial genome (16 569 bp). The black circle highlights the COI gene. Taanman, 1999.

The COI gene lack of introns, has limited exposure to recombination and has haploid mode of inheritance (Hebert *et al.*, 2003a). The 648 bp length of COI are short enough to be sequenced quickly and cheaply and are able to identify at species-level. Also, considering that amino acid sequence changes occur more slowly in COI, this gene marker is more likely to provide deeper phylogenetic insights than alternatives (e.g. cytochrome b) (Hebert *et al.*, 2003a). Due to the evolution of mitochondrial gene, COI are able to discriminate closely allied species and phylogeographic groups within a single species (Cox and Hebert, 2001). A study using *Tetrahymena thermophila* species (ciliate species) demonstrated that species can be identified based on COI gene, revealing high degree of precision (Lynn and Strüder-Kypke, 2006). Hajibabaei and collaborators (2012) used HTS metabarcoding approach to access biodiversity of benthic macroinvertebrate community. HTS approach demonstrated to be effective in environmental studies, increasing the potential of using DNA information (Hajibabaei *et al.*, 2012). In other study,

HTS metabarcoding approach was applied to test the efficacy of COI-pyrosequencing in the detection of arthropods and microbiome from a bulk sample. The authors confirmed that this approach provides biodiversity assessment and environmental monitoring (Gibson *et al.*, 2014). The associated databases to COI gene region have boasts millions of taxonomically verified DNA sequences, which not verified with 18S gene region. Because DNA metabarcoding taxonomic identification is performed by sequence-based identification, the existence of a standard reference library of known organisms is the most important requirement to biodiversity assessment (Aylagas *et al.*, 2014). Therefore, nuclear genome has limitations when compared to mitochondrial genome and as a result the standard mitochondrial DNA barcode region are effective for species identification.

Efficient PCR primers of broad taxonomic scope are fundamental in DNA barcoding research to allow amplification of the same locus across a wide range of taxa from different phyla, with the same efficiency (Lobo *et al.*, 2013). Finding a unique suitable metabarcode within a short variable DNA region to target multiple species on an environmental sample, flanked by two highly conserved regions, (about 20 bp) is a difficult task (Taberlet *et al.*, 2012). A large number of primers have been design for COI amplification from various animal groups. Folmer and colleagues (1994) designed the first "universal" primers, called LCO1490 and HCO2198 ("Folmer primers"), to amplify 658 bp fragments of the COI gene in a broad range of marine metazoan phyla. However, these primers often fail or perform poorly for many taxa (Blankenship *et al.*, 2005; Lohman *et al.*, 2009). The limited amplification success of Folmer primers are possibly related to mismatches occurring in the target annealing position, this led some authors to develop new primers with some level of degeneracy, this is created during primer synthesis by mixing nucleotides at the variable sites, thereby creating a pool of primers containing all variants (Geller *et al.*, 2013; Leray *et al.*, 2013; Lobo *et al.*, 2013). In 2013, Geller and colleagues redesigned "Folmer" primers using degenerate positions and internal inosines. The use of inosines is useful because it can pair with any natural base (adenine, thymine or cytosine), without disrupt the primer's annealing efficiency. The new jgLCO1490 and jgHCO2198 (658 bp) primers showed to be broadly applicable and complement the standard Folmer primers in DNA barcoding applications. Lobo and collaborators (2013) designed new enhanced primers, LoboF and LoboR (658 bp), for COI-5P barcode region to overcome the limitations of Folmer primers, especially in marine invertebrates identification. The primers have a high success rate of amplification of COI-5P gene and revealed to be rapid, practical

and cost-effective (Lobo *et al.*, 2013). The forward primer mlCOIintF were designed within the COI region by Leray and collaborators (2013). In a study using coral reef fish gut contents, they combined mlCOIintF with jgHCO2198 (313 bp) and reported a higher success than using Folmer primers (Leray *et al.*, 2013). Recently, Gibson and colleagues (2014) used HTS and multiple primers sets primers, including the combination ArF2 and ArR5 used in this study to maximize recovery of the arthropod macrobiome and the bacterial and other microbial microbiome of a bulk arthropod sample.

Another limitation to the use of the full length of barcode region is their application on the recovery of museum specimens, since the DNA is often degraded. Short sequences (≈100 bp) can regularly be obtained from old specimens and a new approach based on "mini-barcodes" was developed to identify unknown specimens (e.g. Fishes and Lepidoptera in Hajibabaei *et al.*, 2006) (Meusnier *et al.*, 2008). However, mini-barcode primers demonstrated a limited efficiency for DNA amplification from some taxa (Arif *et al.*, 2011).

PCR amplification can introduce some sources of PCR bias, such as chimeric sequences formation. However, in metabarcoding, amplification failures of a particular taxa are not subject to optimization. These occurs because specimens that initially failed in amplification are masked by the detection of amplicons from other taxa present in the sample. Reference library preparation, detection of the incorporated nucleotides and utilization of primer cocktails can minimize these effects and increase amplification success rates (Shokralla *et al.*, 2012). Therefore, primers designed for COI DNA barcode region has proved to be very robust, allowing routine detection of species segments of COI, and enabling amplification of most animal phyla (Stoeckle, 2003).

Biomonitoring programs, through the employment of biotic surveys, are essential to assess information about species composition, biodiversity changes and ecosystem status and trends (Hajibabaei *et al.*, 2011). Benthic macroinvertebrates communities are routinely used as bioindicators to detect environmental disturbances in aquatic ecosystems. These communities display some of the highest diversity on Earth, yet there is a well-knowledge gap in understanding of their global biodiversity. Only 1% of their biodiversity are estimated to be known (Fonseca *et al.*, 2010). Furthermore, due to the broad taxonomic diversity and a lack of consistently approaches (e.g. efficient primers), macrobenthic communities have been hard to identify (Lobo *et al.*, 2013). Also, these communities contain development stages (e.g. eggs), cryptic species and associated

gut contents which difficult species identification (Leray *et al.*, 2013). The bioassessment of macrobenthic fauna can be improved by novel approaches that significantly speed-up benthic macroinvertebrate monitoring, which is traditionally time-consuming undertaking (Baird and Sweeney, 2011). This is especially important under the European Union's Water Framework Directive (WFD). The WFD was developed to implement an aquatic ecosystem-monitoring network, which commits European Union member states to achieve good qualitative and quantitative status of all water bodies by 2015. A classification for ecologic status (high, good, moderate, poor and bad) in order to define the ecologic and chemical status of aquatic bodies (Costa and Antunes, 2012).

1.3 Aim of the thesis

The main objective of the present work was to prime the development of a DNA metabarcoding methodology for routine species identification and inventory in marine macrobenthic communities, with particular focus on estuaries and coastal areas. In order to attain this objective, the partial goals and associated tasks were:

- To compile a reference library of cytochrome oxidase I DNA barcodes of estuarine and coastal marine invertebrates from Portugal to be used as a central framework for sequenced-based species identification through metabarcoding approaches. The reference library shall include dominant member of the three main marine phyla represented in macrobenthic communities, namely Annelida, Crustacea and Mollusca.
- To evaluate the effect of the amplicon size, and location within the COI-5P barcode region, on the sequenced-based species discrimination ability. For this purpose we carried out a structured *in silico* analysis based on the sequential pruning of the reference library in multiple fragments of different size. This *in silico* analysis was required because the metabarcoding approach typically uses shorter sequences than the full COI-5P barcode region.
- To investigate the ability of different primer sets to amplify, and therefore enable detection, of the diversity of species present in a macrobenthic assemblage of known species composition and abundance, through the use of experimentally assembled communities.

This thesis was developed in the scope of the project BEstBarcode (PTDC/MAR/113435/2009), funded by Fundação para a Ciência e Tecnologia (FCT). Dr. C.

Hollatz executed the laboratory experiments here reported with the assistance of J. Lobo in primer design and preliminary tests. High-throughput sequencing (HTS) was carried out in Genoinseq, UC-Biotech (BioCant Park, Cantanhede, Portugal), under the supervision and support of Dr. C. Egas, together with Dr. H. Froufe in the upstream data treatment and analyses of HTS reads. The sequence data used in the reference library were compiled from published, submitted and unpublished projects led by the Molecular Ecology and Biodiversity group of CBMA, at University of Minho (Antunes *et al.*, 2015; Borges *et al.*, submitted; Gomes, 2014; Lobo *et al.*, 2013; Lobo *et al.*, 2015; Lobo *et al.*, unpublished). The thesis author, B. Leite, executed all the downstream data analyses and annotation, data interpretation and discussion.

This master's thesis is divided into 5 sections. Firstly, one proceeds to the historical context of the study through a general introduction of the topic of DNA metabarcoding. This also includes the objectives and the thesis structure. Secondly, there is an inventory of the materials and methods that were used for all experimental procedures. Lastly the results are presented, being followed by the discussion and the conclusion.

2. MATERIALS AND METHODS

2.1 Overview of the global approach and experimental design

The global experimental approach followed in this study is composed of three main stages. The first stage encompasses the reference library compilation of COI-5P DNA barcodes of marine invertebrates from mainland Portugal and Azores Islands, for sequenced-based species identification. The second stage comprises the evaluation of the amplicon size and location (within the COI-5P barcode), on the sequence-based species discrimination ability. Once defined the discrimination degree for different amplicons, the third stage is to test the species detection success in experimentally assembled macrobenthic communities whose COI-5P barcodes were amplified using 5 different sets of primer pairs. Two different simulated macrobenthic communities (SimCom) with known species composition were created, comprising a same number of species but a different number of specimens per species. Figure 3 provides an overview of the global approach and experimental design here followed.

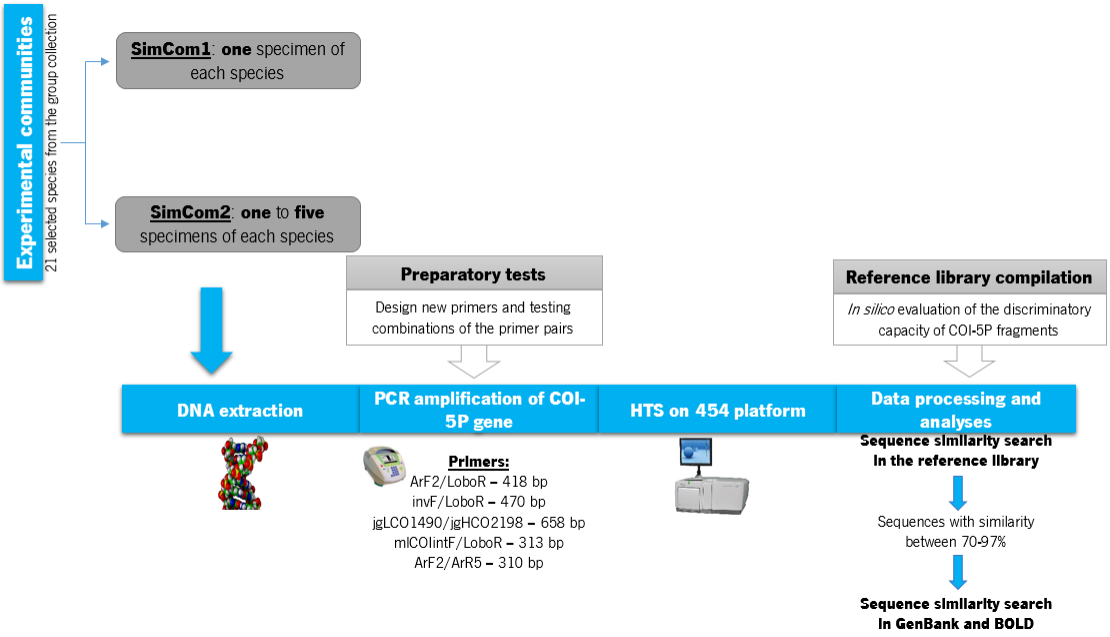

Figure 3 Schematic representation of the experimental design used in this study for testing the application of the metabarcoding approach to species identification in macrobenthic communities. SimCom1 – Simulated Community 1; SimCom2 – Simulated Community 2.

2.2 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. A total of 21 species were selected, in order to embrace the widest possible phylogenetic diversity within the three major phyla typically present in macrobenthic communities. The distribution of species per phyla was respectively 4.8% Annelida, 33.3% Arthropoda, and 61.9% 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 DNA extraction, each community containing different number of specimens per species (88 specimens in total). SimCom1 had one specimen of each species, while SimCom2 had one to five specimens of each species (Table 2). This approach aimed to test whether the relative abundance in the mixture affect the amplification success by the different primer sets.

Table 2 Taxonomic classification and distribution of the 21 marine macrobenthic species among the two different simulated macrobenthic communities. SimCom1 – Simulated Community 1; SimCom2 – Simulated Community 2; n – number of specimens per species.

Phylum	Class	Order	Species	SimCom1 (n)	SimCom2 (n)
Annelida	Polychaeta	Phyllodocida	*Hediste diversicolor* (O.F. Müller, 1776)	1	5
Arthropoda	Malacostraca	Amphipoda	*Apohyale prevostii* (Milne Edwards, 1830)	1	4
			Corophium multisetosum Stock, 1952	1	5
			Echinogammarus marinus (Leach, 1815)	1	4
			Melita palmata (Montagu, 1804)	1	3
		Isopoda	*Cyathura carinata* (Krøyer, 1847)	1	4
			Dynamene bidentata (Adams, 1800)	1	4
			Lekanesphaera rugicauda (Leach, 1814)	1	4
Mollusca	Bivalvia	Mytiloida	*Mytilus* Linnaeus, 1758	1	1
	Gastropoda	Archaeogastropoda	*Gibbula cineraria* (Linnaeus, 1758)	1	3
			Phorcus lineatus (da Costa, 1778)	1	3

	Docoglossa	*Patella aspera* Röding, 1798	1	2
		Patella vulgata Linnaeus, 1758	1	2
	Littorinimorpha	*Alvania mediolittoralis* Gofas, 1989	1	4
	Neogastropoda	*Nassarius incrassatus* (Strøm, 1768)	1	2
		Nassarius reticulatus (Linnaeus, 1758)	1	3
		Nucella lapillus (Linnaeus, 1758)	1	3
		Ocinebrina edwardsii (Payraudeau, 1826)	1	3
	Pulmonata	*Siphonaria pectinata* (Linnaeus, 1758)	1	2
Polyplacophora	Chitonida	*Acanthochitona crinita* (Pennant, 1777)	1	2
		Lepidochitona cinerea (Linnaeus, 1767)	1	4

2.3 DNA extraction

The pooled specimens of each of the two simulated macrobenthic communities were homogenized separately in a grinder and the resultant slurry was incubated at 56 °C to evaporate residual ethanol, for 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 DNA was extracted using E.Z.N.A. Mollusk DNA Kit (Omega Bio-tek), following manufacturer's instructions. After extractions, aliquots of DNA were pooled in a single microtube of 1.5 mL, representing each simulated community (500 µL total volume).

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

A preliminary assessment of the amplification success of a series primer pairs, including the newly designed by J. Lobo and other already published, was conducted using individual test specimens. Based on the results, five primer pair combinations, which amplify different fragments within COI barcode region, were selected for the metabarcoding tests (Table 3 A; Figure 4). The first PCR used the COI specific primers and the second PCR involved 454 fusion-tailed primers, with fusion primers containing the Roche-454 A and B titanium sequencing adapters. In the first

step, each PCR reactions contained 2.23 µL DNA template, 32.77 µL molecular biology grade water, 5 µL 10x Advantage Buffer SA, 2 µL dNTPs (5 mM), 2 µL forward primer (5 mM), 2 µL reverse primer (5 mM), 3 µL DMSO (6%) and 1 µL 50x Advantage2 Taq polymerase mix. The PCR thermal cycling conditions for each primer pair are displayed in Table 3 B.

Table 3 A - Primers used for PCR amplification of fragments of COI-5P gene from the two different simulated communities and B - PCR primer combinations and respective thermal cycling conditions for the five primer pairs.

A

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

B

Primer combinations	PCR conditions
ArF2/LoboR	94 °C 5' 94 °C 30'' | 46 °C 1' | 68 °C 1' 15x 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' 15x 68 °C 10' | 4°C ∞

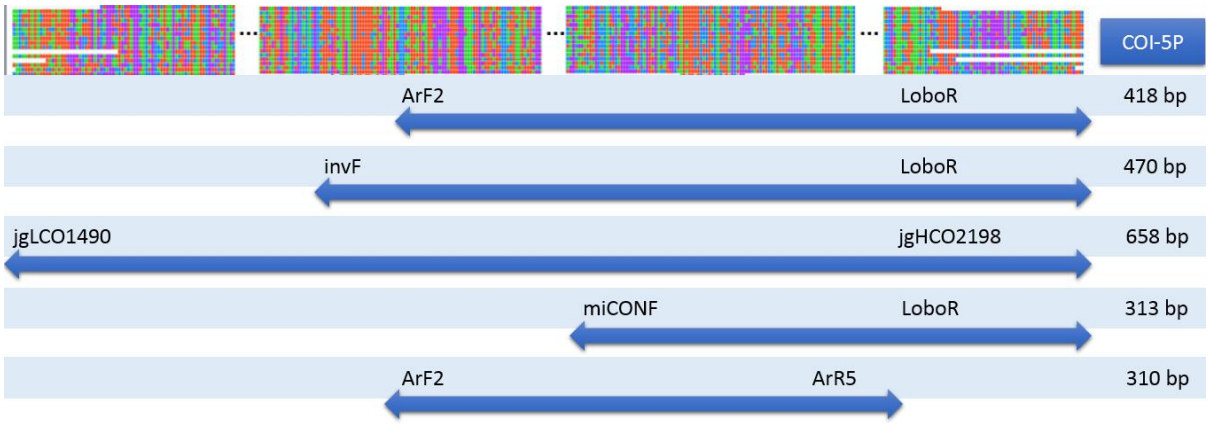

Figure 4 Schematic representation of the amplicons and their size, generated after PCR amplification. The COI-5P barcode and the five primer pairs that were used in PCR amplification within the standard barcode are represented.

The purified amplicons from the first PCR were used as templates in a second PCR with the same amplification condition used in the first PCR with the exception of using 454 fusion-tailed primers in a 30-cycle amplification regime. A negative control reaction (no DNA template) was included in all experiments. PCR success was checked by agarose gel electrophoresis.

2.5 High-throughput 454-pyrosequencing 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 GS 454 FLX Titanium chemistry, following the amplicon sequencing protocol provided by the supplier (Roche, 454 Life Sciences, Branford, CT, USA) at Biocant (Cantanhede, Portugal).

The DNA was fractionated and subsequently bound to beads in a 1:1 proportion to ensure only one fragment per bead. Each segment was amplified in microreactors formed by emulsion PCR. The beads with DNA were distributed over an optical fiber plate and then the sequencing occurs by synthesis (sequencing of a DNA single strand and then synthesizing its complementary strands enzymatically).

2.6 Data processing and analyses

2.6.1 Reference library compilation

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 315 barcode sequences of 300 taxa (species or genus), retrieved from private and public projects of the Molecular Ecology and Biodiversity Research Group (Antunes *et al.*, 2015; Borges *et al.*, submitted; Gomes, 2014; Lobo *et al.*, 2013; Lobo *et al.*, 2015; Lobo *et al.*, unpublished data) and comprising taxa from the three main marine phyla (Annelida, Arthropoda, Mollusca). Species are represented from one to four sequences, which were selected among the longest and of highest quality (absence of ambiguous bases) available and sequences displaying intraspecific distance above 2%. The sequences were aligned using the ClustalW 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 aminoacid patterns.

2.6.2 *In silico* evaluation of the discriminatory capacity of COI-5P fragments

Two *in silico* tests were carried out 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 on 158 pb of the 5' end, with 100 bp increments until 558 bp and then 658 bp. Second, all sequences of the reference library were clipped with the five primers pairs used in this study, with amplicon sizes of 310, 313, 418, 470 and 658 bp.

The Neighbor Joining (NJ) method was used to constructed phenograms (Saitou and Nei, 1987) in the program MEGA v.6.0, using the Kimura 2-parameter (K2P) substitution model (Kimura, 1980), the most used for analysis of DNA barcodes. Node support was assessed through 1000 bootstrap replicates. This provided a graphic representation of the divergence patterns among species allowing the visual inspection of clusters to determine the percentage of monophyletic clades. The monophyletic clades were evaluated in two different phases: (1) percentage of monophyletic clades with internal divergence higher than 3%; (2) percentage of

different species that were grouped in the same clade, in which case the genetic distance among species was verified using the, p-distance metric, calculated using MEGA v.6.0 program.

2.6.3 High-throughput data processing

The pyrosequencing reads (fasta files) were processed using an automated pipeline implemented at Genoinseq (Nex Gen Sequencing Unit, BioCant Park, Cantanhede, Portugal). The sequencing reads were assigned to the appropriate sample libraries (separately by primer and SimCom tested) based on the respective sequencing tags. To minimize the effects of random sequencing errors the sequencing reads were initially checked for quality and filtered (elimination of the sequence reads with less than 150 bp and the sequences that contained more than two undetermined nucleotides). Still at BioCant, the filtered reads obtained for each community were aligned against a reference library using the Usearch 6.1 software (Edgar, 2010). Finally sequence similarity searches at 97% minimum identity were performed against the reference library to assign a primary taxonomic identification.

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. The BOLD-IDS for COI accepts sequences from the 5' region of the mitochondrial cytochrome c oxidase subunit I gene and returns a species-level identification when one is possible (Ratnasingham and Hebert, 2007). GenBank® (http://www.ncbi.nlm.nih.gov) is a comprehensive database that contains publicly available nucleotide sequences for formally described species (Benson *et al.*, 2013). GenBank data retrieval is possible, for example, through the use of "The Basic Local Alignment Search Tool (BLAST)", which finds regions of local similarity between sequences. The program compares nucleotide (BLASTn) or protein sequences (BLASTp) to sequence databases and calculates the statistical significance of matches (Altschul *et al.*, 1990). Only matches > 97% similarity were considered for taxon identification in this analysis.

3. RESULTS

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

The reference library encompasses 315 sequences of marine and estuarine macrobenthic specimens, representing 266 taxa. The distribution of barcode sequences across the three main marine phyla, were: Annelida (19.68%), Arthropoda (60.32%) and Mollusca (16.51%). Other phyla with minor representations (< 4%) in the library were: Chordata (1.90%), Cnidaria (0.32%), Echinodermata (0.95%) and Nermetea (0.32%) (Figure 5 A).

The vast majority of the COI-5P barcodes included in the reference library were identified to species (266) but some were only to genus (34) or family (15) level only (Figure 5 B).

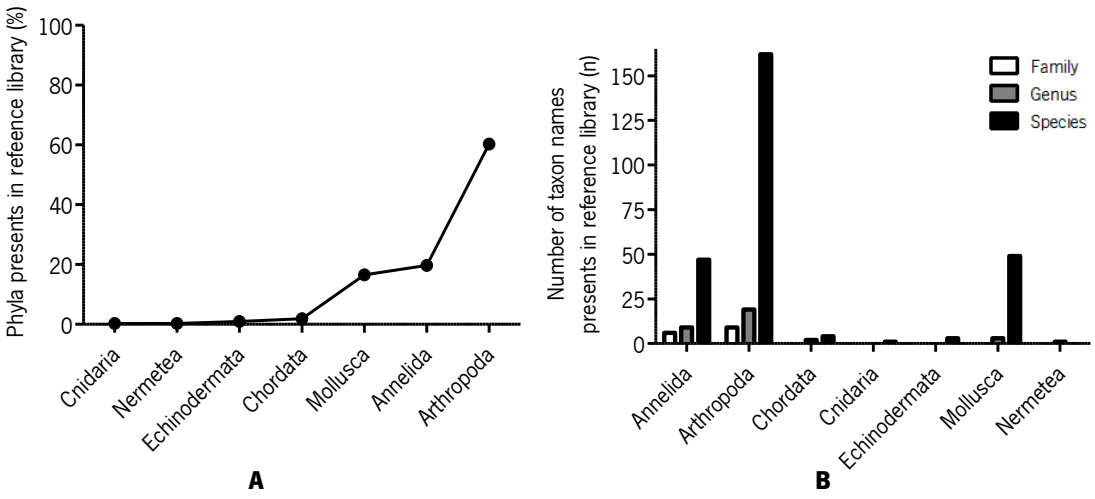

Figure 5 Number of specimens per phyla (A) and number of taxon names present in COI-5P reference library for seven representative phyla (B).

The NJ tree showed that regarding the fragment size, almost all species in the reference library were separated similarly in distinct clusters. Although minor shifts on the clade distances were noted among the different amplicons, as illustrated by Figure 6.

It is important to note that some species that were not previously resolved using the full COI-5P barcode region kept the same clustering pattern when compared to the other amplicons. In the full COI-5P barcode region 1.13% of the 266 total species were not distinguished, grouped in the

same clade with divergence lower than 3% (e.g. *Mytilus galloprovincialis*, *Mytilus edulis* and *Mytilus sp.*). The same clustering pattern were observed in three amplicons A (418 bp), B (470 bp) and E (310 bp), while in amplicon D (313 bp) 1.50% of the 266 total species were not distinguished (Figure 7). In this last amplicon there were three more cases of species not resolved.

The full length of the barcode region was divided into multiple fragments. Between the 258 bp and the 658 bp no reduction in the species discrimination was detected due to the fragment size reduction, i. e. the inconsistences that appear using the full COI-5P barcode region kept the same when compared to smaller fragments. However, in the minor fragment, 158 bp, two more cases were observed. This resulted in an increase up to 1.50% of the species that were not distinguished, and which grouped in the same clade with divergences lower than 3%.

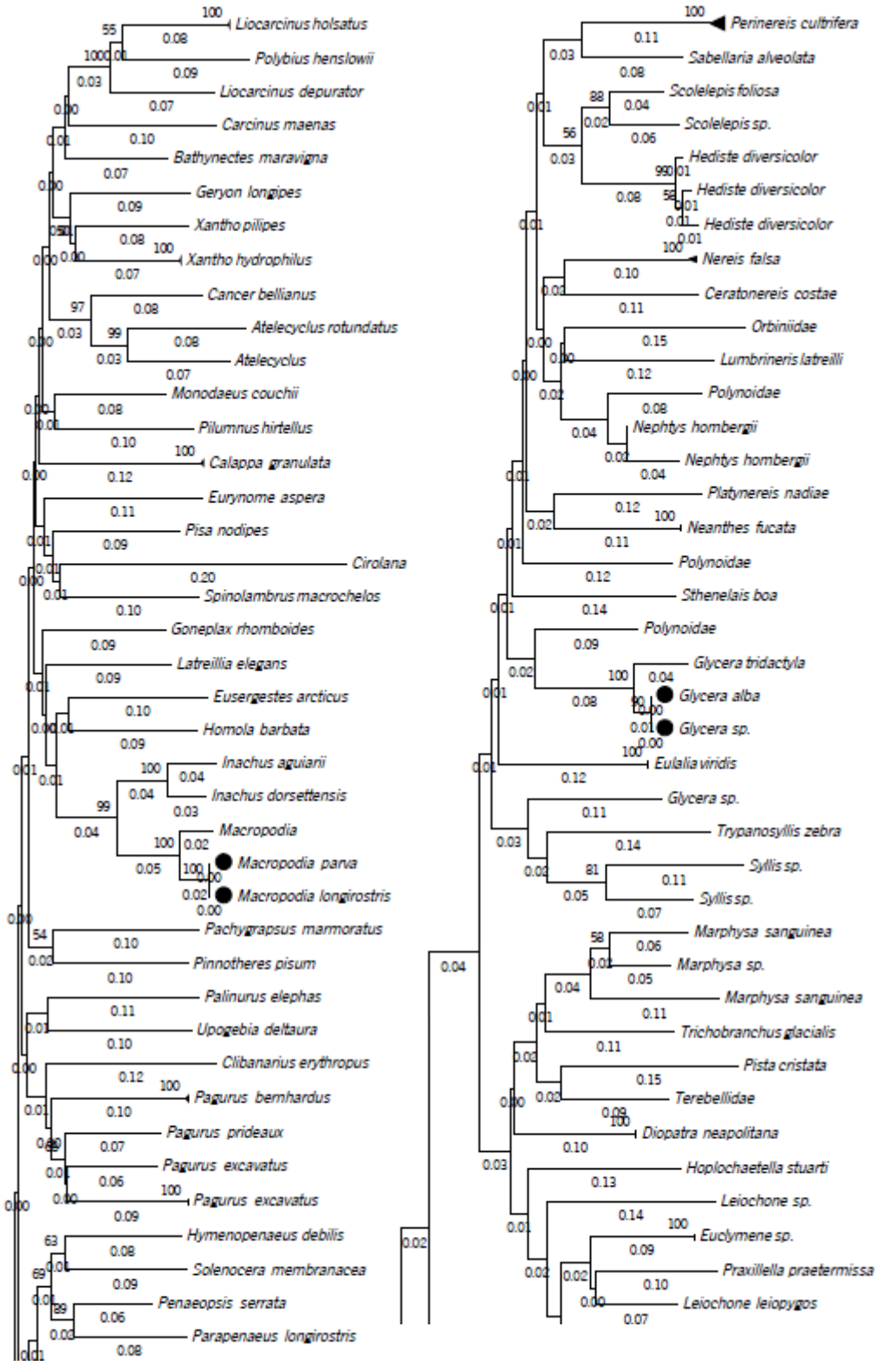

Figure 6 Phylogenetic NJ tree created from 315 sequences of A - full COI-5P DNA barcodes (658 bp) and B - short COI-5P fragments (158 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ● - Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree.

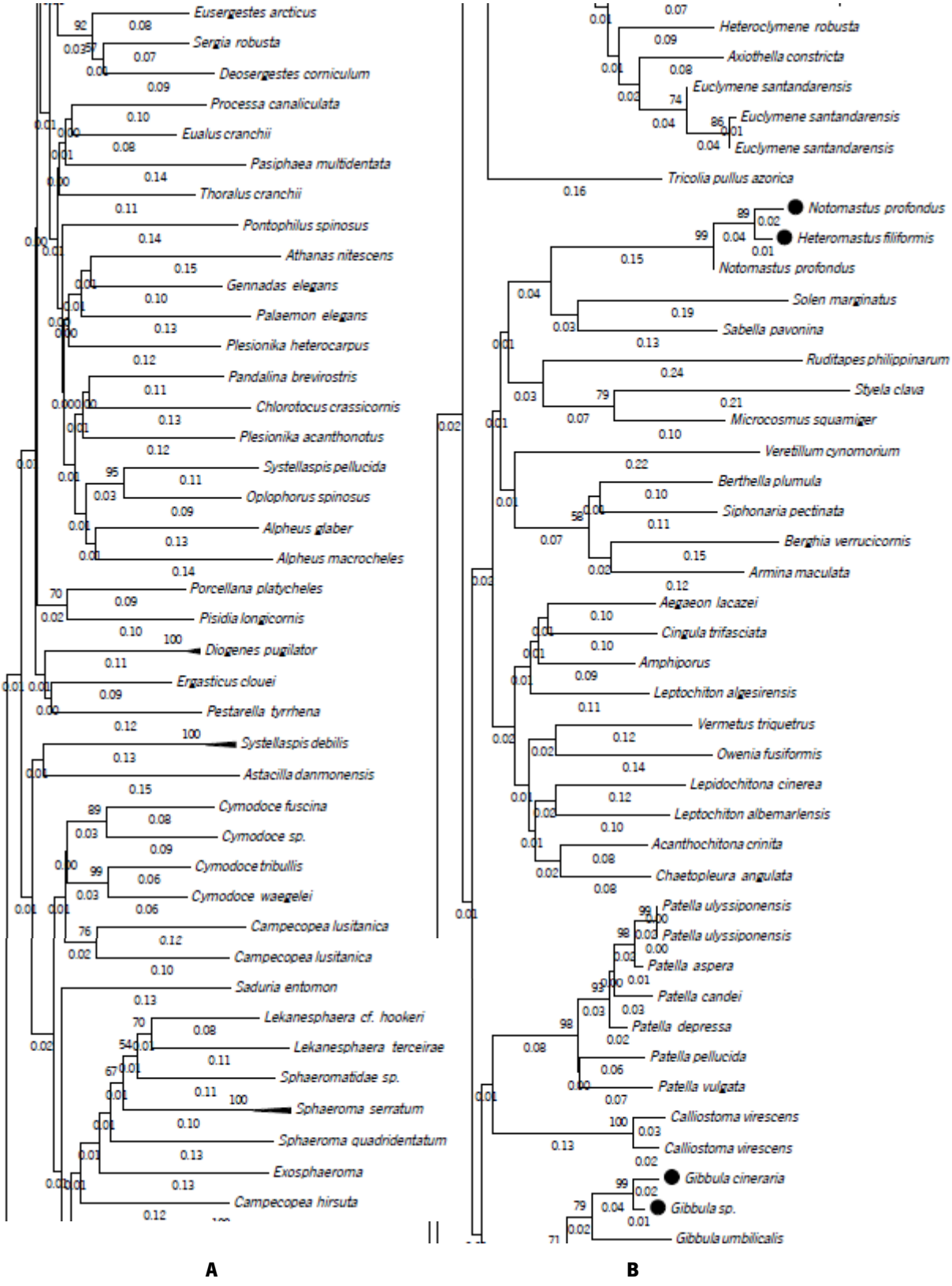

Figure 6 Phylogenetic NJ tree created from 315 sequences of A - full COI-5P DNA barcodes (658 bp) and B - short COI-5P fragments (158 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ● - Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

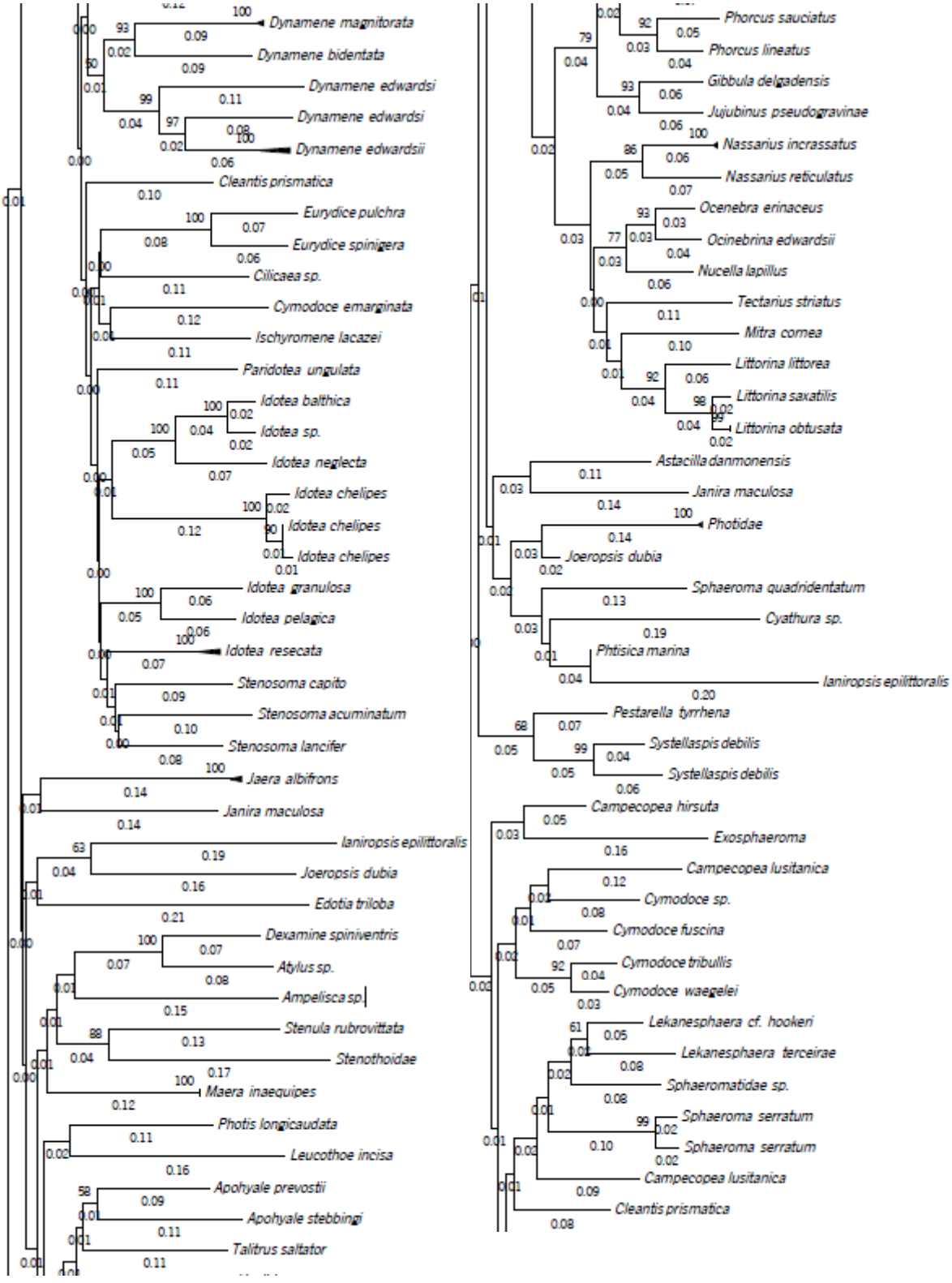

Figure 6 Phylogenetic NJ tree created from 315 sequences of A - full COI-5P DNA barcodes (658 bp) and B - short COI-5P fragments (158 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ● - Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

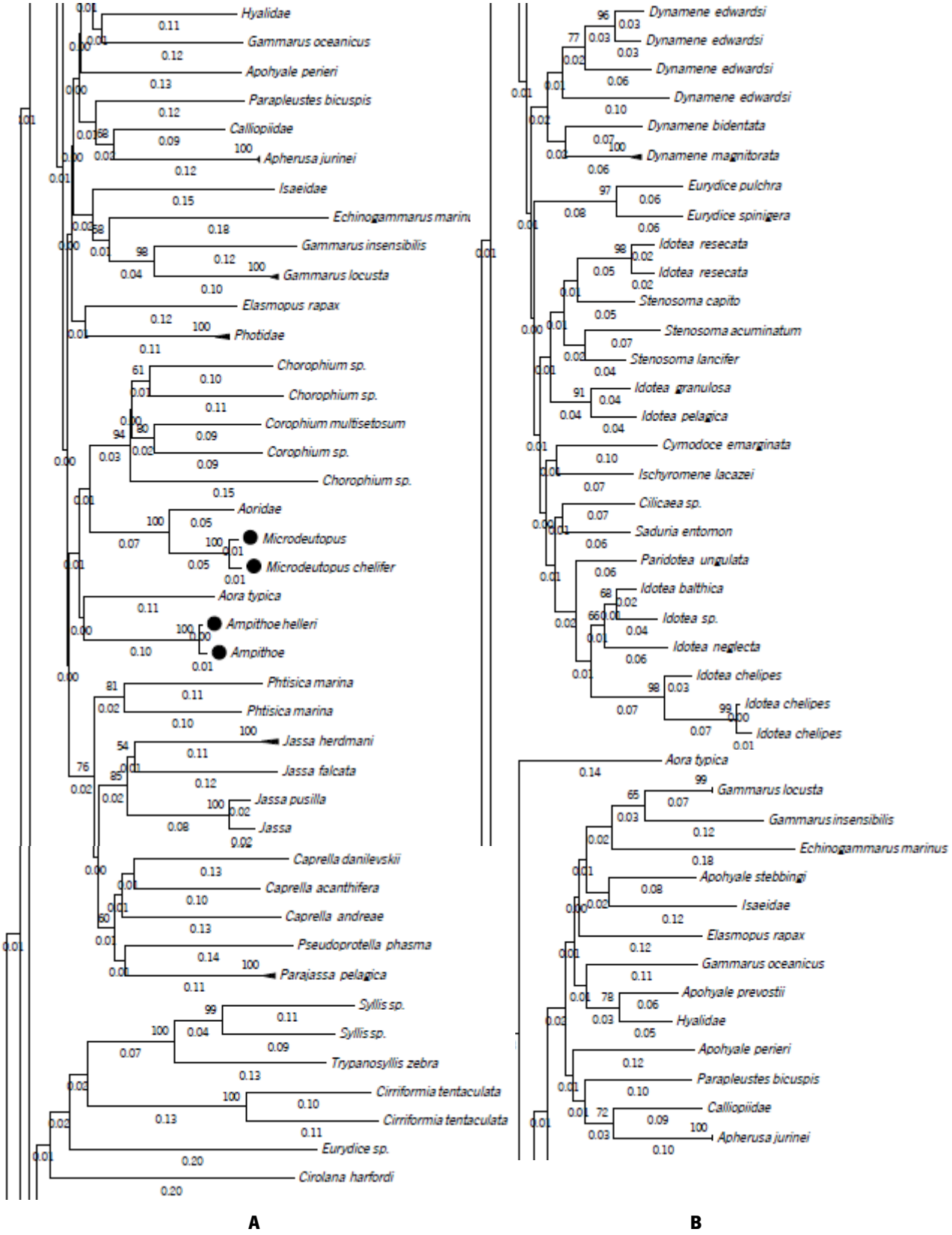

Figure 6 Phylogenetic NJ tree created from 315 sequences of A - full COI-5P DNA barcodes (658 bp) and B - short COI-5P fragments (158 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

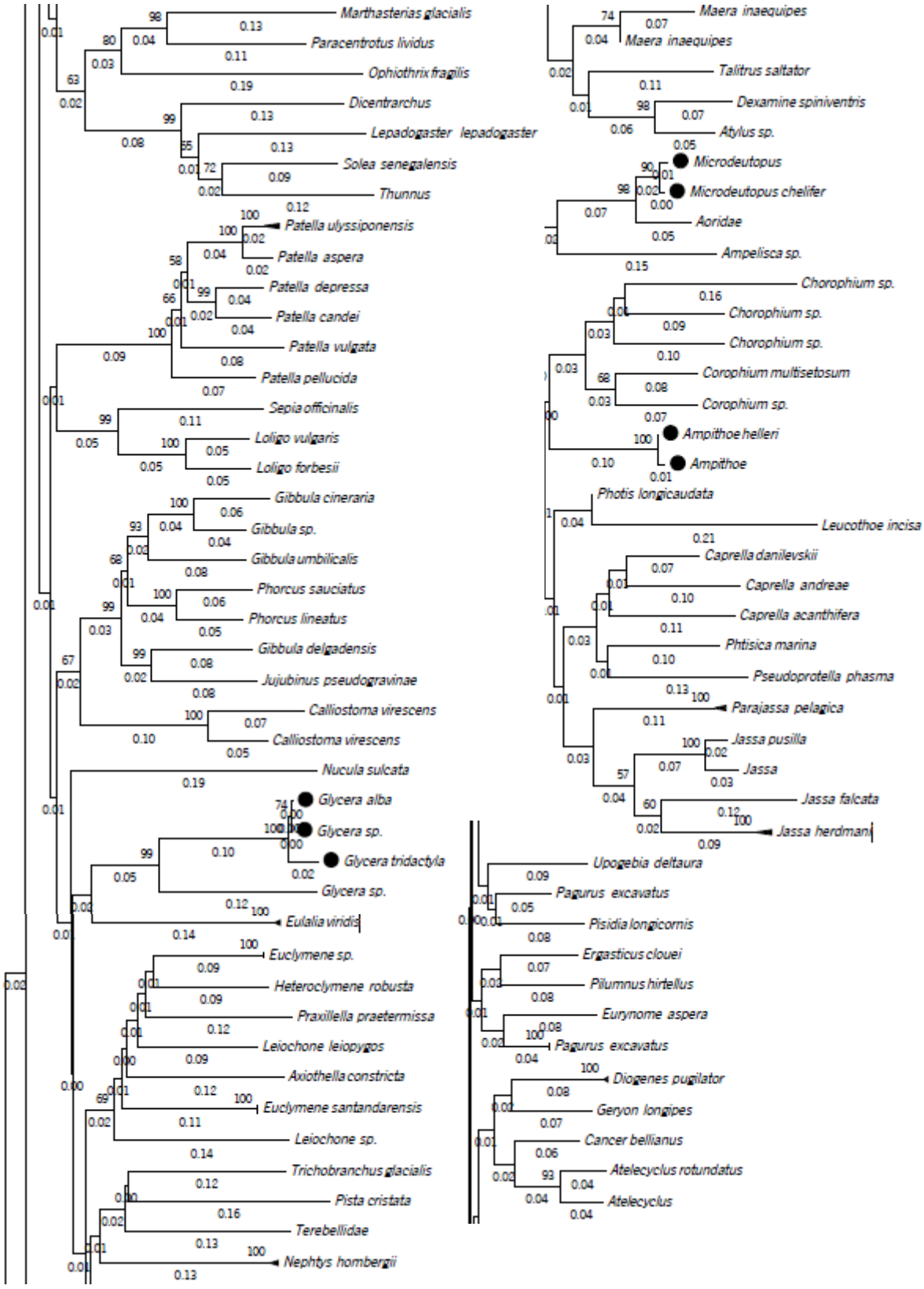

Figure 6 Phylogenetic NJ tree created from 315 sequences of A - full COI-5P DNA barcodes (658 bp) and B - short COI-5P fragments (158 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ● - Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

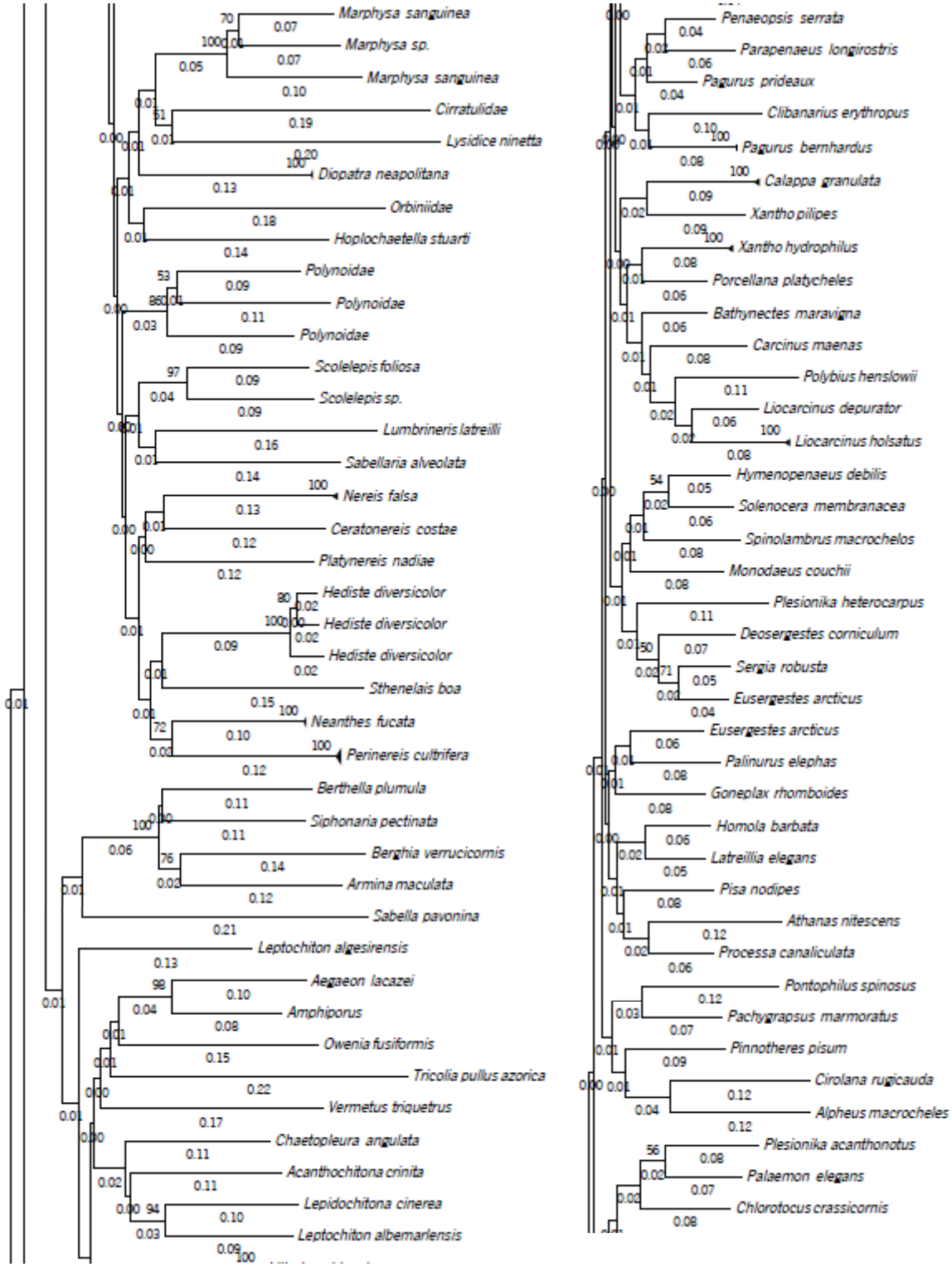

Figure 6 Phylogenetic NJ tree created from 315 sequences of A - full COI-5P DNA barcodes (658 bp) and B - short COI-5P fragments (158 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

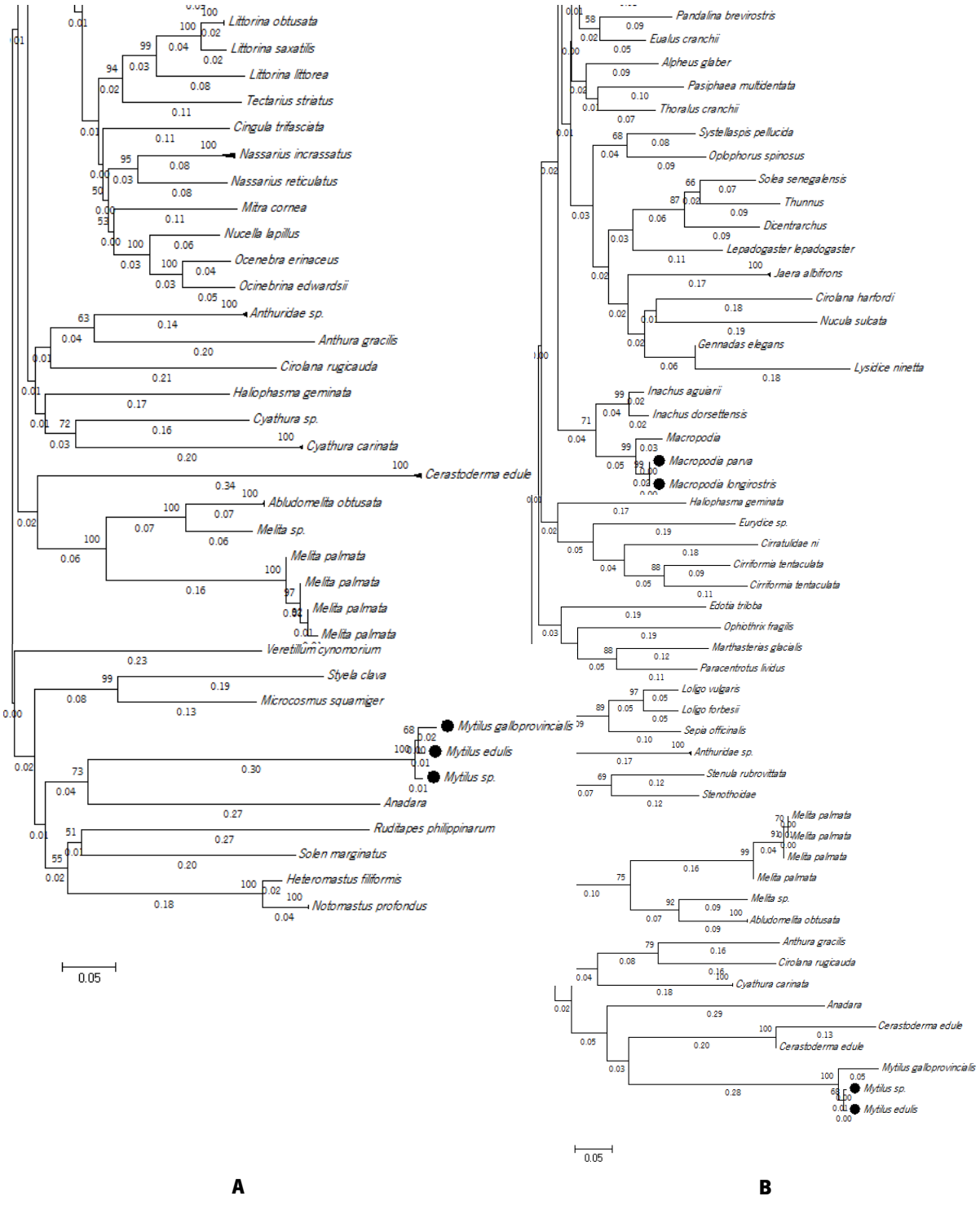

Figure 6 Phylogenetic NJ tree created from 315 sequences of A - full COI-5P DNA barcodes (658 bp) and B - short COI-5P fragments (158 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

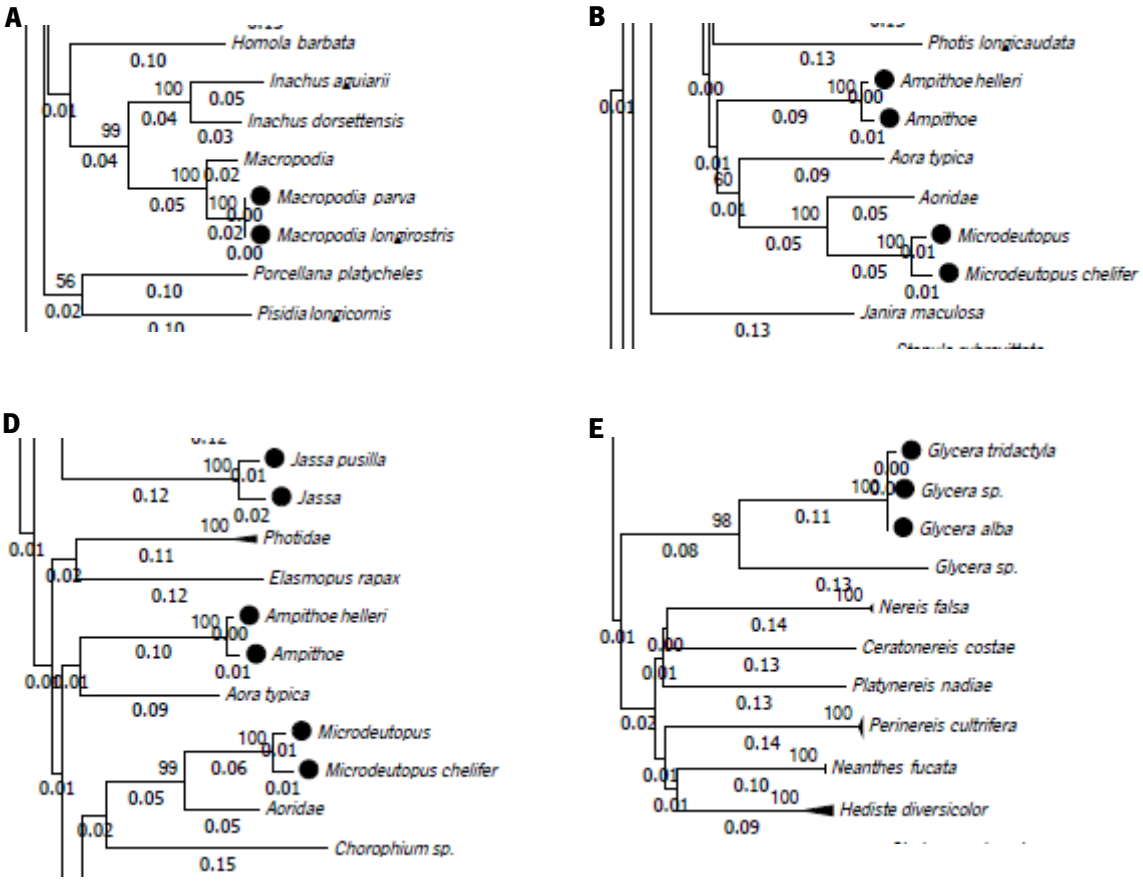

Figure 7 Phylogenetic NJ tree created from 315 sequences of COI DNA barcodes reference library clipped with the primer pairs. A – ArF2/LoboR (418 bp); B – invF/LoboR (470 bp); D – mlCOIintF/LoboR (313 bp); E – ArF2/ArR5 (310 bp). The Neighbor Joining (NJ) method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree.

3.2 Sequenced-based species identification through HTS

3.2.1 Global appraisal of HTS output

A total of 24198 454-pyrosequencing reads were generated: 12221 for SimCom1 and 11977 for SimCom2. Following trimming, filtering and quality checking 7709 (63%) sequences for SimCom1 and 7084 (59%) sequences for SimCom2 were used for our analysis. Of these sequences, 7499 (97%) for SimCom1 and 6282 (87%) for SimCom2 were assigned to a single species, if the 454 read shared >97% of sequence similarity to a Sanger generated COI-5P sequence of our reference library, or to barcode sequences archived in the public databases BOLD and GenBank. The number 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 number of sequences with match increase to 97% in SimCom1 and to 89% in SimCom2 (Figure 8). For more details about the total number of 454-pyrosequencing reads generated and the number of reads assigned to a single species see Table A1 (annex).

The increase of the number of sequences assigned to a species was observed using all five primer pairs, for the two simulated communities. The primer pair D generated more usable reads with sequence similarity higher than 97% in both simulated communities, while in the primers C (SimCom1) and B (SimCom2) less usable reads were obtained.

The primer pair E was the one that had a more significant variation in the number of usable reads, before and after sequence similarity search, with an increase of 1247 sequenced reads in SimCom1 and 784 in SimCom2.

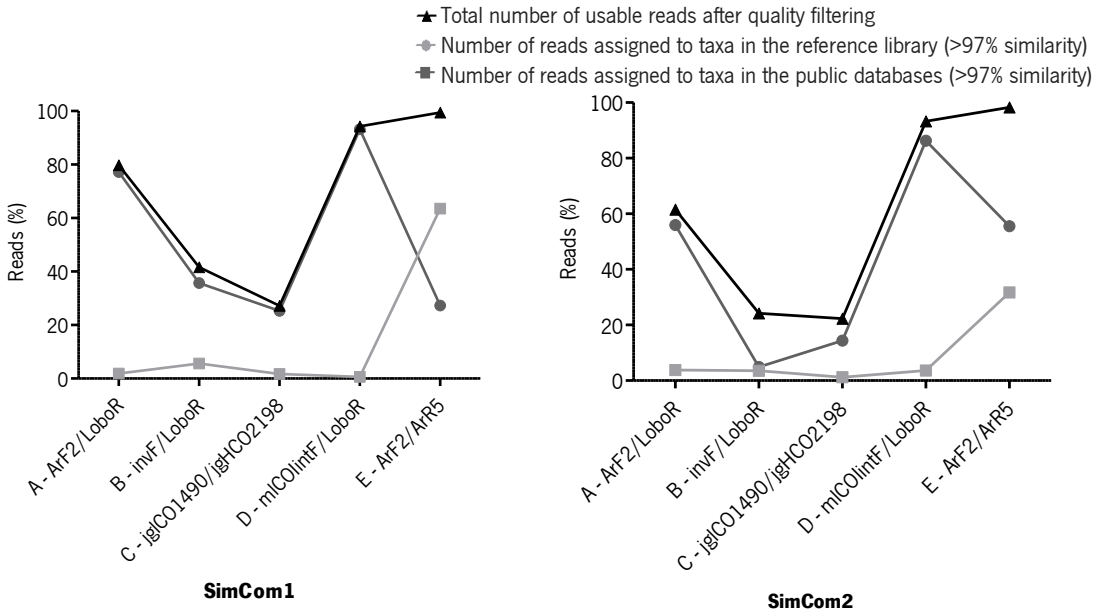

Figure 8 Sequence reads abundances generated by 454 pyrosequencing for two simulated macrobenthic communities.

The number of 454-pyrosequencing usable reads with sequence similarity higher than 97% per 21 selected species are presented in Figure 9. Globally, the number of reads varied between the two simulated macrobenthic communities.

Our results showed the predominance of some species with representative reads. The limpet *Patella aspera* was the most represented species with 6158 sequence reads in total. The species *Patella vulgata* and *Phorcus lineatus* were the next species in number of reads. Contrariwise, a

high number of species had fewer number of representative reads. Three species were identify only by a single read: *Lekanesphaera rugicauda* (SimCom1), *Hediste diversicolor* (SimCom2) and *Cyathura carinata* (SimCom2).

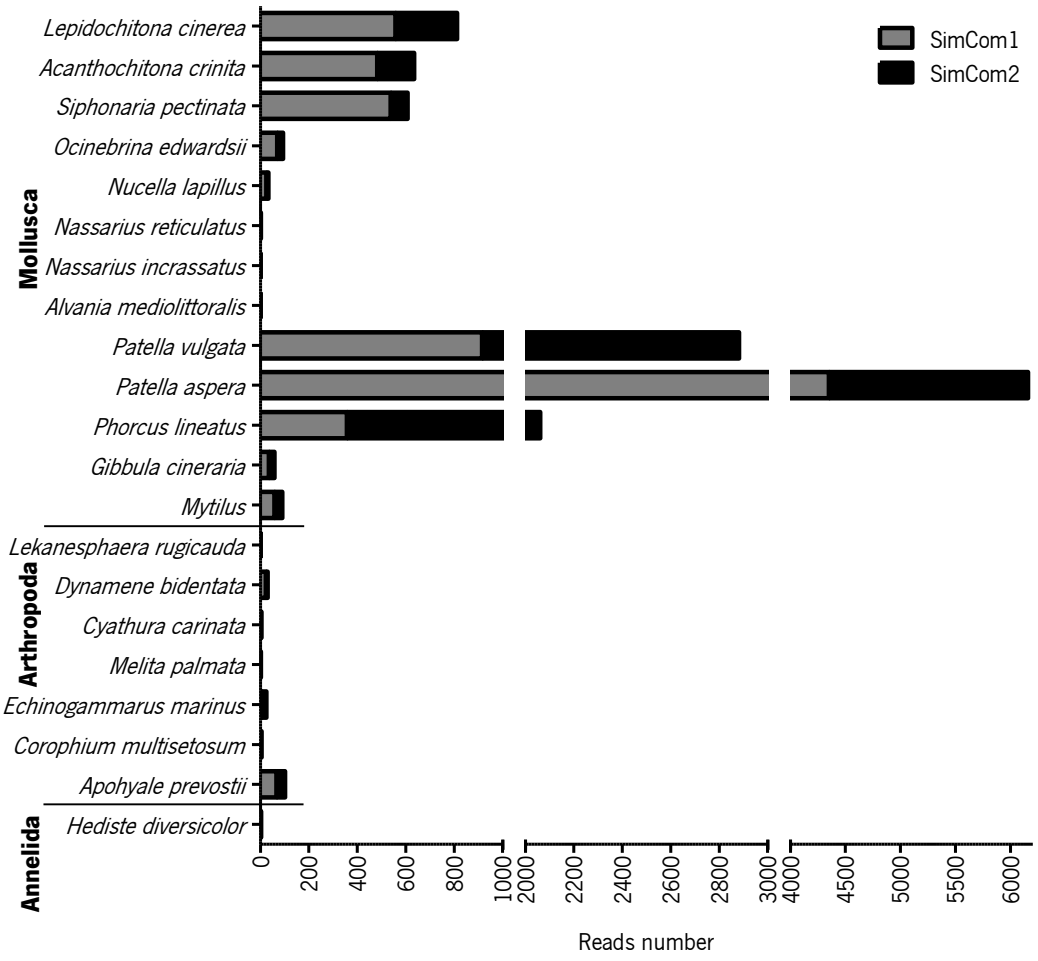

Figure 9 Number of sequenced reads generated by 454 pyrosequencing for each of the 21 species of three phyla and each of the two simulated macrobenthic communities.

3.3 Assessing the comparative success of primer pairs in taxa detection from simulated macrobenthic communities

3.3.1 Differential taxa detection among the primers and simulated macrobenthic communities

The results of taxa detection for each primer combination after pyrosequencing of simulated macrobenthic communities are displayed in Table 4. The effectiveness of a primer set in the detection of a taxon was confirmed if at least one representative read with similarity higher than

97%. By using a combination of five PCR-amplification primer pairs, we were able to recover 18 species from SimCom1 and 16 species from SimCom2. No single or combined primer set was able to recover 100% of species present in any of the simulated communities. However, two out of 21 species (*Alvania mediolittoralis* and *Nassarius incrassatus*) used in the pooled samples of each community did not also amplify by individual Sanger sequencing. Those specimens were stored at 4°C for a long period and we suspect that the DNA could be extensively degraded.

Interestingly, 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 biomass specimens did not seem to pose an impediment for that species detection. Species, such as *Echinogammarus marinus*, *Melita palmata, L. rugicauda* and *C. carinata*, which had just one representative in Sim Com1, 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*, *P. vulgata* and *Ocinebrina edwardsii* were recovered for all five primer pairs, while in SimCom2, *P. lineatus*, was the only species detected in all primer combinations.

Table 4 Species detection (1) or failed detection (0) for each primer pair after HTS of SimCom1 and SimCom2. Dark grey: species that was detected with the five primers in the two simulated communities; Light grey: 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.

Species \ Primers	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

3.3.2 Taxa recovery success rates among the simulated macrobenthic communities

The global taxa recovery success was slightly different between the two simulated macrobenthic communities (Figure 10). The combined five primer sets were able to recover 85.7% of the species in SimCom1 and 76.2% in SimCom2.

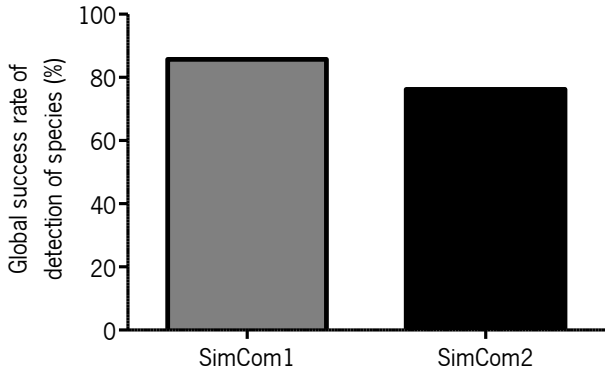

Figure 10 Global success rate of species detection of simulated macrobenthic communities.

In SimCom1, the most successful primer pairs were C (658 bp) and D (313 bp) with 61.9% of recovered species, while in SimCom2, the primer A (418 bp) detected 57.1% of species. The less successful result was obtained using primer E (310 bp) with 42.9% of species detected in SimCom1, against primers B (470 bp), which were able to recover only 26.6% of the species in SimCom2 (Figure 11).

In most cases, the success of species detection was different among the two simulated macrobenthic communities: a single primer had more success in SimCom1 when compared to SimCom2. Whereas the primer D had a detection success in SimCom1 of 61.9%, in SimCom2 detected less than five species, obtained less successful result of 38.1%. The only exceptions were the primers A. which increased detection success level from SimCom1 (52.4%) to SimCom2

(57.1%), and the primer E, which detected the same number of species in the two simulated macrobenthic communities.

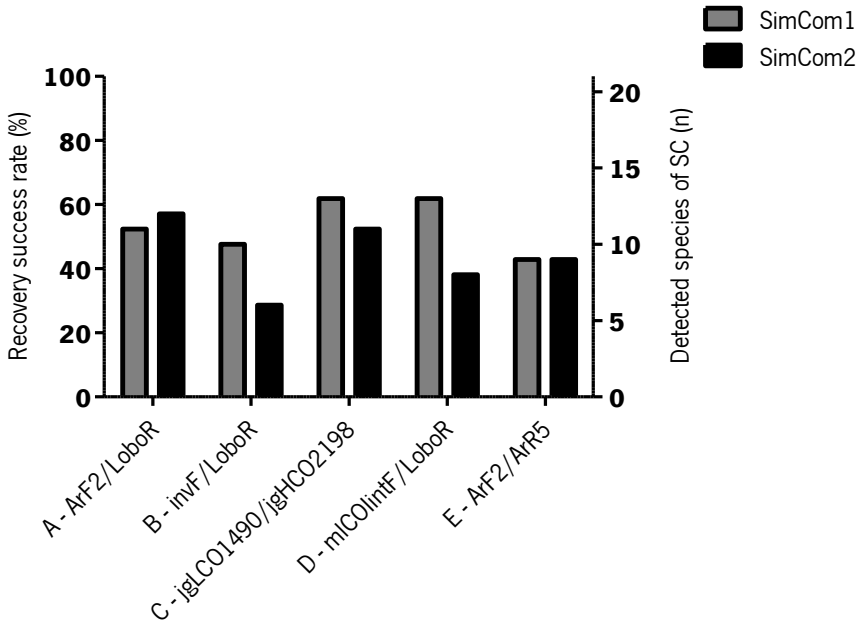

Figure 11 Recovery success rate and number of species detected for each of the five primer pairs in the two simulated macrobenthic communities.

The Figure 12 showing cumulative number of taxa recovered versus the primer set tested and indicates clear differences between the primer pairs. It is visible significant differences between the numbers of detected species, and consequently differences in primers affinity among simulated communities. Seen that curve increase proportionally among communities, through the addition of more primer pairs the success of species detection tends to increase.

Looking into our results, we observed that three primer sets in each simulated community were essential to acquire the highest number of species in each simulated community. Primers C and D for SimCom1 and primers A, C and E for SimCom2 were the most successful. Despite some complementarity between the primers, they were indispensable to have success in species detection. For example, the primer D in SimCom1 were the only that detected *L. rugicauda* and *M. plamata*. On the contrary, were some primers sets less relevant which showed total complementarity with other primers sets. Due this redundancy the primer A in SimCom1 and the primers B and D in SimCom2 are unnecessary to use. Furthermore, our results showed that although primer E detected fewer species in SimCom1, is relevant because they amplify species

that also only primer B: the polychaete *H. diversicolor*. These suggest the possibility to use others enhanced primers specifically designed.

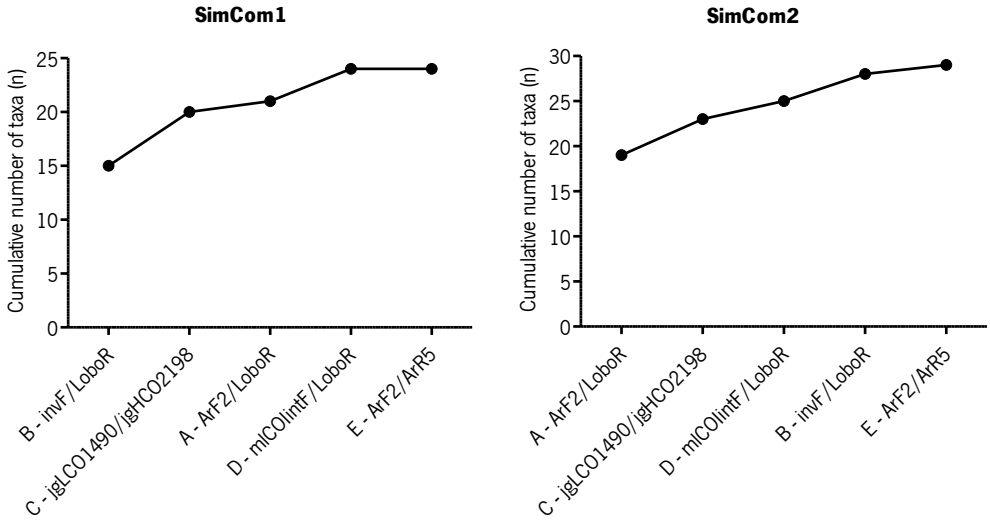

Figure 12 Accumulation curve of number of taxa successfully recovered by each primer set in the two simulated macrobenthic communities.

3.4 Detection of species not listed in the simulated communities

Surprisingly, the sequence similarity search in GenBank® and BOLD detected a total of 18 taxa identified at species or genus level (at >97% sequence similarity), which were not part of the listed species in any of the simulated macrobenthic communities (Table 5). 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 unlisted 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 for more reads (4 and 63, respectively). A total of 10 among the unlisted species detected had only one read.

Table 5 Detected taxa after the sequence similarity search in public databases (at 97%) that were not listed in the simulated communities. P: primer pair used: A – ArF2/LoboR; B – invF/LoboR; C – jgLCO1490/jgHCO2198; D – mlCOIintF/LoboR; E – ArF2/ArR5. R: number of sequence 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**
		Gigartinales	*Jania sp. 1MX*	-		**D**	**2**
			Peyssonnelia sp. 1WA	-		**B**	**4**

4. DISCUSSION

The aim of this study was to evaluate the efficiency of primer set in the assessment of the biodiversity of estuarine and marine macrobenthic communities. Although it has been used 454 massively parallelized pyrosequencing, this test is valid for others HTS technologies based on PCR amplification, since the technical limitation under investigation was the ability of primers amplification. Indeed, other platforms may even lead to a more profound sequencing ability. This information is especially important for biomonitoring programs, as the macrobenthic community structure is often used as indicator of aquatic ecosystems health, making these organisms good predictors of environmental changes. Biodiversity analysis can benefit from the use of DNA barcoding of individual specimens. However, while DNA barcoding uses traditional Sanger sequencing method to gather sequence information from single specimens, the application of HTS coupled with DNA barcoding can deliver information on assemblages of specimens at a much faster pace. In this context, a primary goal is to evaluate the capacity and advantages of HTS to retrieve complete and accurate DNA information from whole communities, when compared to the current taxonomic identification approaches, which are based on organisms' morphology.

Although a number of studies have used DNA metabarcoding to assess macrobenthic invertebrate biodiversity in freshwater ecosystems (e.g. Carew *et al.*, 2013, Hajibabaei *et al.*, 2011, Hajibabaei *et al.*, 2012), very few have applied the DNA barcode standard region (COI-5P) to examine marine or estuarine macrobenthic communities. To accomplish our objectives the main steps involved: (1) the compilation of a reference library of COI DNA barcodes of marine and estuarine coastal marine invertebrates, (2) evaluation of the effect of the amplicon size and location within the COI-5P barcode region on the species discrimination ability and (3) investigation of the performance of different primer sets for detection of species in the scope of a metabarcoding-based macrobenthos inventories. We assembled a standard DNA barcode reference library including 315 specimens, corresponding to 266 species of estuarine and coastal marine invertebrates from mainland Portugal and Azores Islands. This step is crucial to determine the taxonomic identity of individuals present in the simulated communities. In recent years, several studies have compiled DNA barcode reference libraries, providing a valuable resource for the identification of different taxa (e.g. Barco *et al.*, 2015; Borges *et al.*, submitted; Landi *et al.*, 2014; Lobo *et al.*, 2015).

One of the potentially limiting points under investigation in this study was the impact of the barcode length, as well the target region within the full barcode fragment, in the ability to discriminate species. The fragment size should be suitable for the desired HTS platforms and return an accurate species-level taxonomic identification. To this end, we constructed NJ phenograms, using COI-5P barcode sequences, available in our macrobenthic reference library, to inspect the species discrimination capacity of different fragment sizes. We found some incongruences in our dataset that were patent in the full COI-5P phenograms and which were originated from specimens that could not be morphologically identified to species, or that displayed intra-specific divergences higher than 3%. Hence these ambiguities were already present upfront in the full barcode reference library, and require further examination aside from this study. However, when compared to the full barcode region, all phenograms constructed for the different COI-5P fragments displayed similar clustering patterns with high bootstrap values and nearly any loss in the species discrimination ability compared to the full barcode. This results indicate the suitability of smaller fragments, from different regions within the COI-5P, for species-level resolution using our dataset (see Figure A 4, 5, 6, 7). 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 anticipated limitation to overcome 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. We tested different combinations of previously published primers targeting the barcode region (Geller *et al.*, 2013; Gibson *et al.*, 2014; Leray *et al.*, 2013; Lobo *et al.*, 2013). In addition, one newly designed forward primer (InvF) was included in the analyses. For this purpose, two experimentally assembled communities were used to evaluate the performance of five primer sets and their success in the species detection. 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. There were two cases of recalcitrant species (two gastropods: *A. mediolittoralis* and *N. incrassatus*) which were not detected with any primer set. However, even when testing specimens of these species individually, no PCR products were generated (data not shown). Aside from that, this study newly presents 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 samples which present at greater than 0.64% abundance of species presents in the contrived communities they could be detected. Other studies reported failures in sequence-based species identification 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 sequence generation somehow, as we found that SimCom1, which was composed by only one specimen per species (e.g. SimCom1: *H. diversicolor*, *A. prevosti*, *L. rugicauda*) had the best recovery results regarding small specimens, when compared to SimCom2 containing higher number of specimens. Deeper sequencing in a higher throughput platform (e.g. Illumina) may help to overcome potential bias that originated from the over 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, C. carinata* and *Lepidochitona cinerea,* failed to amplify by the same single primer in both communities. Since the goal is 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; Gibson *et al.*, 2014; Leray *et al.*, 2013). Looking into our results, we observed that only three primer sets were sufficient to recover the total number of species detected, although in different combinations for each simulated community. While in SimCom1 the primer combinations: A, C and D recovered more species, in SimCom2 the primer sets: A, C and E were the most successful combination. 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 (PCR amplification with combination of primers sets) 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. Pochon and collaborators (2013) used NGS sequencing for detecting the presence of various invasive species in marine ecosystems. They found that four distinct primer sets were required to obtain positive PCR amplifications for the COI gene across the five taxonomic groups under investigation. They observed that the addition of a third and fourth primer set substantially

improved their findings. Similarly, Gibson and collaborators (2014) used 11 primer sets to amplify three gene regions (COI, 16S and 18S) in order to investigate the diversity found in malaise trap samples taken from tropical Costa Rica. They found a much higher recovery rate across taxa when all 11 primer sets were used compared to any single primer set. However, it was observed that all eleven together provided little additional information over the two best sets.

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. However, we did failed to detect species that were present. In spite of the fact that some primers sets might not be adequate for a target species, we observed variations on a single primer set and its ability 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 results indicates that some additional work is needed to test detection limits variations in samples containing a diverse taxa at different abundances. Moreover, some adjustments in HTS sequencing protocols could be made in order to tune sequencing depth and coverage. In other words, by increasing or decreasing the number of sequence reads, researchers can tune the sensitivity of an experiment to accommodate their objectives.

Our results showed considerable variability in the number of sequence reads obtained between species and between the simulated communities. The genus *Patella* yielded the highest number of reads in both communities. However its biomass did not seem to be a contributing factor to these results, as the genus *Mytilus* represent a specimen with similar biomass in the mixture and displayed lower number of reads (see SimCom1). Also, some species like the crustacean *Apohyale prevostii,* which is a small species represented by lower biomass, obtained a higher number of reads comparatively with higher species like genus *Mytilus*. The number of individuals and the number of reads in the simulated communities could not be positively associated, as our results showed a higher number of reads for SimCom1, with lower specimen abundance. This results contrasts with the findings by Carew and collaborators (2013), they found a positive correlation in field-collected Chironomidae. Hajibabaeii *et al.*, (2011) suggested that species with higher affinity in 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 sequence reads and species detection.

In our study, a taxon was considered present in the 454 dataset when a sequence was > 97% similar to a Sanger method identified through our DNA barcode reference library, BOLD-IDS or BLASTn. In our reference library, we had full COI-5P sequences originated by Sanger method for all species represented in the communities, hence enabling an accurate taxonomic identification. A new cut-off threshold at 70% was adopted thereafter, aiming to find new information about species that could be possibly associated to others presented in our simulated communities. To this end, a new similarity search on BOLD and GenBank was conducted 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 new taxa identified at species or genus level. The polychaeta worm *Eulalia viridis* was recovered in SimCom1. Nereididae species such as *H. diversicolor* represented in our sample, are mainly omnivorous, but depending on nutrient availability may present a cannibalistic behavior (Caron *et al.*, 2004; Costa *et al.*, 2006; Fauchald *et al.*, 1979; Herrigshaw *et al.*, 2010; Scaps, 2002). A possible explanation is the identification of the DNA sequence of *E. viridis* through the gut content of *H. diversicolor*.

In SimCom2 one sequence read was identified as *Homo sapiens*. This result was possibly due to DNA contamination during laboratory experiments and careful laboratory procedures can minimize this result. The arthropod *Mytilicola intestinalis* was detected in SimCom2 by a significant number of sequence 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 in particular mussels (*M. galloprovincialis* and *M. edulis*, represented in our simulated macrobenthic communities) (Dethlefsen, 1985; Trotti *et al.*, 1998). This parasite causes overall reduction of condition, which affects the quality of meat in marketable mussels and it was associated with past mass mortalities of their hosts, leading to significant economic loss (Shinn *et al.*, 2015).

In addition, five species of Ochrophyta (brown alga) and six species of Rhodophyta (red alga) were detected, but most of them yield a small number of sequenced reads. The species: *Bangia atropurpurea*, *Corallina caespitosa*, *M. strangulans*, *Porphyra umbilicalis* and *Zonaria tournefortii* are found along Portuguese coast (Guiry and Guiry 2015, http://www.algaebase.org/; Pereira and Neto, 2015). There are no records for the remainder identified algae species in our coast. Many species in our simulated communities, including molluscan (e.g. *P. vulgata*, *P. aspera*, *Mytylus* sp) and crustacean species (e.g. *E. marinus*, *C. carinata*) that may feed on algae (Martins *et al.*, 2010). Moreover, species of algae are known to be able to live in epibiosis (i.e. any relationship between two organisms in which one grows on the other but is not parasitic on it) with groups of organisms such as crustaceans and molluscs. An association among nine epibiontes were reported by (Martins *et al.*, 2014), including the taxa *Jania sp.* and *P. aspera*, both identified in our study. Similar studies report the presence of the algae of the genus *Corallina* in *P. aspera* on the Portuguese coast (Guerra and Gaudêncio, 1986).

For the bulk DNA extraction we homogenized whole specimens, without any manipulation or removal of body parts. The goal was to test the metabarcoding procedure in a realistic way, which would mimic the intended procedure for analyzing benthic communities without need for specimen sorting or other type of time consuming manipulation. Therefore, either barnacles, algae or other epibiontes, could have been growing in the shells of the mussels included in the simulated communities. 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. Epibiosis could also be a possible explanation for the detection of *Patella depressa*, a limpet species which was not included in the simulated communities, but that was identified by the primers A and E, in SimCom2 (three sequence reads in both). This species is easy to distinguish morphologically from the *Patella* species present in our sample, and their barcode sequences group into distinct and well defined clusters. Moreover, most of the primer sets showed a high level of specificity for the amplification of the *Patella* species (our own observations), suggesting that *P. depressa* could not possibly be misidentified as *P. aspera* and *P. vulgata*. Another possibility to explain this result, is that the DNA of *P. depressa* 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 is 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).

5. CONCLUSION

Our exploratory investigation have demonstrated that the application of combined primer sets coupled with high-throughput technologies may enhance species identification throughput in marine macrobenthic communities inventories, even for low-abundance species, overtaking the need for deep technical competency in taxa identification. The sensitivity of this approach may easily outperform traditional morphological identification methods, as it revealed to be practical and objective, with a great potential to detect a "hidden" biodiversity that could not be possibly identified based on morphology.

The reference library compiled with COI-5P DNA barcodes of estuarine and coastal marine invertebrates from Portugal demonstrated to be a vital framework for efficient sequenced-based species identification. The COI-5P reference library must continue to grow through the addition of new DNA barcodes for the numerous macrobenthic species that are still missing. This includes the compilation of compliant sequences from multiple studies and sources available in public databases, and must be complemented with a thorough inspection and annotation of every species records to reduce ambiguities.

The five primers pairs used in this study proved to be able to amplify, and therefore to detect, a high very proportion of the species present in the two simulated macrobenthic communities, including those that would have been missed by conventional procedures, such as internal parasites. Following further refinement, this methodology has great potential for application in future biomonitoring studies, such as large-scale marine and estuarine macrobenthic biodiversity assessments. The results here collected are readily extensible to other HTS platforms than Roche 454, since the key technical limitation under investigation was the primer amplification ability. In fact, other platforms may even provide deeper sequencing capacity while still allowing full barcode sequencing, if required. A logical follow-up to this study would be the comparison of the species compositions of unknown bulk environmental samples collected from different sites. To this end, by comparing morphology-based identifications with the HTS-metabarcoding approach, the success rate of species detection over known communities and unknown communities can be investigated.

This study opens prospects to much cheaper, objective and practical methods for the detection and inventorying of species diversity present in macrobenthic assemblages. We trust that by using combined primer sets coupled with high-throughput technologies the capacity for ecological and evolutionary studies can be greatly increased. Also, the implementation of biomonitoring programs, extended to habitats and biota groups, can be done due to technical competency.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. Journal of Molecular Biology 215, 403-410.

Antunes I, Ferreira MSG, Lobo J, Teixeira MAL, Borges LMS, Sousa R, Gomes PA, Costa MH, Cunha MR, Costa FO (2015). Comparison between morphological and DNA barcode-suggested species boundaries among shallow water amphipod fauna from the southern European Atlantic coast. Scientific abstracts from the 6th International Barcode of Life Conference. Genome 58,187.

Arif IA, Khan HA, Al Sadoon M, Shobrak M (2011). Limited efficiency of universal mini-barcode primers for DNA amplification from desert reptiles, birds and mammals. Genetics and Molecular Research 10, 3559-3564.

Aylagas E, Borja Á, Rodríguez-Ezpeleta N (2014). Environmental status assessment using DNA metabarcoding: towards a genetics based Marine Biotic Index (gAMBI). PLoS ONE 9, e90529.

Baird DJ and Sweeney BW (2011). Applying DNA barcoding in benthology: the state of the science. Journal of the North American Benthological Society 30, 122-124.

Barco A, Evans J, Schembri PJ, Taviani M, Oliverio M (2013). Testing the applicability of DNA barcoding for Mediterranean species of top-shells (Gastropoda, Trochidae, *Gibbula* sl). Marine Biology Research 9, 785-793.

Barco A, Raupach MJ, Laakmann S, Neumann H, Knebelsberger T (2015). Identification of North Sea molluscs with DNA barcoding. Molecular Ecology Resources. doi: 10.1111/1755-0998.12440.

Barroso R, Klautau M, Solé-Cava AM, Paiva PC (2010). *Eurythoe complanata* (Polychaeta: Amphinomidae), the 'cosmopolitan' fireworm, consists of at least three cryptic species. Marine Biology 157, 69-80.

Begerow D, Nilsson H, Unterseher M, Maier W (2010). Current state and perspectives of fungal DNA barcoding and rapid identification procedures. Applied Microbiology and Biotechnology 87, 99-108.

Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I (2007). Cryptic species as a window on diversity and conservation. Trends in Ecology and Evolution 22, 148-155.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK (2012). Sequencing our way towards understanding global eukaryotic biodiversity. Trends in Ecology and Evolution 27, 233-243.

Blanco-Bercial L, Cornils A, Copley N, Bucklin A (2014). DNA barcoding of marine copepods: assessment of analytical approaches to species identification. PLoS Currents Tree of Life. doi: 10.1371/currents.tol.cdf8b74881f87e3b01d56b43791626d2.

Blankenship LE and Yayanos AA (2005). Universal primers and PCR of gut contents to study marine invertebrate diets. Molecular Ecology 14, 891-899.

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005). Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society B: Biological Sciences 360, 1935-1943.

Borges LMS, Hollatz C, Lobo J, Cunha AM, Vilela AP, Calado G, Coelho R, Costa AC, Ferreira MSG, Costa MH, Costa FO (submetido). With a little help from DNA barcoding: investigating the diversity of Gastropoda from the Portuguese coast. Scientific Reports.

Bouchet P (2006). (eds.) The magnitude of marine biodiversity in Exploration of Marine Biodiversity. Scientific and Technological challenges, pp. 31–62. Fundación BBVA.

Bucklin A, Hopcroft RR, Kosobokova KN, Nigro LM, Ortman BD, Jennings RM, Sweetman CJ (2010). DNA barcoding of Arctic Ocean holozooplankton for species identification and recognition. Deep Sea Research Part II: Topical Studies in Oceanography 57, 40-48.

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

Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA (2013). Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. Frontiers in Zoology 10, 45.

Caron A, Desrosiers G, Olive PJW, Retière C, Nozais C (2004). Comparison of diet and feeding activity of two polychaetes, *Nephtys caeca* (Fabricius) and *Nereis virens* (Sars), in an estuarine intertidal environment in Québec, Canada. Journal of Experimental Marine Biology and Ecology 304, 225-242.

CBOL Plant Working Group: Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim KJ, Kress WJ, Schneider H, van AlphenStahl J, Barrett SCH, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML, Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim YD, Lahaye R, Lee HL, Long DG, Madriñán S, Maurin O, Meusnier I, Newmaster SG, Park CW, Percy DM, Petersen G, Richardson JE, Salazar GA, Savolainen V, Seberg O, Wilkinson MJ, Yi DK, Little DP (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences 106, 12794-1277.

Chapman AD (2009). Numbers of living species in Australia and the world. Australian Biological Resources Study.

Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V (2005). Land plants and DNA barcodes: short-term and long-term goals. Philosophical Transactions of the Royal Society B: Biological Sciences 360, 1889-1895.

Chen H and Jiang W (2014). Application of high-throughput sequencing in understanding human oral microbiome related with health and disease. Frontiers in Microbiology 5, 508.

Cho Y, Mower JP, Qiu YL, Palmer JD (2004). Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proceedings of the National Academy of Sciences of the United States of America 101, 17741-17746.

Corell J and Rodríguez-Ezpeleta N (2014). Tuning of protocols and marker selection to evaluate the diversity of zooplankton using metabarcoding. Revista de Investigación Marine, AZTI-Tecnalia 21, 19-39.

Costa FO and Antunes PM (2012). The contribution of the Barcode of Life initiative to the discovery and monitoring of Biodiversity. In: Mendonça A, Chakrabarti R and Cunha A (eds.) Natural Resources, Sustainability and Humanity – A comprehensive View, pp. 37-68. Springer, Dordrecht.

Costa FO and Carvalho GR (2007). The Barcode of Life Initiative: synopsis and prospective societal impacts of DNA barcoding of fish. Life Sciences Society and Policy 3, 29.

Costa FO and Carvalho GR (2010). New insights into molecular evolution: prospects from the Barcode of Life Initiative (BOLI). Theory in Biosciences 129, 149-157.

Costa FO, deWaard JR, Boutillier J, Ratnasingham S, Dooh RT, Hajibabaei M, Hebert PDN (2007). Biological identifications through DNA barcodes: the case of the Crustacea. Canadian Journal of Fisheries and Aquatic Sciences 64, 272-295.

Costa FO, Landi M, Martins R, Costa MH, Costa ME, Carneiro M, Alves MJ, Steinke D, Carvalho GR (2012). A ranking system for reference libraries of DNA barcodes: application to marine fish species from Portugal. PLoS ONE 7, 1-9.

Costa PF, Oliveira RF, Fonseca LC (2006). Feeding ecology of *Nereis diversicolor* (OF Müller) (Annelida, Polychaeta) on estuarine and lagoon environments in the southwest coast of Portugal. Pan-American Journal of Aquatic Sciences 1, 114-126.

Cowart DA, Pinheiro M, Mouchel O, Maguer M, Grall J, Miné J, Arnaud-Haond S (2015). Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of Morphological and Molecular Surveys of Seagrass Communities. PLoS ONE 10, e0117562.

Cox AJ and Hebert PDN (2001). Colonization, extinction, and phylogeographic patterning in a freshwater crustacean. Molecular Ecology 10, 371-386.

Creer S, Fonseca VG, Porazinska DL, Giblin-Davis RM, Sung W, Power DM, Packer M, Carvalho GR, Blaxter ML, Lambshead PJD, Thomas WK (2010). Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Molecular Ecology 19, 4-20.

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biology Letters 10, 20140562.

Deagle BE, Kirkwood R, Jarman SN (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. Molecular Ecology 18, 2022-2038.

Dethlefsen V (1985). *Mytilicola intestinalis*, parasitism. International Council for the Exploration of the Sea.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.

Elsner NO, Jacobsen S, Thieltges DW, Reise K (2011). Alien parasitic copepods in mussels and oysters of the Wadden Sea. Helgoland Marine Research 65, 299-307.

Fauchald K and Jumars PA (1979). The diet of worms: a study of polychaete feeding guilds. Oceanography and Marine Biology Annual Review 17, 193-284.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3, 294-299.

Fonseca VG, Carvalho GR, Nichols B, Quince C, Johnson HF, Neill SP, Lambshead JD, Thomas WK, Power DM, Creer S (2014). Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes. Global Ecology and Biogeography 23, 1293-1302.

Fonseca VG, Carvalho GR, Sung W, Johnson HF, Power DM, Neill SP, Packer M, Blaxter ML, Lambshead PJD, Thomas WK, Creer S (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nature communications 1, 98.

Fonseca VG, Nichols B, Lallias D, Quince C, Carvalho GR, Power DM, Creer S (2012). Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. Nucleic Acids Research 40, e66.

Geller J, Meyer C, Parker M, Hawk H (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources 13, 851-861.

Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical

arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences 111, 8007-8012.

Gomes NMA (2014). Construção de uma biblioteca de referência de DNA barcodes para Isópodes marinhos (Crustacea: Isopoda) de Portugal e da Macronésia. Dissertação de Mestrado em Ecologia, Universidade do Minho.

Guerra MT and Gaudencio MJ (1986). Aspects of the ecology of *Patella* spp. on the Portuguese coast. Hydrobiologia 142, 57-69.

Guiry MD and Guiry GM (2015). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. Available from http://www.algaebase.org. Accessed at 2015-10-17.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS ONE 6, e17497.

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006). A minimalist barcode can identify a specimen whose DNA is degraded. Molecular Ecology Notes 6, 959-964.

Hajibabaei M, Spall JL, Shokralla S, van Konynenburg S (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BioMed Central Ecology 12, 28.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003). Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B: Biological Sciences 270, 313-321.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proceedings of the National Academy of Sciences of the United States of America 101, 14812-14817.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004). Identification of birds through DNA barcodes. PLoS Biology 2, 1657-1663.

Herringshaw LG, Sherwood OA, McILroy D (2010). Ecosystem engineering by bioturbating polychaetes in event bed microcosms. Palaios 25, 46-58.

Ishii K and Fukui M (2001). Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. Applied and Environmental Microbiology 67, 3753-3755.

Kekkonen M and Hebert PDN (2014). DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Molecular Ecology Resources 14, 706-715.

Kimura M (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16, 111-120.

Knox MA, Hogg ID, Pilditch CA, Lörz AN, Hebert PDN, Steinke D (2012). Mitochondrial DNA (COI) analyses reveal that amphipod diversity is associated with environmental heterogeneity in deep-sea habitats. Molecular Ecology 21, 4885-4897.

Kurata S, Kanagawa T, Magariyama Y, Takatsu K, Yamada K, Yokomaku T, Kamagata Y (2004). Reevaluation and reduction of a PCR bias caused by reannealing of templates. Applied and Environmental Microbiology 70, 7545-7549.

La Rosa M, Fiannaca A, Rizzo R, Urso A (2013). Alignment-free analysis of barcode sequences by means of compression-based methods. BioMed Central Bioinformatics 14, S4.

Landi M, Dimech M, Arculeo M, Biondo G, Martins R, Carneiro M, Carvalho GR, Lo Brutto S, Costa FO (2014). DNA barcodes for the identification of marine fish species from southern Europe. Can a reference library of DNA barcodes of fish from Portugal be used to identify marine fish from the central Mediterranean? PLoS One 9, e106135.

Le Gall L and Saunders GW (2010). DNA barcoding is a powerful tool to uncover algal diversity: a case study of the phyllophoraceae (Gigartinales, Rhodophyta) in the Canadian flora. Journal of Phycology 46, 374-389.

Lejzerowicz F, Esling P, Pillet L, Wilding TA, Black KD, Pawlowski J (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. Scientific Reports 5, 13932.

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (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. Frontiers in Zoology 10, 34.

Lindeque PK, Parry HE, Harmer RA, Somerfield PJ, Atkinson A (2013). Next generation sequencing reveals the hidden diversity of zooplankton assemblages. PLoS ONE 8, e81327.

Lobo J, Costa PM, Teixeira MAL, Ferreira MSG, Costa MH, Costa FO (2013). Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. BioMed Central Ecology 13, 34.

Lobo J, Teixeira MAL, Borges LMS, Ferreira MSG, Hollatz C, Gomes PA, Sousa R, Ravara A, Costa MH, Costa FO (2015). Starting a DNA barcode reference library for shallow water polychaetes from the southern European Atlantic coast. Molecular Ecology Resources doi: 10.1111/1755-0998.12441.

Lohman DJ, Prawiradilaga DM, Meier R (2009). Improved COI barcoding primers for Southeast Asian perching birds (Aves: Passeriformes). Molecular Ecology Resources 9, 37-40.

Lynn DH and Strüder-Kypke MC (2006). Species of *Tetrahymena* identical by small subunit rRNA gene sequences are discriminated by mitochondrial cytochrome c oxidase I gene sequences. Journal of Eukaryotic Microbiology 53, 385-387.

Mardis ER (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9, 387-402.

Martins GM, Faria J, Furtado M, Neto AI (2014). Shells of *Patella aspera* as 'islands' for epibionts. Journal of the Marine Biological Association of the United Kingdom 94, 1027-1032.

Martins GM, Thompson RC, Neto AI, Hawkins SJ, Jenkins SR (2010). Exploitation of intertidal grazers as a driver of community divergence. Journal of Applied Ecology 47, 1282-1289.

Meusnier I, Singer GAC, Landry JF, Hickey DA, Hebert PDN, Hajibabaei M (2008). A universal DNA mini-barcode for biodiversity analysis. BioMed Central Genomics 9, 214.

Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson KH (2008). Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. Evolutionary Bioinformatics 4, 193.

Padial JM, Miralles A, De la Riva I, Vences M (2010). The integrative future of taxonomy. Frontiers in Zoology 7, 1-14.

Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010). Names are key to the big new biology. Trends in Ecology and Evolution 25, 686-691.

Pereira L and Neto JM (eds.) (2014). Marine algae: biodiversity, taxonomy, environmental assessment, and biotechnology. CRC Press.

Pochon X, Bott NJ, Smith KF, Wood SA (2013) Evaluating Detection Limits of Next-Generation Sequencing for the Surveillance and Monitoring of International Marine Pests. PLoS ONE 8, e73935.

Porazinska DL, Giblin-Davis RM, Faller L, Farmerie W, Kanzaki N, Morris K, Powers TP, Tucker AE, Sung W, Thomas W K (2009). Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. Molecular Ecology Resources 9, 1439-1450.

Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, Zhou J (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Applied and Environmental Microbiology 67, 880-887.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011). Removing noise from pyrosequenced amplicons. BioMed Central Bioinformatics 12, 38.

Ratnasingham S and Hebert PDN (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). Molecular Ecology Notes 7, 355-364.

Ratnasingham S and Hebert PDN (2013). A DNA-based registry for all animal species: The Barcode Index Number (BIN) System. PLoS ONE 8, e66213.

Rodman JE and Cody JH (2003). The taxonomic impediment overcome: NSF's partnerships for enhancing expertise in taxonomy (PEET) as a model. Systematic Biology 52, 428-435.

Rothberg JM and Leamon JH (2008). The development and impact of 454 sequencing. Nature Biotechnology 26, 1117-1124.

Saitou N and Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406-425.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74, 5463-5467.

Scaps P (2002). A review of the biology, ecology and potential use of the common ragworm *Hediste diversicolor* (OF Müller) (Annelida: Polychaeta). Hydrobiologia 470, 203-218.

Seifert KA (2009). Progress towards DNA barcoding of fungi. Molecular Ecology Resources 9, 83-89.

Shendure J and Ji H (2008). Next-generation DNA sequencing. Nature Biotechnology 26, 1135-1145.

Shinn AP, Pratoomyot J, Bron JE, Paladini G, Brooker EE, Brooker AJ (2015). Economic costs of protistan and metazoan parasites to global mariculture. Parasitology 142, 196-270.

Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M (2014). Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Molecular Ecology Resources 14, 892-901.

Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB, Hajibabaei M (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports 5, 9687.

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012). Next-generation sequencing technologies for environmental DNA research. Molecular ecology 21, 1794-1805.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences 103, 12115-12120.

Song H, Buhay JE, Whiting MF, Crandall KA (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences 105, 13486-13491.

Sønstebø JH, Gielly L, Brysting AK (2010). Using next generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Molecular Ecology Resources 10, 1009–1018.

Stoeckle M (2003). Taxonomy, DNA, and the bar code of life. BioScience 53, 796-797.

Taanman JW (1999). The mitochondrial genome: structure, transcription, translation and replication. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1410, 103-123.

Taberlet P, Coissac E, Hajibabei M, Rieseberg LH (2012). Environmental DNA. Molecular Ecology 21, 1789-1793.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21, 2045-2050.

Taberlet P, Prud'homme SM, Campione E, Roy J, Miquel C, Shehzad W, Gielly L, Rioux D, Choler P, Clément J, Melodelima C, Pompanon F, Coissac E (2012). Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Molecular Ecology 21, 1816-1820.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution 30, 2725-2729.

Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D (2012). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences 109, 16208-16212.

Teletchea F (2010). After 7 years and 1000 citations: comparative assessment of the DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. Mitochondrial DNA 21, 206-226.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673-4680.

Trotti GC, Baccarani EM, Giannetto S, Giuffrida A, Paesanti F (1998). Prevalence of *Mytilicola intestinalis* (Copepoda: Mytilicolidae) and *Urastoma cyprinae* (Turbellaria: Hypotrichinidae) in marketable mussels *Mytilus galloprovincialis* in Italy. Diseases of Aquatic Organisms 32, 145-149.

Wheeler QD (Ed.) (2008). The new taxonomy. CRC Press.

Worms Editorial Board (2015). World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed at 2015-10-17.

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3, 613-623.

ANNEX

Table A 1 Number of sequence reads generated by 454 pyrosequencing. A – ArF2/LoboR; B – invF/LoboR; C – jgLCO1490/jgHCO2198; D – mlCOIintF/LoboR; E – ArF2/ArR5.

	SimCom1					SimCom2				
	A	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
Total reads number	2648	1816	3840	1955	1962	3736	1374	2927	1466	2474
Usable reads	2113	756	1045	1843	1952	2299	333	652	1368	2432
Reads assigned to taxa in the reference library (>97% similarity)	2044	649	971	1822	535	2090	67	421	1265	1375
Reads assigned to taxa in the public databases (>97% similarity)	49	102	67	13	1247	143	49	35	53	784

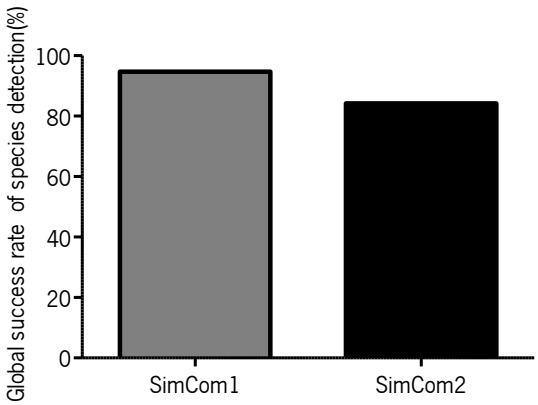

Figure A 1 Global success rate of species detection for 19 total species of simulated macrobenthic communities (excluded the 2 recalcitrant species).

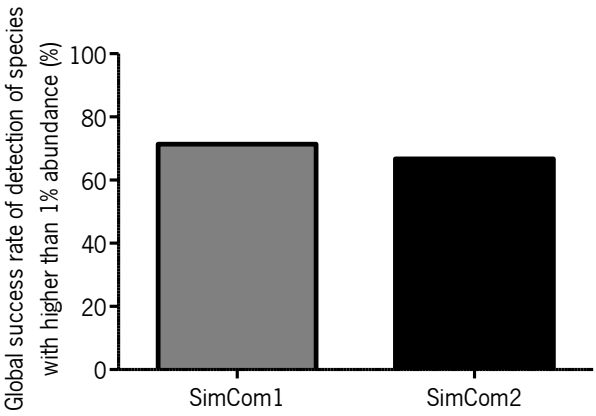

Figure A 2 Global success rate of detection of species excluding singletons.

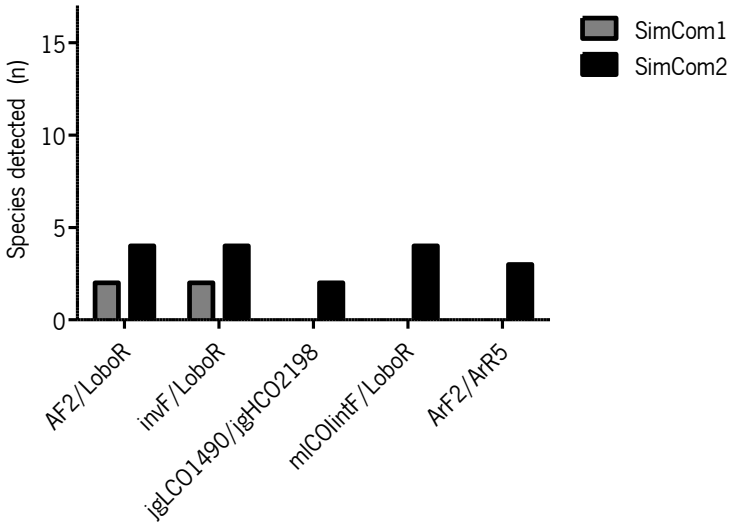

Figure A 3 Number of species detected which were not present in simulated communities, excluding singletons.

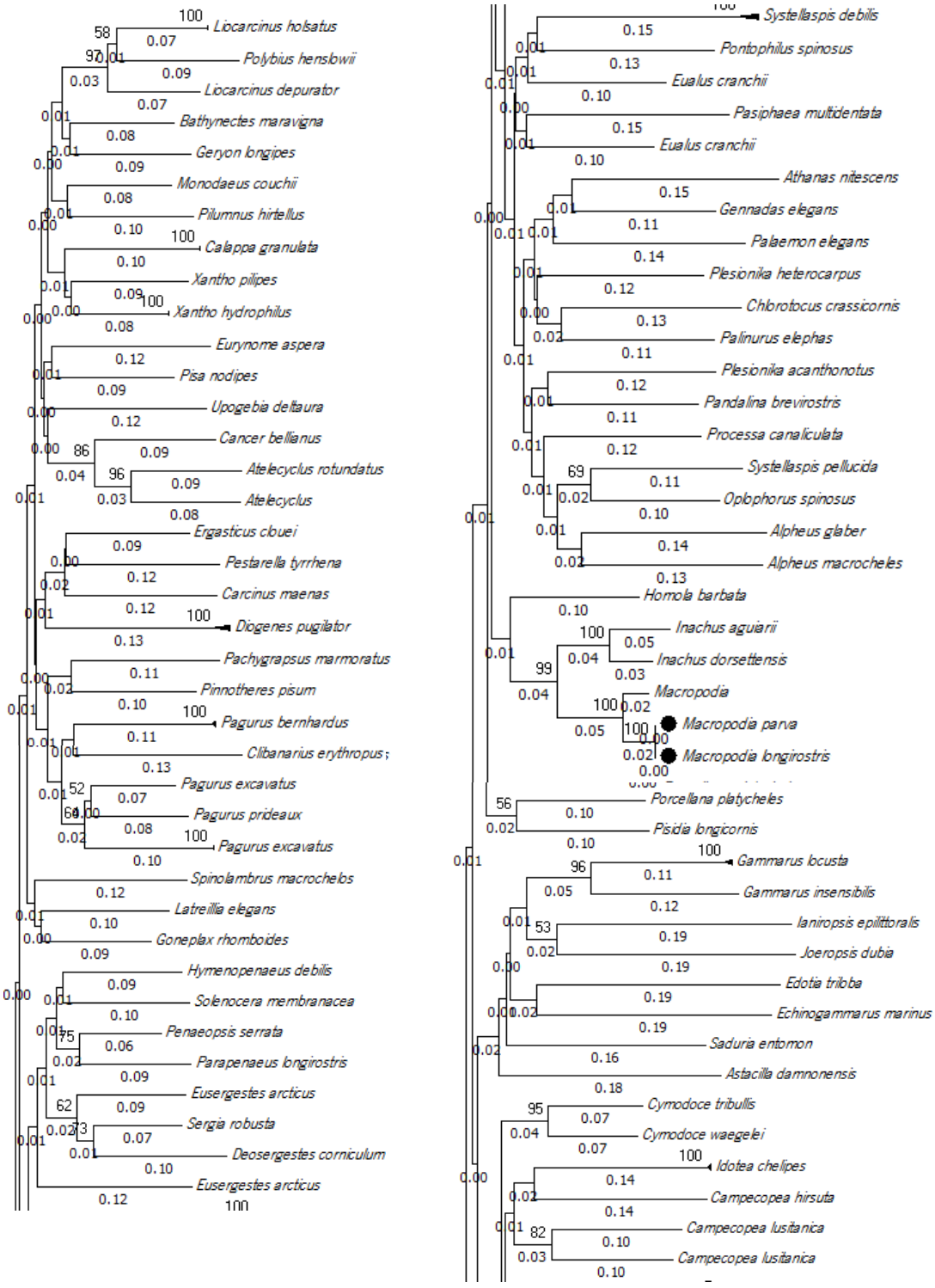

Figure A 4 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/LoboR (418 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree.

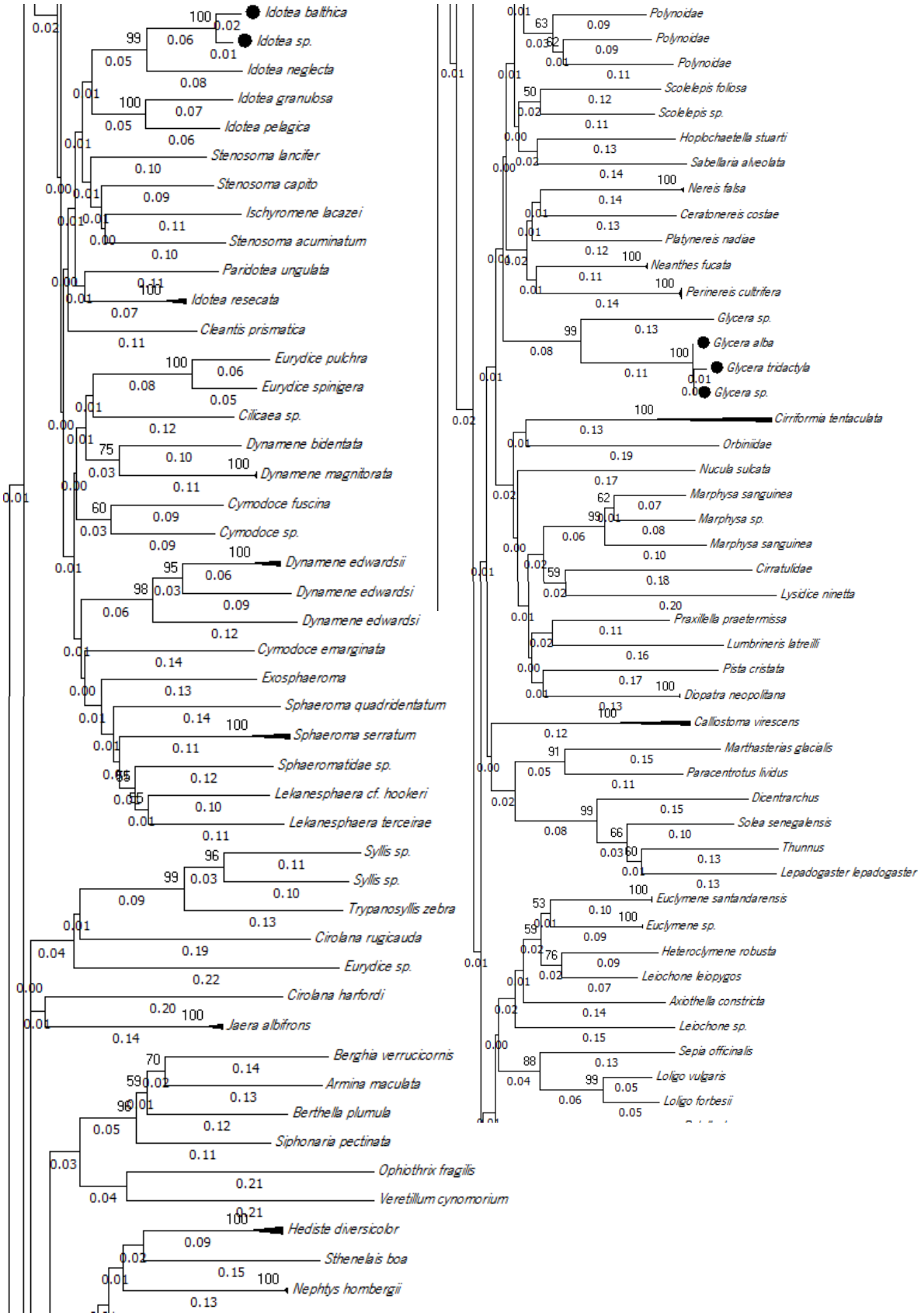

Figure A 4 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/LoboR (418 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

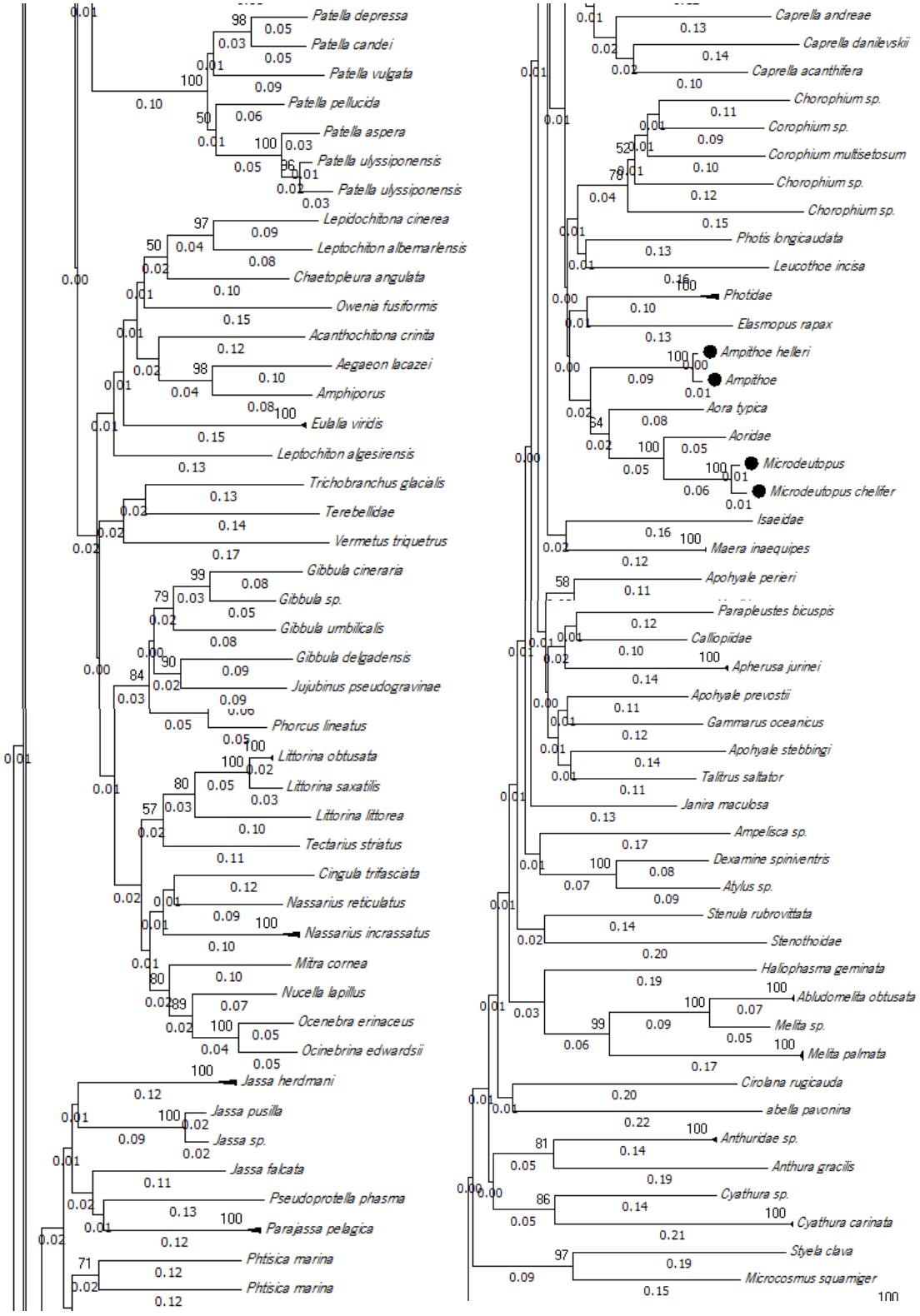

Figure A 4 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/LoboR (418 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

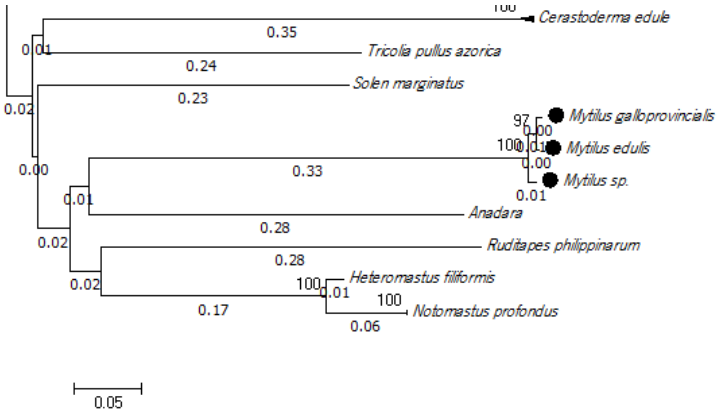

Figure A 4 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/LoboR (418 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

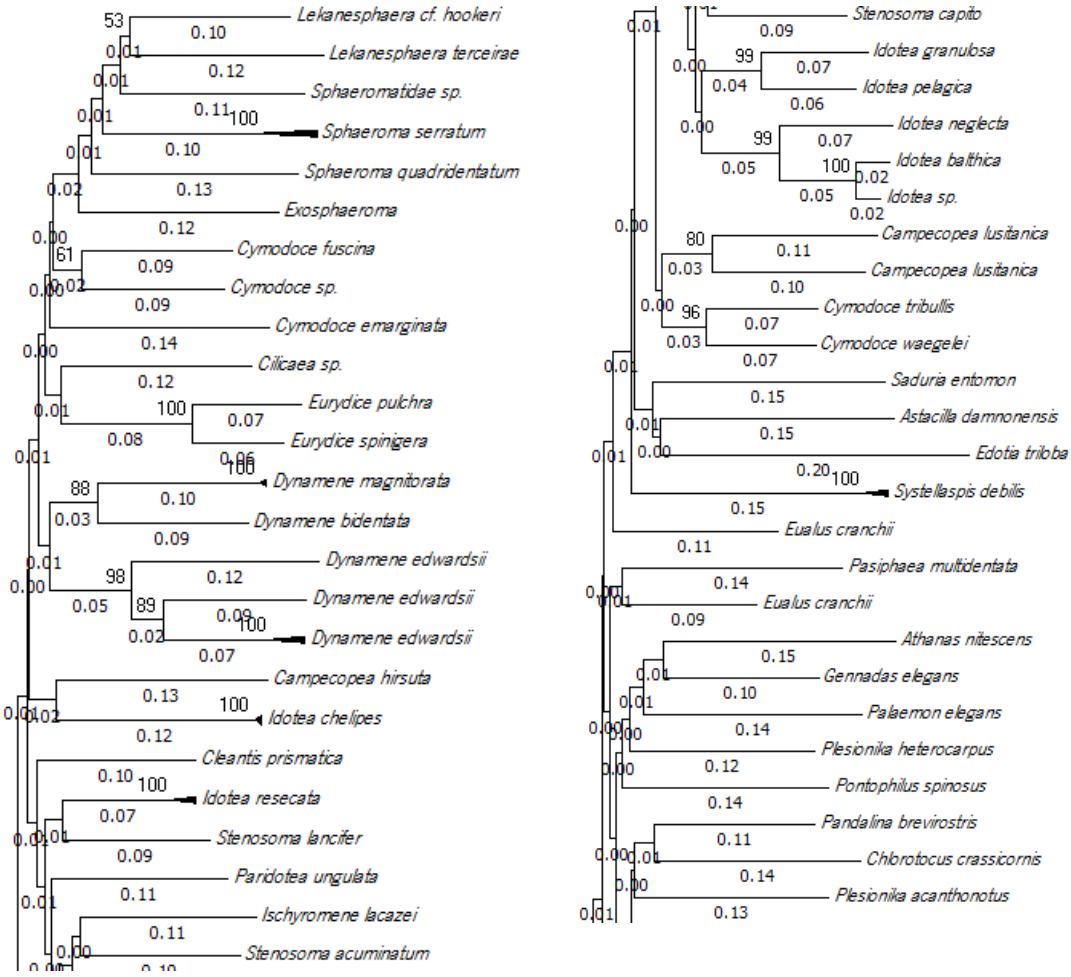

Figure A 5 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair invF/LoboR (470 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree.

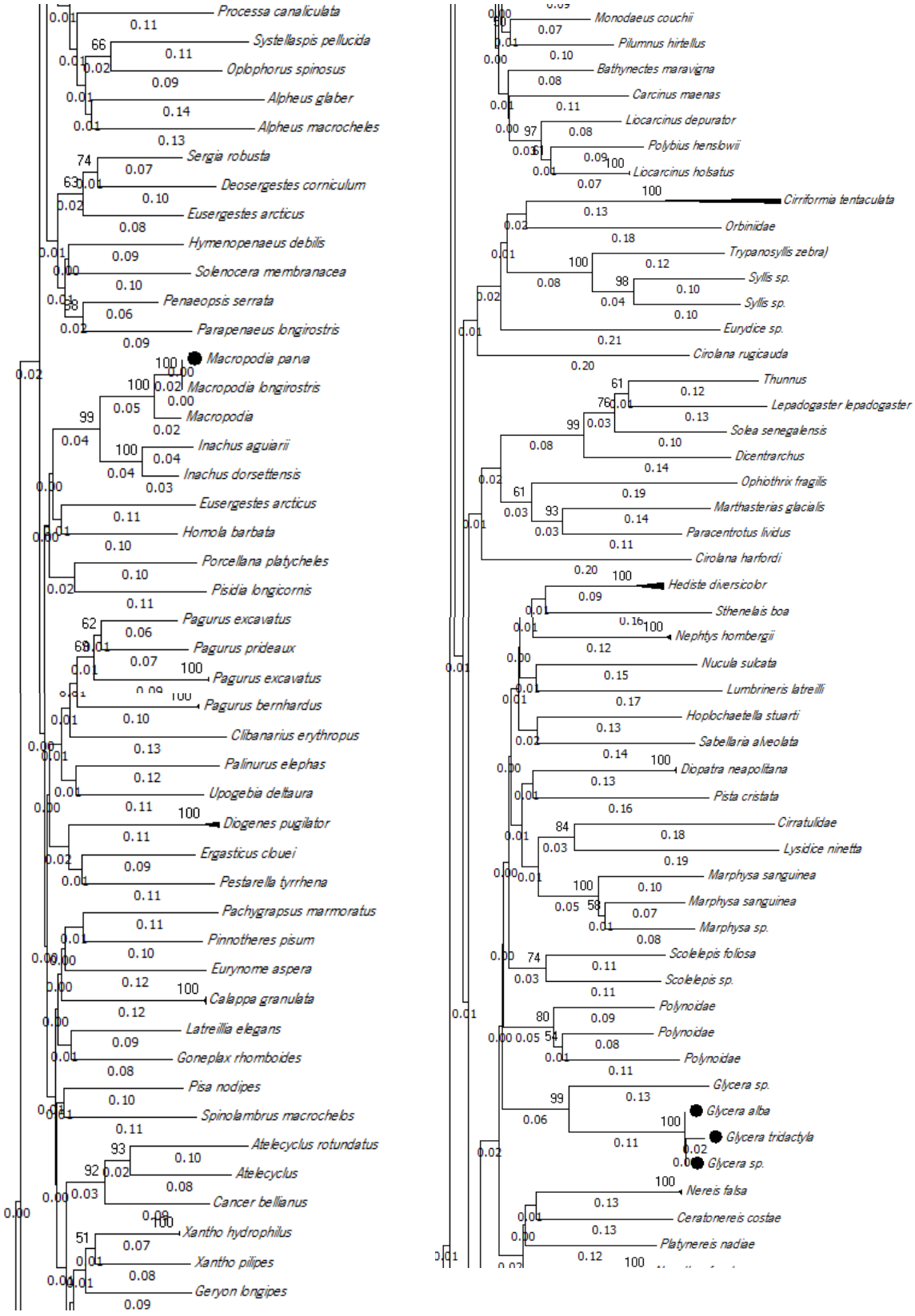

Figure A 5 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair invF/LoboR (470 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

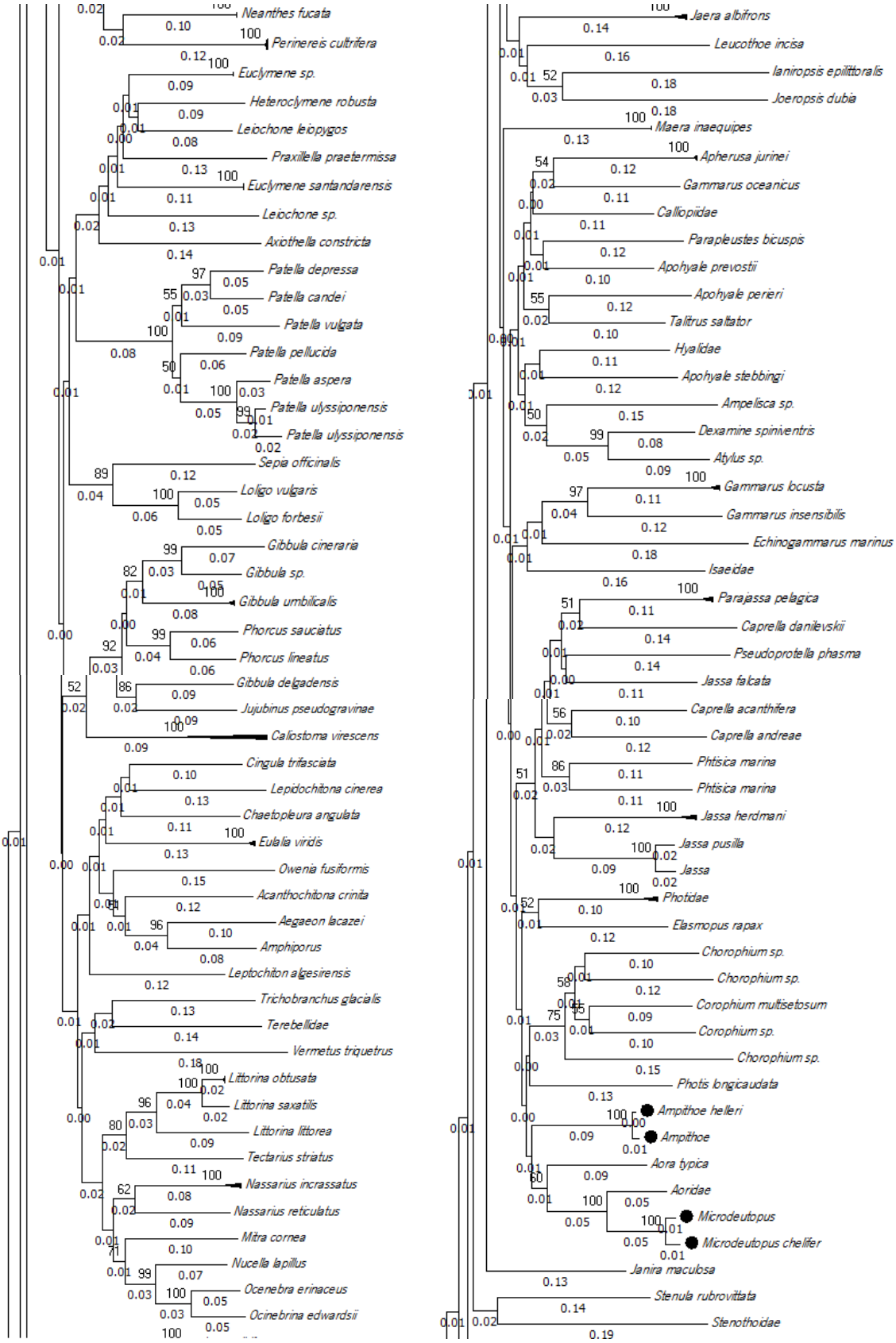

Figure A 5 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair invF/LoboR (470 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

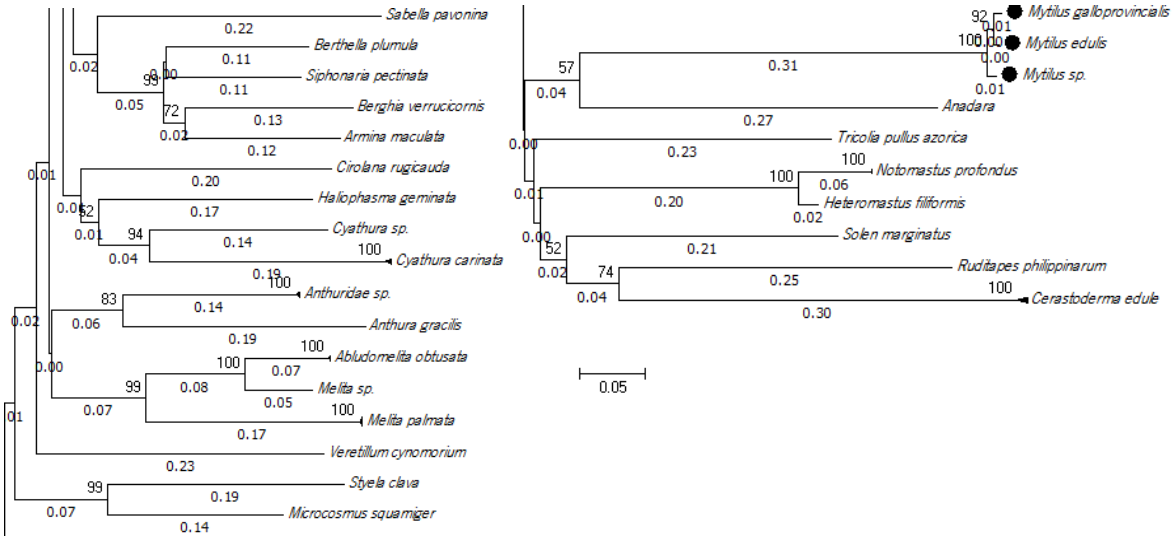

Figure A 5 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair invF/LoboR (470 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

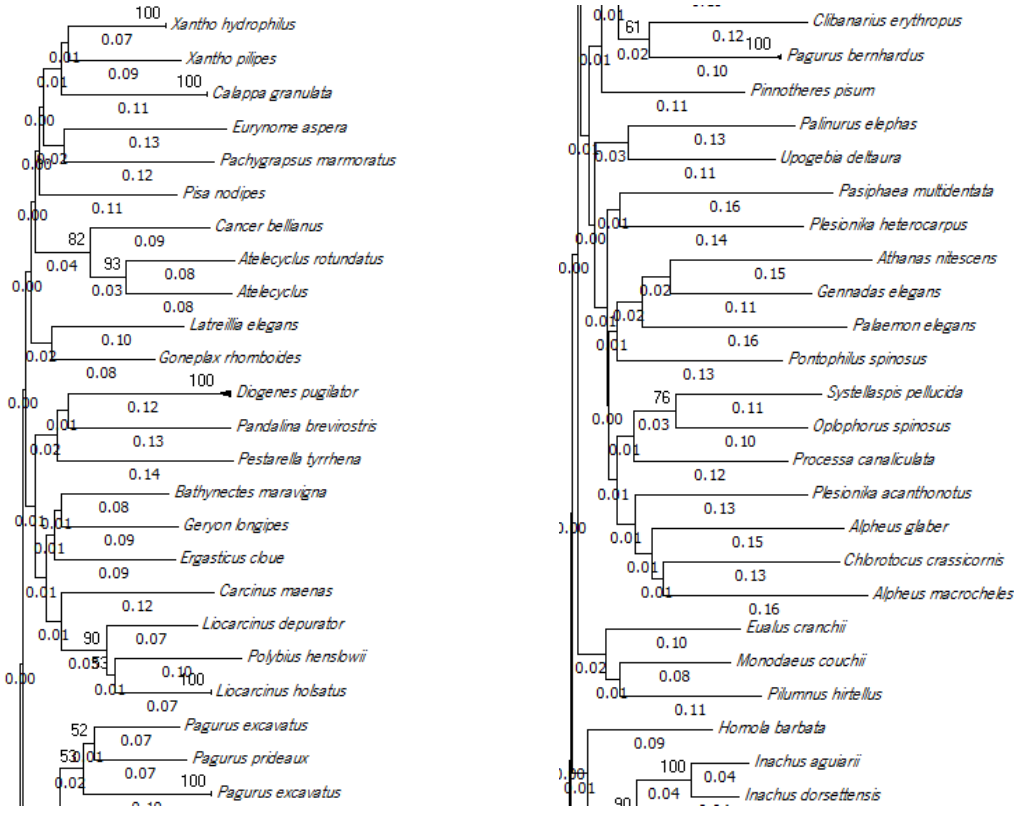

Figure A 6 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair mlCOIintF/LoboR (313 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree.

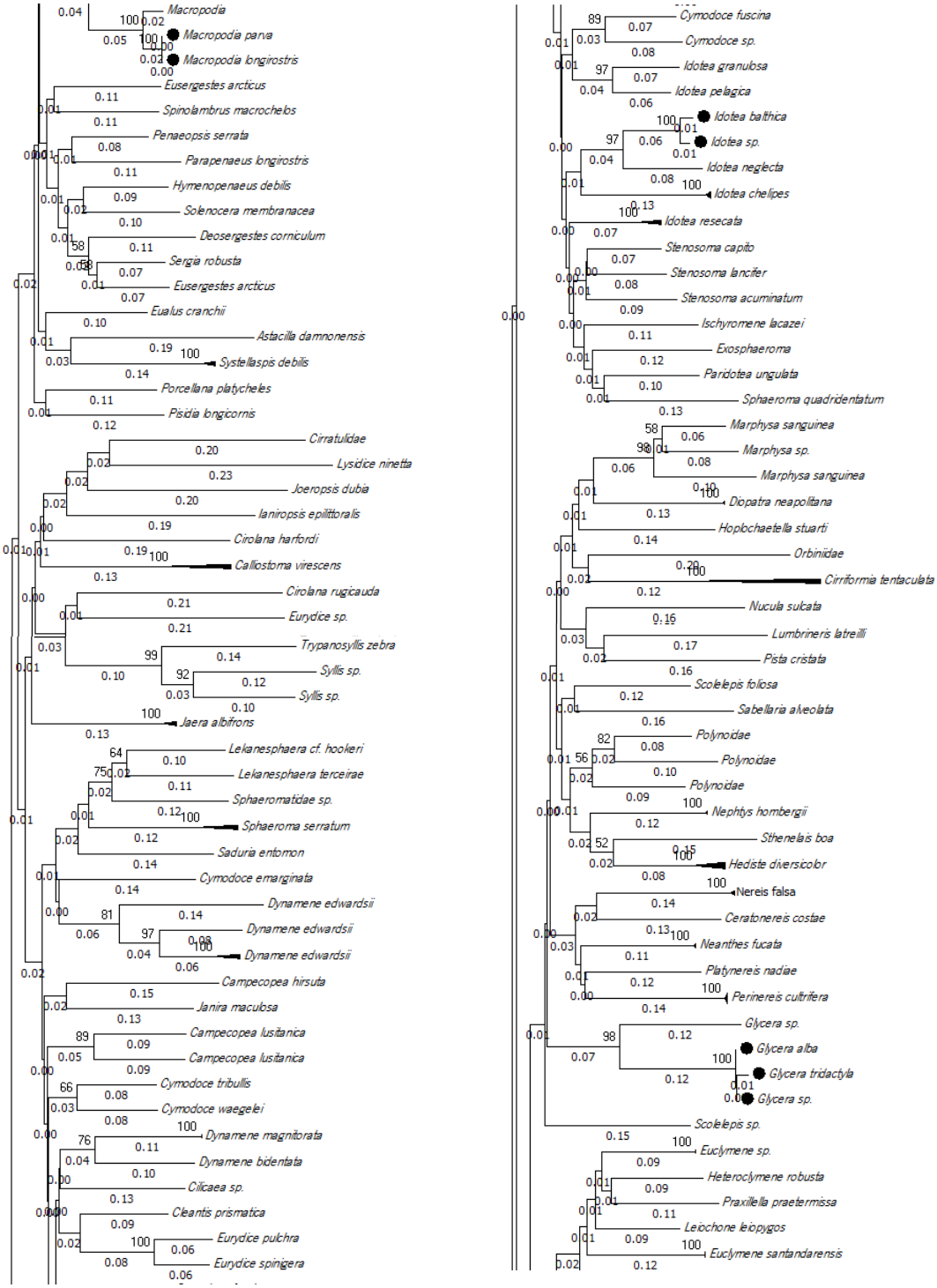

Figure A 6 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair mlCOIintF/LoboR (313 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

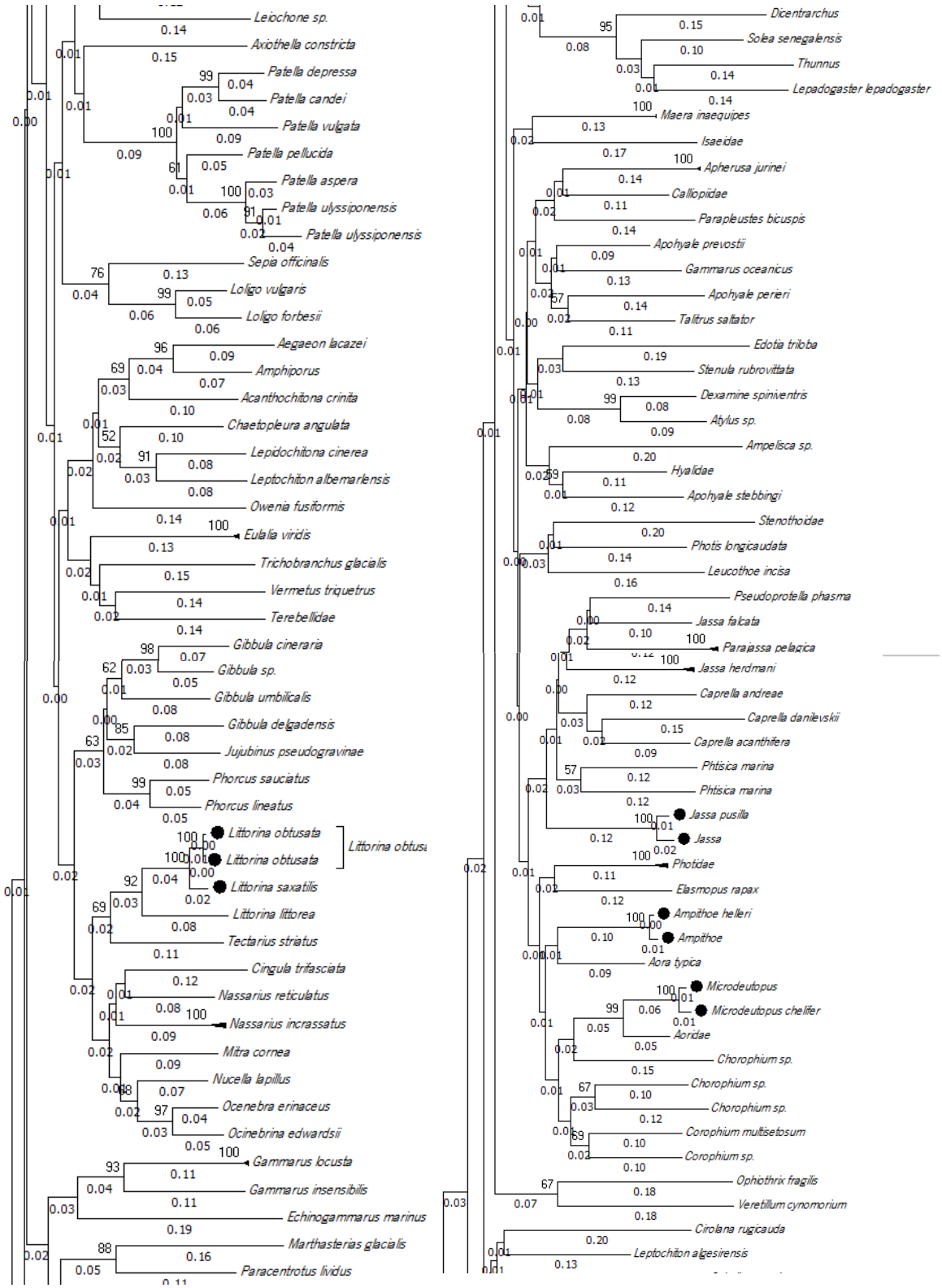

Figure A 6 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair mlCOIintF/LoboR (313 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

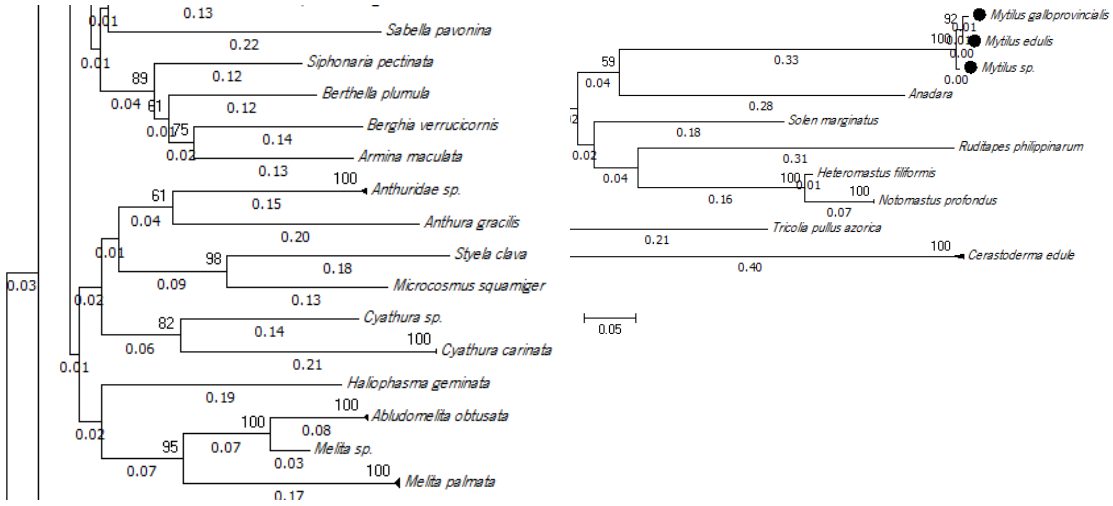

Figure A 6 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair mlCOIintF/LoboR (313 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

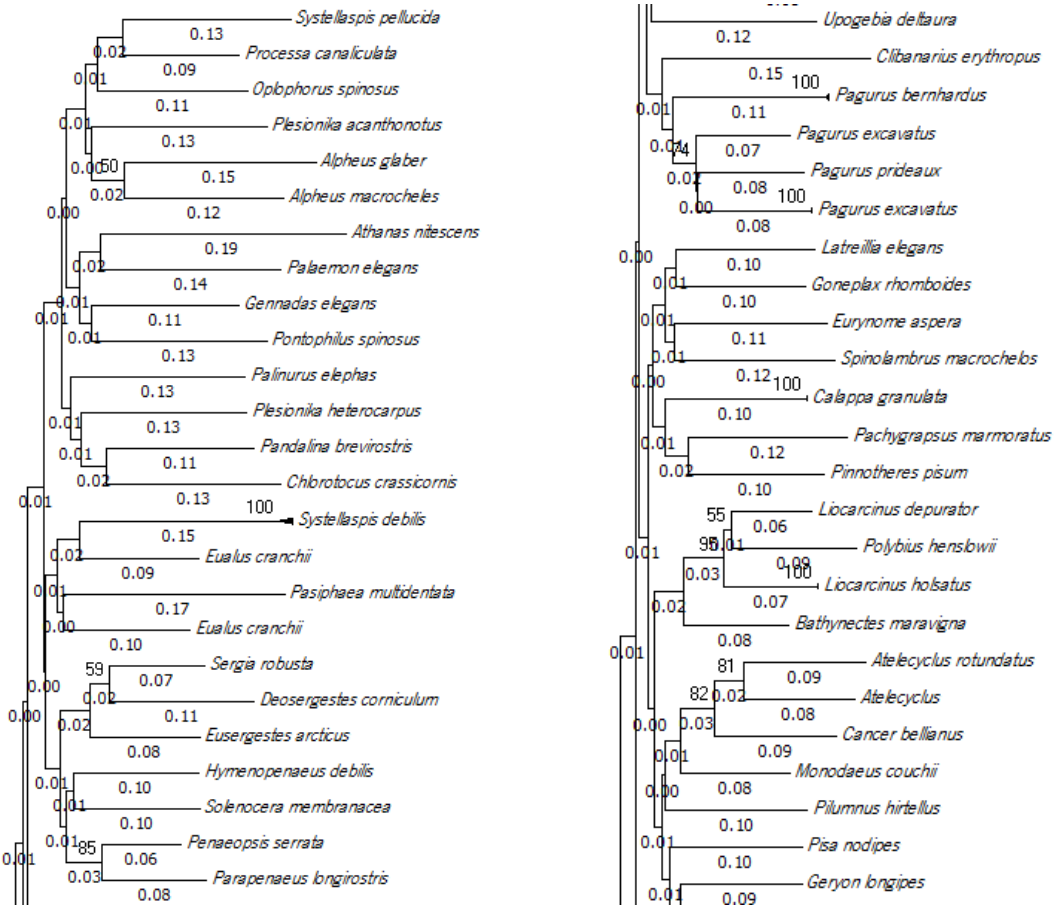

Figure A 7 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/ArR5 (310 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree.

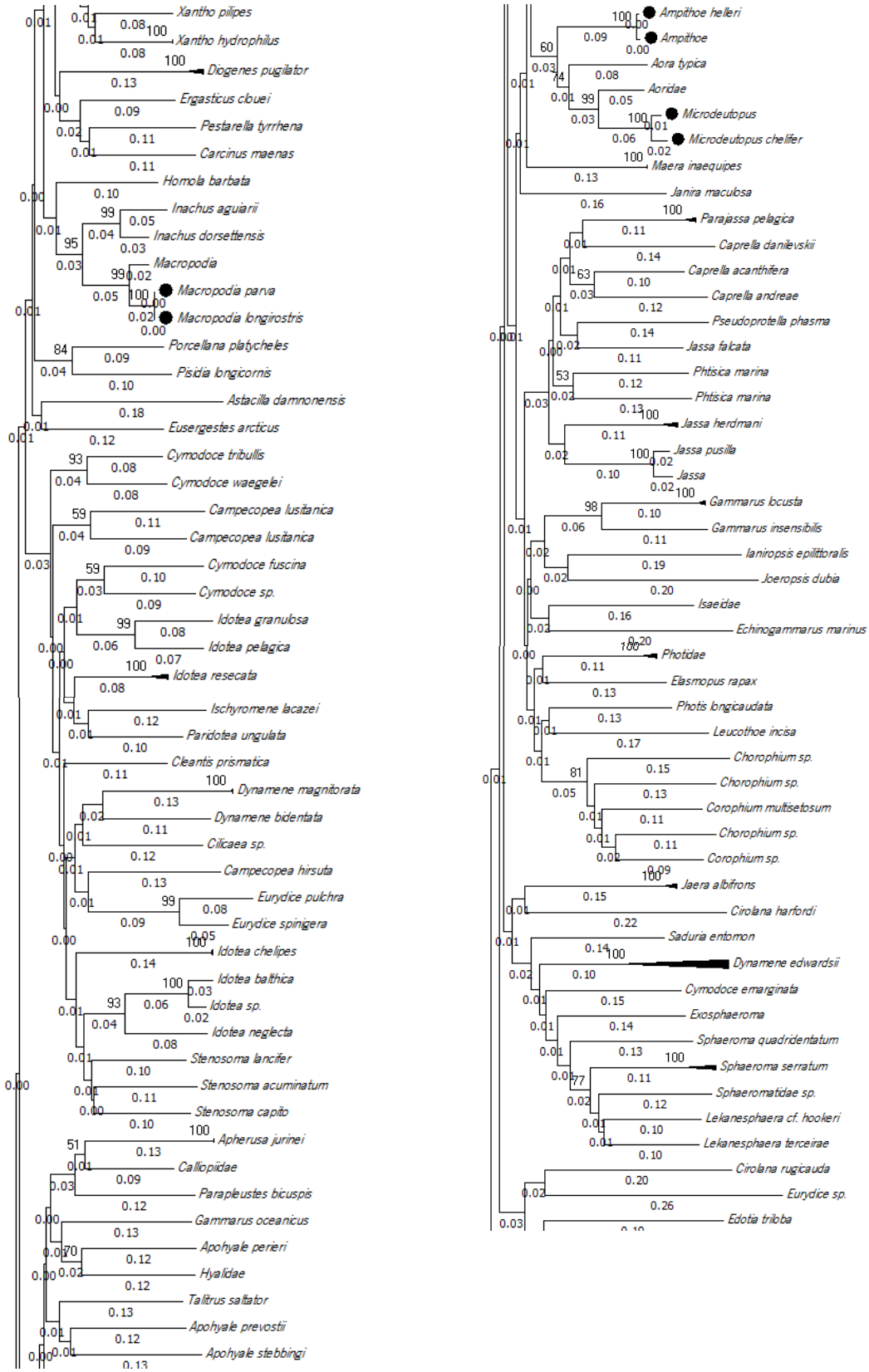

Figure A 7 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/ArR5 (310 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

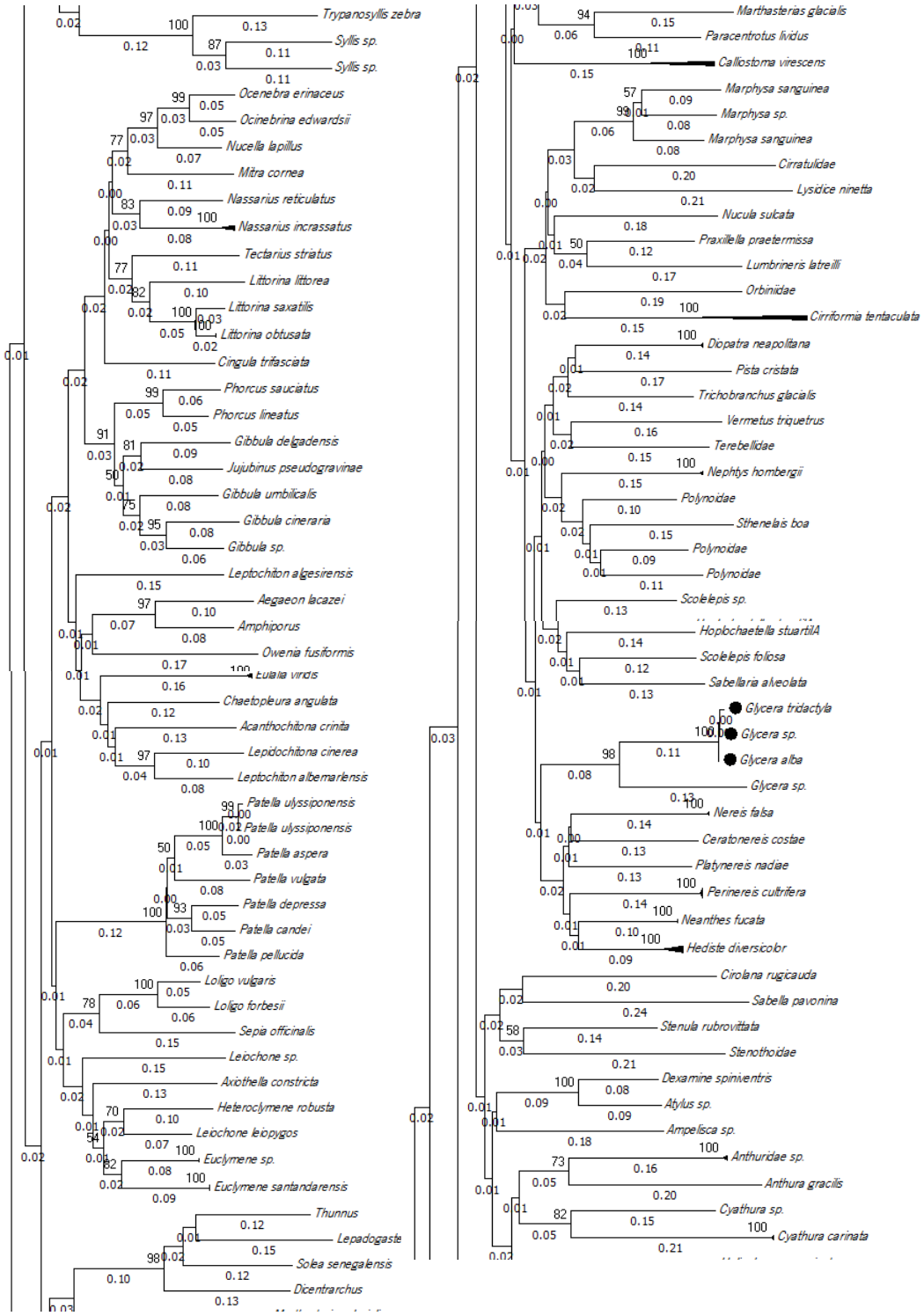

Figure A 7 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/ArR5 (310 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)

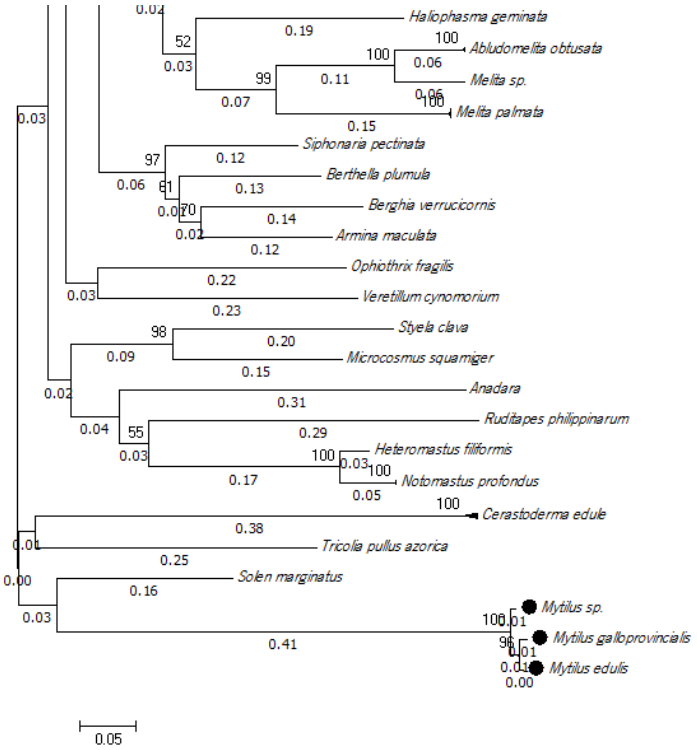

Figure A 7 Phylogenetic NJ tree created from 315 sequences of full COI-5P DNA barcodes clipped with the primer pair ArF2/ArR5 (310 bp) of our reference library. The NJ method was used and the node support was assessed through 1000 bootstrap replicates. ●- Species non discriminated by morphological analyses and species non discriminated in phylogenetic tree. (continued)