

Universidade do Minho
Escola de Ciências

João Tadeu Silva Fontes

**Improving auditing and annotation of DNA
barcode reference libraries of animal COI
sequences for Molecular Ecology applications**



Universidade do Minho
Escola de Ciências

João Tadeu Silva Fontes

**Improving auditing and annotation of DNA
barcode reference libraries of animal COI
sequences for Molecular Ecology
applications**

Master Thesis
Master in Ecology

Work developed under the supervision of
Professor Doctor Filipe José Oliveira Costa
and
Doctor Pedro Alexandre Dias Soares

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho

Atribuição-NãoComercial-SemDerivações



CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisors, Professor Filipe Costa and Doctor Pedro Soares, for accepting me from the beginning and supporting me at every step of the way with patience, helpful guidance, constructive suggestions and encouragement.

I would also like to thank Doctor Pedro Vieira for having such a huge part in the completion of this project, and for always being extremely thoughtful whenever I needed any help.

I would also like to express my profound appreciation for my parents, without whom none of my academic accomplishments would even be possible to begin with. I want to thank them for their unwavering support, motivation and understanding throughout my entire academic life.

I also have to show the biggest appreciation for my girlfriend, Cátia, for fully supporting me at every time and always managing to calmly help me deal with every stressful situation.

I would also like to thank my high school friends and my university friends for helping alleviate my stress and providing me with fun moments whenever I needed.

I also want to thank all members of the ME-Barcode group for always being so courteous and kind with me.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Título: Otimização da auditoria e anotação de bibliotecas de referência de DNA *barcodes* de sequências COI de animais para aplicações em Ecologia Molecular

RESUMO

A descoberta e a descrição da biodiversidade na Terra constituem um capítulo contínuo e incompleto da atividade científica. Felizmente, os estudos de biodiversidade têm beneficiado de abordagens moleculares como o *DNA (meta)barcoding*, que constituem instrumentos de identificação eficientes para a biomonitorização e conservação. A precisão das identificações de espécies, e a sua abrangência taxonômica, dependem de bibliotecas de referência bem representadas. No entanto, a ocorrência de falhas operacionais, ou a incerteza taxonômica nos registros das bibliotecas, pode comprometer a precisão dessas identificações.

Neste estudo foi desenvolvida uma aplicação *web* baseada em *R* - BAGS (*Barcode, Audit & Grade System*) - que executa a auditoria e a anotação automatizada de bibliotecas de *barcodes* da subunidade I do citocromo c oxidase (COI), obtidas através da *Barcode of Life Data System* (BOLD), para um dado grupo taxonômico de animais. A aplicação integra etapas iniciais de controlo de qualidade, bem como a opção de seleção ou exclusão de espécies marinhas das bibliotecas a auditar. A auditoria e triagem por espécie da biblioteca de referência é realizada de acordo com cinco categorias qualitativas (A a E), que dependem dos atributos dos dados e da congruência entre os nomes de espécie e as sequências agrupadas em *Barcode Index Numbers* (BINs). Finalmente, a criação de um relatório de auditoria permite que os utilizadores percecionem rapidamente a qualidade da biblioteca, segregando automaticamente os registos mais úteis e confiáveis de acordo com a sua congruência. Para verificar a performance da precisão da anotação das categorias do BAGS, realizámos testes em três grandes conjuntos de dados: a) peixes marinhos de todo o mundo, b) Chironomidae da Europa (Insecta), e c) anfípodos marinhos de todo o mundo (Crustacea).

Esta ferramenta tem potencial para preencher uma lacuna significativa no paradigma atual das ferramentas de investigação para *DNA barcoding*, através do rastreamento das bibliotecas de referência de forma a avaliar o estado de congruência dos dados e consequentemente, facilitar a triagem de dados ambíguos para serem revistos. Deste modo, o BAGS pode tornar-se um complemento relevante em estudos de *DNA (meta)barcoding*, podendo a longo prazo contribuir para o aumento da qualidade e confiabilidade de bibliotecas de referência.

Palavras-chave: *DNA barcoding*, *DNA metabarcoding*, bibliotecas de referência, *BOLD*, *R*

Title: Improving auditing and annotation of DNA barcode reference libraries of animal COI sequences for Molecular Ecology applications

ABSTRACT

The uncovering and description of Earth's biodiversity constitute an ongoing and incomplete chapter of the scientific endeavour. Fortunately, biodiversity studies have been greatly benefiting from molecular tools, such as DNA (meta)barcoding, which provide efficient identification tools for biomonitoring and conservation programmes. The accuracy of species-level assignments, and the taxonomic span of the identifications, relies on comprehensive DNA barcode reference libraries. However, the occurrence of accidental errors in libraries' records may compromise the accuracy of species' assignments, including the fortuitous operational flaws in the generation of the barcodes, the eventual taxonomic uncertainty or the occurrence of undescribed diversity.

This study describes a web-accessible R-based application - BAGS (Barcode, Audit & Grade System) - that performs automated auditing and annotation of cytochrome c oxidase subunit I (COI) sequences libraries, retrieved from the Barcode of Life Data System (BOLD), for a given taxonomic group of animals. Several initial quality-filtering steps are implemented, as well as the optional filtering of species by their presence in marine and non-marine habitats. This is followed by the auditing and sorting of the barcode records for each species in the library, according to five qualitative grades (A to E) that depend on the attributes of the data and congruency of species names with sequences clustered in Barcode Index Numbers (BINs). Finally, BAGS' reporting tool allows researchers to quickly audit and set aside the most useful and reliable data from the reference libraries, highlighting and segregating records according to their congruency. To verify BAGS' performance and accuracy in grade annotation, successful tests were carried out in three large datasets comprising a) marine fishes of the world, b) Chironomidae of Europe (Insecta), and c) marine Amphipoda of the world (Crustacea).

BAGS has the potential to fulfil a significant gap in the current landscape of DNA barcoding research tools by quickly screening reference libraries to gauge the congruence status of data and facilitate the triage of ambiguous data for posterior review. Thereby, BAGS may become a valuable addition in forthcoming DNA (meta)barcoding studies, in the long term contributing to globally improve the quality and reliability of the public reference libraries.

Keywords: DNA barcoding, DNA metabarcoding, reference libraries, BOLD, R

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	III
RESUMO	V
ABSTRACT	VI
LIST OF FIGURES	IX
LIST OF TABLES.....	X
LIST OF ABBREVIATIONS AND ACRONYMS.....	XI
1. INTRODUCTION.....	12
1.1. DNA BARCODING AND DNA METABARCODING	12
1.1.1. The taxonomic impediment.....	12
1.1.2. DNA barcoding	13
1.1.3. DNA metabarcoding.....	15
1.2. DNA BARCODE REFERENCE LIBRARIES	16
1.2.1. Overview of DNA barcode reference libraries	16
1.2.2. Possible errors and discordances in reference libraries.....	17
1.2.3. Validation and curation of DNA barcode reference libraries	17
1.2.4. Molecular Operational Taxonomic Units and the Barcode Index Number	18
1.2.5. Auditing and annotation system for reference libraries	20
1.3. OBJECTIVES.....	22
2. METHODOLOGY.....	23
2.1. DEVELOPMENT OF THE R SCRIPT AND WEB APPLICATION	23
2.1.1. Overview of BAGS	23
2.1.2. BAGS pipeline.....	24
2.1.2.1. Data mining and library compilation	24
2.1.2.2. Auditing and annotation.....	24
2.1.2.3. Output and annotation-based file sorting.....	26
2.1.2.4. Informatic implementation.....	26
2.2. PERFORMANCE ASSESSMENT	27
2.2.1. Marine taxa selection or exclusion filters test	27
2.2.2. Grade assignment test.....	28
3. RESULTS.....	29
3.1. PERFORMANCE TESTS	29
3.1.1. Marine taxa selection or exclusion filter	29
3.1.2. Trial datasets.....	29
4. DISCUSSION.....	32
4.1. BAGS PERFORMANCE ASSESSMENT TESTS.....	32
4.1.1. Marine taxa selection or exclusion filter test.....	32
4.1.2. Trial datasets.....	33
4.1.2.1. Cases of possible hidden diversity (Grade C).....	33

4.1.2.2. Cases of discordance (Grade E).....	36
5. CONCLUSIONS AND FUTURE PERSPECTIVES.....	38
5.1. CONCLUSIONS AND BAGS LIMITATIONS	38
5.2. FUTURE PERSPECTIVES.....	39
6. BIBLIOGRAPHY.....	41
7. ANNEXES	49

LIST OF FIGURES

Figure 1 – Overview of the main stages of the DNA barcoding workflow for production of reference DNA barcode records.	13
Figure 2 - Position of the COI gene in the mitochondrial genome (Adapted from Trivedi et al., 2016).....	14
Figure 3 - Overview of the main stages of the DNA metabarcoding workflow for production of reference DNA barcode records.	15
Figure 4 - Two distinct outcomes of the BIN assignment pipeline, A: assuming there is a complete congruency between the morphospecies and their respective BINs (i.e. MOTUs); B: assignment of a single morphospecies to several BINs, suggesting the existence of hidden taxonomic diversity.	20
Figure 5 - Overview of BAGS' four main features and their arrangement along the informatics pipeline.	23
Figure 6 - Workflow for automated auditing and annotation of qualitative grades to each species in a BAGS compiled reference library (adapted from Oliveira et al. 2016).....	25
Figure 7 - Print screen of BAGS home page.....	27
Figure 8 - Barplots displaying the distribution of the number of species assigned to each qualitative grade for the three taxonomic groups tested. From top to bottom: marine Amphipoda, Chironomidae and marine fish (Actinopterygii, Elasmobranchii and Holocephali).....	30
Figure 9 - Number of species per BIN in the grade E dataset generated through BAGS for each tested taxonomic group (marine Amphipoda, Chironomidae and marine fish).	31
Figure 10 - Subset of the neighbour-joining tree created for the grade C species belonging to the Chironomidae reference library.	35
Figure 11 - Subset of the neighbour-joining tree created for the grade C species belonging to the marine fish reference library.	35
Figure 12 - Subset of the neighbour-joining tree created for the grade C species belonging to the marine Amphipoda reference library.....	36

LIST OF TABLES

Table 1 - Number of specimens, species and BINs for each of the three libraries created with BAGS for the family Palaemonidae.	29
Table 2 - Number of specimens, species and BINs for each of the three libraries used for the grade assignment test (Marine Amphipoda, Chironomidae, Marine fish).	29
Table 3 - Percentage of monophyletic or non-monophyletic tested species assigned to grade C of each tested taxonomic group, according to their position in the Neighbour-Joining trees constructed.	31
Table 4 - Percentage of the different plausible origins for the assignment of grade E to species for each tested taxonomic group.	31

LIST OF ABBREVIATIONS AND ACRONYMS

BAGS – Barcode Audit & Grade System;

Bp – Base pair;

BIN – Barcode Index Number;

BOLD – Barcode of Life Data system;

COI – Cytochrome c oxidase subunit I;

DNA - Deoxyribonucleic acid

eDNA – Environmental DNA;

GBIF - Global Biodiversity Information Facility;

HMM – Hidden Markov Model;

HTS – High-throughput Sequencing;

IDE – Integrated Development Environment;

MCL – Markov Cluster Algorithm;

MOTU – Molecular Operational Taxonomic Unit;

mtDNA – Mitochondrial DNA;

OTU – Operational Taxonomic Unit;

QA – Quality Assurance;

QC – Quality Control;

WoRMS – World Register of Marine Species;

1. INTRODUCTION

1.1. DNA barcoding and DNA metabarcoding

1.1.1. The taxonomic impediment

The sheer scale of the Earth's biosphere and the percentage of which that is still awaiting discovery, make the study of biodiversity one of the most fundamental subjects in science. Although the question of how many species there are on Earth is still largely unanswered, some recent and extreme estimates claim between 1 to 6 billion species in total, although the majority of them would be bacteria (Larsen, *et al.*, 2017). On the other hand, estimates for the existing number of eukaryotic species alone have been placed in the range of 2 to 100 million species (Costello *et al.*, 2012). One mathematical model estimated that the number of eukaryotic species is ~8.7 million, ~2.2 million of which are marine, suggesting that approximately 86% non-marine species and 91% of marine species, are yet to be described (Mora *et al.*, 2011). This huge gap in our knowledge, coupled with the fact that species extinction rates have become between 100 and 1000 times greater than they were during pre-human history (Lamkin & Miller, 2016) due to phenomena such as pollution (Maiti & Chowdhury, 2013) or climate change (Bellard *et al.*, 2012), encourage the study of biodiversity and taxonomy to become more accurate and comprehensive.

In spite of the fact that a huge percentage of the Earth's biodiversity is currently undiscovered, the number of described species is already providing taxonomists with large volumes of data to work with. So far, the Catalogue of Life registers more than 1.8 million living species (Roskov *et al.*, 2019), while the Global Biodiversity Information Facility (GBIF) currently registers close to 3.5 million species names in total, with 43% of those overlapping with the Catalogue of Life (GBIF Backbone Taxonomy, 2020). These numbers include both extensively studied and described species, as well as species which lack proper description and detail. For instance, each of these biological data bases includes, respectively, 1.7 million and 2.4 million synonym species names, showing that ambiguity and redundancy increase along with the unravelling of biodiversity. Similarly, the World Register of Marine Species (WoRMS) currently registers almost 500,000 species, with only around 240,000 of these having an accepted species name (WoRMS, 2019). Additionally, with the establishment of modern taxonomic methodologies through the use of molecular and computational tools, biodiversity has definitely moved on to the field of big data. Thus, the study of biodiversity is relying progressively more on a proper auditing of biological data, especially in the case of biological databases (Moudrý & Devillers, 2020; Horta *et al.*, 2007; Blair *et al.*, 2020; Ball-Damerow *et al.*, 2019).

1.1.2. DNA barcoding

Our knowledge about biodiversity, taxonomy and phylogenetics has been greatly increasing in recent years, largely due to the development of useful molecular and computational tools such as DNA barcoding and metabarcoding (DeSalle & Goldstein, 2019; Subbotin *et al.*, 2018; Vieira *et al.*, 2019; Weber *et al.*, 2019). DNA barcoding (Figure 1) consists, generally, in sampling an organism, identifying it taxonomically according to its morphology, amplifying and sequencing a short specific region of its genome and subsequently compile the sequence and specimen data in reference libraries and biological data bases such as GenBank or BOLD (Barcode of Life Data system) (Hebert *et al.*, 2003; Ratnasingham & Hebert, 2007; Costa & Antunes, 2012; Cariani *et al.*, 2017; Sayers *et al.*, 2019). Conventionally, DNA barcoding utilizes Sanger sequencing as the standard approach to identify individual specimens, sequencing only one DNA amplicon at a time (Sanger *et al.*, 1977; Shokralla *et al.*, 2014). However, given the current refinement of DNA sequencing methodologies (e.g. High-throughput sequencing [HTS]), extremely large volumes of sequence data are being identified through searches

for matching sequences in BOLD and GenBank (Meiklejohn *et al.*, 2019; Porter & Hajibabaei, 2018). This is made possible because HTS technologies are capable of sequencing millions of DNA fragments simultaneously per run (Shokralla *et al.*, 2014). Therefore, researchers are enabled to access an unprecedented amount of data and conduct analysis using sequences generated from independent sources, granting them access to sequence data on most taxonomic groups, from most geographic regions (Ratnasingham and Hebert, 2007; Hebert *et al.*, 2016). Consequently, DNA barcoding has been consistently establishing itself as one of the primary driving forces behind the uncovering of biodiversity (Costa & Antunes, 2012; DeSalle & Goldstein, 2019; Trivedi *et al.*, 2016), as well as a very important

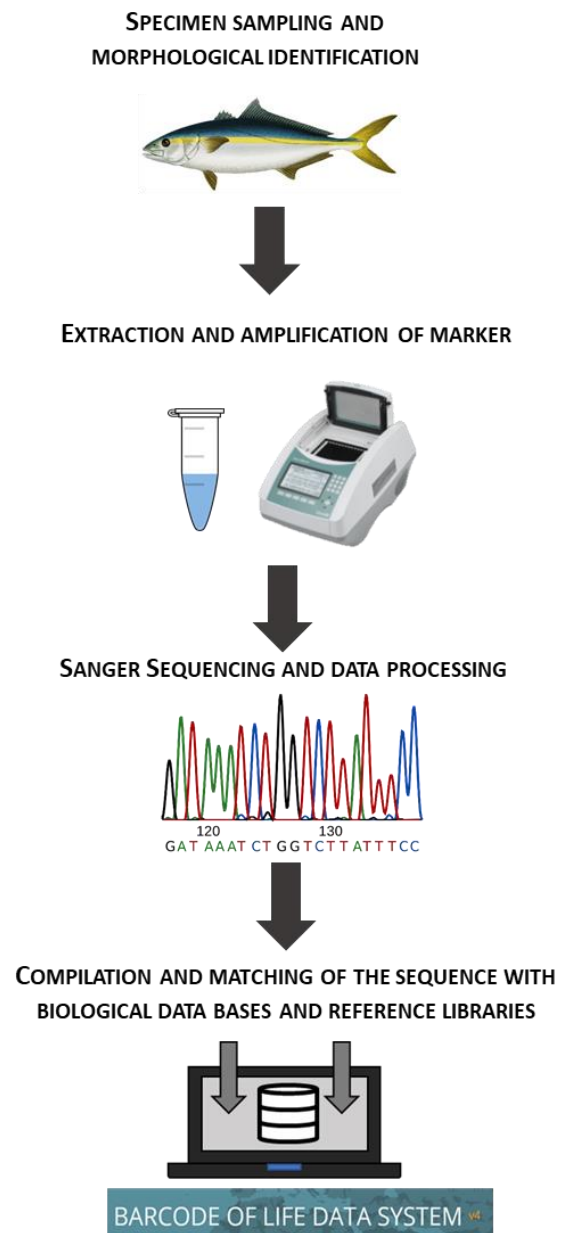


Figure 1 – Overview of the main stages of the DNA barcoding workflow for production of reference DNA barcode records.

tool for discovering evolutionary patterns and the phylogenetic relations between taxa (Costa & Carvalho, 2010; Liu *et al.*, 2017; Subbotin *et al.*, 2018; Wong *et al.*, 2011).

In the particular case of the animal kingdom, the standardized marker for DNA barcoding consists of a ~650 base pair (bp) DNA section from the 5' end of the mitochondrial gene coding for cytochrome *c* oxidase subunit I (COI-5P or COI; Figure 2). The reasoning behind choosing a gene from the mitochondrial genome (mtDNA) over the nuclear genome relies on several factors, namely that mtDNA is non-recombinant in most organisms, is transmitted mostly through haploidy, lacks the presence of introns and has a high substitution rate (Hebert *et al.*, 2003; Saccone *et al.*, 1999). Moreover, usually COI sequences do not possess indels, making operations such as sequence alignments easier to compute and run analysis on (Hebert *et al.*, 2003; Mardulyn & Whitfield, 1999). In addition, the usefulness of this genetic marker is also corroborated by the large volumes of sequences being uploaded to data bases, with over 2.5 million COI sequences in GenBank alone (Porter & Hajibabaei, 2018). Given all these factors, the COI gene has been extensively used in barcoding studies (Durand & Borsa, 2015; Kress *et al.*, 2015; Subbotin *et al.*, 2018), proving its worth as a marker for both uncovering biodiversity and biomonitoring.

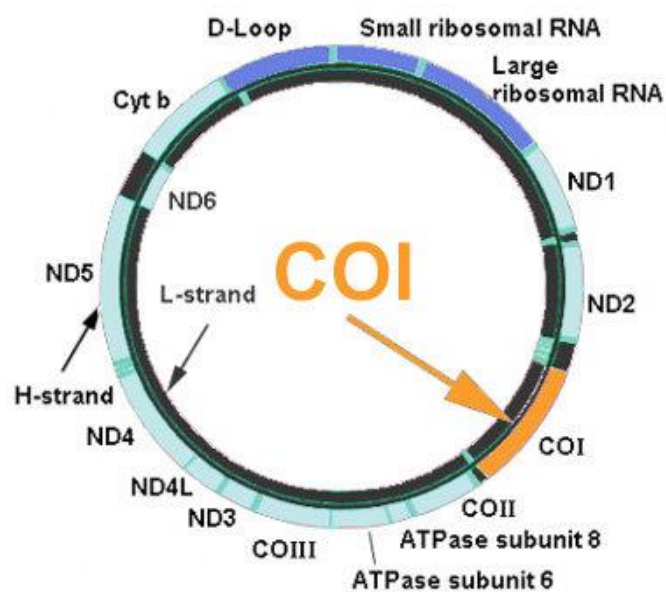


Figure 2 - Position of the COI gene in the mitochondrial genome (Adapted from Trivedi *et al.*, 2016).

1.1.3. DNA metabarcoding

DNA metabarcoding (Figure 3) takes one step further in comparison to DNA barcoding, by targeting whole communities or environmental bulk samples (e.g. soil, water, air, faeces, etc.), as opposed to a single specimen. This is followed by extracting DNA from the community or from the environmental DNA (eDNA), which comprises a complex mixture of genetic material from many different organisms (Taberlet, *et al.*, 2012). Once the (e)DNA is extracted, it is amplified using primers with a broad taxonomic range and the sequence reads are generated through HTS (Ruppert *et al.*, 2019). This approach is carried out in order to assess the composition, diversity and species richness of whole communities of organisms. (Djurhuus *et al.*, 2018; Leese *et al.*, 2018; Taberlet *et al.*, 2012).

Given the high diversity and high number of species present in either bulk community DNA or eDNA, this technique especially benefits from HTS analysis, outputting large numbers of reads, with sequences belonging to a myriad of taxonomic groups which coexist within a unique sample (Rimet *et al.*, 2018). This way, (e)DNA allows for the assessment of otherwise elusive or rare species which are difficult to assess using conventional taxonomy, or species which occur in low abundances. Although DNA metabarcoding was originally used to study mainly microbial communities (Sogin *et al.*, 2006), it has been employed for studying several other groups such as invertebrates (Curry *et al.*, 2018; Porazinska *et al.*, 2010), zooplankton (Djurhuus *et al.*, 2018), plants (Hiiesalu *et al.*, 2012) and vertebrates (Brown *et al.*, 2012; Rayé *et al.*, 2011). Additionally, the versatility of DNA metabarcoding allows for more than standard biodiversity assessments, as it is also possible to use (e)DNA to assess the presence and abundance of non-native invasive species (NIS) (Duarte *et al.*, 2020) or even complement the taxonomic identification of fossils (Grealy *et al.*, 2015).

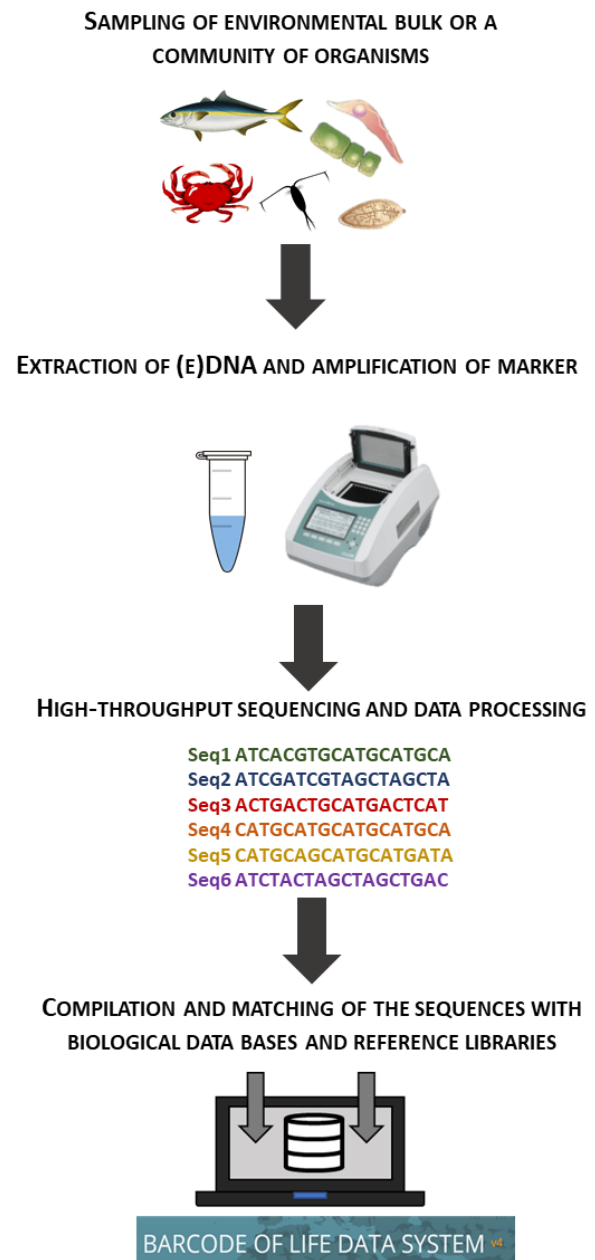


Figure 3 - Overview of the main stages of the DNA metabarcoding workflow for production of reference DNA barcode records.

Since the fundamental basis of DNA metabarcoding relies on a large number of sequences reads with mostly unknown origin, in order to identify them taxonomically, it is imperative to have comprehensive sets of reference sequences.

1.2. DNA barcode reference libraries

1.2.1. Overview of DNA barcode reference libraries

DNA barcode reference libraries consist, generally, in a compilation of DNA barcode sequences for a given group of organisms, as well as their respective metadata and features. In order to have accurate DNA barcode-based species identification, it is essential that well-curated comprehensive reference libraries are available (Leite *et al.*, 2020; Cariani *et al.*, 2017; Leese *et al.*, 2016; Oliveira *et al.*, 2016). The creation of these reference libraries aims to ease species identification by reducing possible errors and incongruencies, as well as aiding in further analysis by optimizing the availability of valid sequence data for researchers. Ultimately, reference libraries allow newly generated barcodes from unknown specimens to be compared with previously published barcode sequences, in order to identify them. Although the primary purpose of constructing comprehensive DNA barcode libraries is to match and compare sequences of unknown origin in order to identify them in biodiversity studies, they can also be used for other purposes. For instance, reference libraries can be utilized in phylogenetic reconstruction (Kress *et al.*, 2015), detecting the illegal use of protected species (Rasmussen & Morrissey, 2008) or for the authentication of animal products in the food industry (Carvalho *et al.*, 2015).

The demand for high-quality barcode reference libraries has increased considerably with the introduction and extended use of DNA metabarcoding for biodiversity assessments and biomonitoring (Leese *et al.*, 2018; Weigand *et al.*, 2019). Due to the large number of reads obtained from HTS instruments, the required bioinformatics often include automated systems to match query sequences to reference sequences in DNA sequence repositories (e.g. Bengtsson-Palme *et al.*, 2018) such as BOLD and GenBank. Furthermore, in the case of metabarcoding studies, since the taxonomy of the organism from which the sequences originated is almost always unknown, the necessity for validated reference libraries gains even more significance.

Typically, DNA barcode reference libraries are used without supervision or quality control of the sequences and specimen data associated with them. There are a few exceptions, such as R-Syst::diatom (Rimet *et al.*, 2016) or MIDORI (Heller *et al.*, 2018), which are reference libraries created and curated for specific taxonomic groups, although overall, the quality control of reference libraries is still lacking. Given the absence of quality control measures, inaccurate records can arise and result in recurrent identification

errors which can be perpetuated over time and across studies without being detected (Keller *et al.*, 2020; Leese *et al.*, 2016; Weigand *et al.*, 2019).

1.2.2. Possible errors and discordances in reference libraries

In order to correctly allocate a newly generated barcode sequence to its respective morphospecies, the first step consists in taxonomically identifying the sampled specimen. However, conventional taxonomic identification tools show clear limitations, especially when they are solely based on species morphology. This happens because frequently, available data on species identification can be incomplete or incorrect. Misidentification can be the result of biological factors, namely the difficulty in differentiating cryptic species (species which present identical morphology, but can be clearly differentiated through molecular tools) (Janzen *et al.*, 2017; Saitoh *et al.*, 2015); homoplasy, which happens when a taxonomic group gains or loses a trait independently in separate lineages of their evolution (Vences *et al.*, 2005); phenotypic plasticity, which are morphological or physiological changes that occur in response to specific environmental factors (Stampar *et al.*, 2017; Weigand *et al.*, 2011); the fact that certain morphological traits only occur during a specific stage of a specie's life cycle (Pegg *et al.*, 2008); recently diverged species and incomplete sorting (Weber *et al.*, 2019); among others. On the other hand, apart from biological errors, there are other sources of inconsistencies and errors that can arise and compromise subsequent analysis. Operational errors can occur due to the mislabelling of DNA barcode sequences; cross contamination of samples with "alien DNA" from a different species; low-quality or short sized sequences or primers; accidental mistakes when recording data; technical errors during sequencing, among others that can easily go unnoticed and become potential liabilities (Packer *et al.*, 2009; Pentinsaari *et al.*, 2019; Rulik *et al.*, 2017).

1.2.3. Validation and curation of DNA barcode reference libraries

In light of these facts, reference libraries need to be validated and curated in order to achieve optimal taxonomic identification by matching newly generated sequences with the existing sequence data in a biological data base. For the purpose of validating a reference library, two main components can be distinguished: Quality assurance (QA), which consists in guaranteeing that the sequences and their respective data and metadata are valid and assigned correctly to a species, achieved by following pre-determined and universal quality standards; and Quality control (QC) which is more user-oriented and consists of cross-validation and the search for possible errors and incongruencies that managed to persist after the previous quality measures (Rulik *et al.*, 2017; Weigand *et al.*, 2019). So far, a few QA and QC criteria have been implemented upstream and along the DNA barcoding pipeline (e.g. Hanner, 2005).

However, the downstream quality control of the taxonomic accuracy in DNA barcode reference libraries has not been implemented in a standardized way. In the BOLD data base, some QA/QC measures are currently implemented such as: Labelling of barcode compliant records, flagging of sequences that are likely contaminations or based of specimens which were misidentified, flagging of sequences with stop codons (Ratnasingham & Hebert, 2007), and the possibility to run BIN-discordance reports (Ratnasingham & Hebert, 2013). Nonetheless, several sources of potential discordance or errors remain unscreened or unexplored in existing systems, with their origins being generally well known (Meiklejohn *et al.*, 2019; Mioduchowska *et al.*, 2018; Siddall *et al.*, 2009; Weigand *et al.*, 2019). Regardless, few systems and studies have addressed the issue of reference library compilation, especially concerning taxonomic reliability. For instance, the “coil” R package (Nugent *et al.*, 2020) helps in detecting incongruencies in animal barcoding and metabarcoding data by placing the sequences in a reading frame and translating them to amino acids. CO-ARBitrator (Heller *et al.*, 2018) detects sequences mislabelled as COI which originate from non-homologous loci. However, although useful, neither of these systems address the issue of taxonomic congruency. A pre-processing system for large dataset has been proposed by Rulik *et al.* (2017), with the goal to generate high quality DNA barcodes by verifying taxonomic consistency. Even so, this system requires a phylogenetic backbone for implementation, and is meant to be used before uploading data to reference libraries, thus not considering global congruence with other data already available in either BOLD or GenBank. A reproducible pipeline for auditing marine eukaryote barcoding sequences has also been proposed by Arranz *et al.* (2020), providing tools for the curation of sequences and detection of synonym species across data bases. Nonetheless, this pipeline is exclusively for marine eukaryotes, thus not being applicable for a large number of taxonomic groups. Moreover, the pipeline is not user-friendly as it requires the user to utilize an extensive set of scripts and softwares in order to complete the entire workflow.

1.2.4. Molecular Operational Taxonomic Units and the Barcode Index Number

Species boundaries are often troublesome and difficult to establish accurately, especially in the case of the biological phenomena previously mentioned such as introgressive hybridization or phenotypic plasticity. Additionally, several well-established species in the past have been found to actually comprise relatively large complexes of species (Durand & Borsa, 2015; Fennessy *et al.*, 2016; Packer *et al.*, 2009; Saitoh *et al.*, 2015; Desiderato *et al.*, 2019; Teixeira *et al.*, 2020). One solution to this issue has been the assignment of presumptive or inferred species, Molecular Operational Taxonomic Units (MOTUs), to groups of sequences. MOTUs are obtained by grouping sequences together through the use of clustering algorithms (Ratnasingham & Hebert, 2013), according to their similarities and pre-determined

parameters. Ideally, in a library or data base of barcode sequences, each MOTU should correspond to a pre-existing morphospecies, thus validating the use of this technique and allowing to easily match newly generated query sequences. Nonetheless, that scenario is not always the case given that, as previously mentioned, certain taxonomic groups can display hidden diversity, where a clear distinction between two species can be achieved algorithmically, but not morphologically (Lin *et al.*, 2015; Vieira *et al.*, 2019). On the other hand, certain species are difficult to identify according to their morphology or require extreme proficiency in taxonomy to do so, which results in misidentifications which can potentially lead to incongruencies in attributing MOTUs.

An approach to carry out the correspondence of MOTUs to existing morphospecies has been implemented in the BOLD platform through the creation of the Barcode Index Number (BIN) system (Ratnasingham & Hebert, 2013). The BIN system implements algorithms that perform single-linkage clustering of DNA barcode sequences, generating groups of sequences clustered in a manner that ideally mirrors a corresponding morphospecies (Ratnasingham & Hebert, 2013). This system is the basis for how animal COI sequences are organized within BOLD, and it overcomes the need for taxonomists to routinely perform species identifications, by assigning a specific number to a given cluster of sequences, which in some cases represent a morphospecies, and in other cases, a cryptic species which is only recognizable through their BIN (Figure 4).

The BIN system workflow consists, generally, in the following steps: 1) Quality checks by excluding sequences with less than 500 bp coverage for the barcode region of COI or with more than 1% of ambiguous bases, checking for stop codons and unlikely peptides, screening for sequences which derive from bacterial or other possible external contaminants, screening for possible chimera sequences; 2) Sequence alignment, where every sequence which passes the previous step is translated to amino acids and aligned to a Hidden Markov Model (HMM) of the COI protein, followed by a back translation to nucleotides in order to perform a multiple sequence alignment; 3) Single Linkage Clustering, that groups together all sequences under a pairwise distance threshold of 2.2%; 4) Markov Clustering which refines the single linkage clustering by collapsing neighbour Operational Taxonomic Units (OTUs) which show a distance of less than 4.4% between each other, followed by using the Markov Cluster Algorithm (MCL) to produce 8 refinement options for each OTU generated; 5) Silhouette Criterion, which takes the 8

candidate refinement options and generates a score for each in order to select the one with maximum score (Ratnasingham & Hebert, 2013).

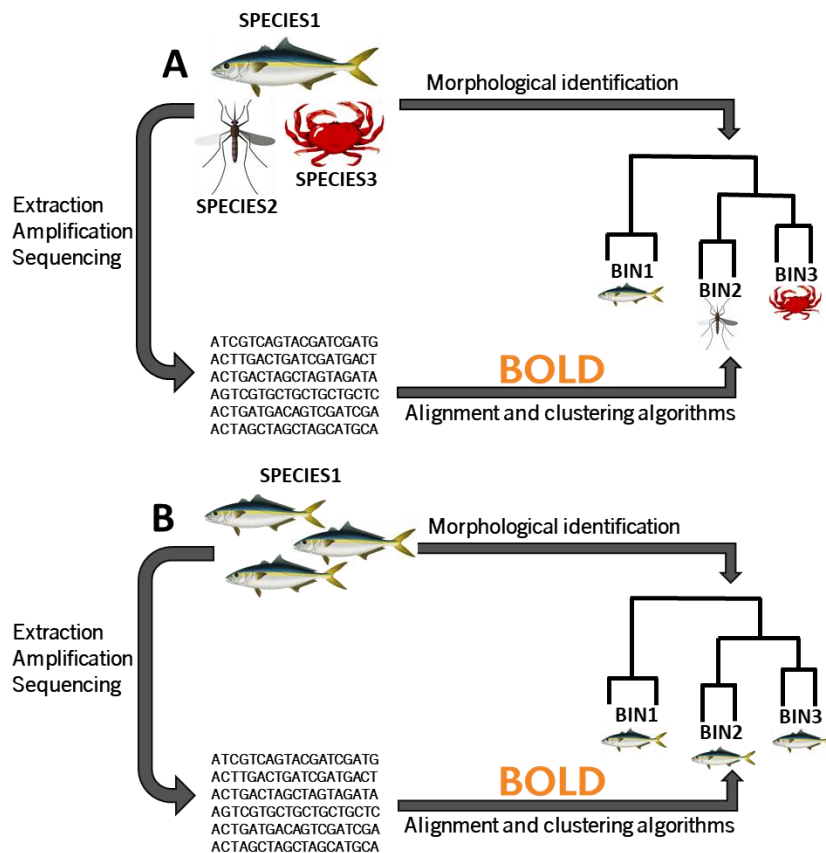


Figure 4 - Two distinct outcomes of the BIN assignment pipeline, A: assuming there is a complete congruency between the morphospecies and their respective BINs (i.e. MOTUs); B: assignment of a single morphospecies to several BINs, suggesting the existence of hidden taxonomic diversity.

1.2.5. Auditing and annotation system for reference libraries

GenBank currently holds a very large number of COI sequences (Porter & Hajibabaei, 2018), which eventually end up being mined to BOLD. Although many of these records do not abide to the formal barcode data standards (Ratnasingham & Hebert, 2013), they are still useful as a resource and should not be overlooked. In fact, many metabarcoding-based studies report taxonomic assignments based on all available COI data, thereby including non-compliant barcode records. This reinforces the need for a compilation, auditing and annotation system which considers all available barcodes at BOLD, in order to provide an indication of the taxonomic reliability of the records for end-users of reference libraries.

Considering these facts and all the intricacies and susceptibilities to errors of DNA barcode reference libraries and biological databases, Costa *et al.* (2012) proposed a ranking system to be implemented at the post-barcoding end of the DNA barcoding pipeline, considering all available sequence data for a given species, both barcode compliant and non-compliant. This ranking system attributes one

of five qualitative grades (A-E) to each species present in a reference library, according to the level of congruency between morphospecies and their respective sequence clusters or MOTU. Later, Kneibelsberger *et al.* (2014) and Oliveira *et al.* (2016) updated the system to use BINs as the reference clustering method. The main goals of this system were to provide end-users of reference libraries with a system which could sort out possible incongruent records, species which are identifiable with current data, ambiguous records or to uncover possible cases of hidden diversity. However, this ranking system relied on individual attribution of the qualitative grades to each species by a user or operator, a method which would prove impractical if used for large DNA metabarcoding reference libraries comprised of hundreds or thousands of species.

In order to reduce these limitations, an R-based application, BAGS (Barcode, Audit & Grade System), was developed for automated auditing and annotation of DNA barcode reference libraries. Here the ranking system proposed by Oliveira *et al.* (2016) was adapted to essentially root the attribution of the grades on match/mismatch between BINs and morphospecies or species names. BAGS is able to apply the auditing and annotation system to user-provided species lists or large taxon-specific datasets composed of all available COI barcode sequences in BOLD, including those mined from GenBank. BAGS also aims to facilitate revision and curation of barcode reference libraries, thereby contributing to improve their quality.

1.3. Objectives

The main goal of this study was the creation of an automated auditing and annotation system for DNA barcode reference libraries of COI sequences for all members of the kingdom Animalia. In a second stage, the goal was to implement this system in a web application with a user-friendly graphical interface. Lastly, the aim was to thoroughly test the application in order to perceive its usefulness and importance for the fields of molecular ecology and taxonomy. In order to meet these goals, the tasks of this project consisted on the following:

- Development of an R script to perform the data mining, auditing and annotation of DNA barcode reference libraries comprising COI sequences and specimen data.
- Implement the R script into a user-friendly web application to be easily accessed and used by researchers for DNA barcoding and metabarcoding studies.
- Perform tests on a set of three taxonomically diverse datasets to assess the accuracy, usefulness and limitations of the automated system here developed.

2. METHODOLOGY

2.1. Development of the R script and web application

2.1.1. Overview of BAGS

BAGS (Figure 5) is an R-based system which features automated compilation of quality-filtered COI sequence datasets from BOLD, allowing for selection or exclusion of marine taxa through matching with the WoRMS checklists (WoRMS, 2019). It delivers taxon-selected libraries annotated with qualitative grades (A to E) based on BIN/morphospecies congruence and on the amount of available data for each species, which can be downloaded whole or sorted by grade. A user-friendly interface allows for minimal operation for users non-familiar with R (R Development Core Team, 2017), while providing a grasp of the overall quality of the reference library through a graphical output of the proportion of records and species assigned to each of the five grades. However, since BAGS can also be locally run, the more experienced R users have the option to make adjustments to the code. The users may then (frequently if necessary) use the annotated datasets to compile their own personalized and reviewed libraries (e.g. BOLD datasets) and use them for taxonomic assignment of HTS metabarcoding-generated reads.

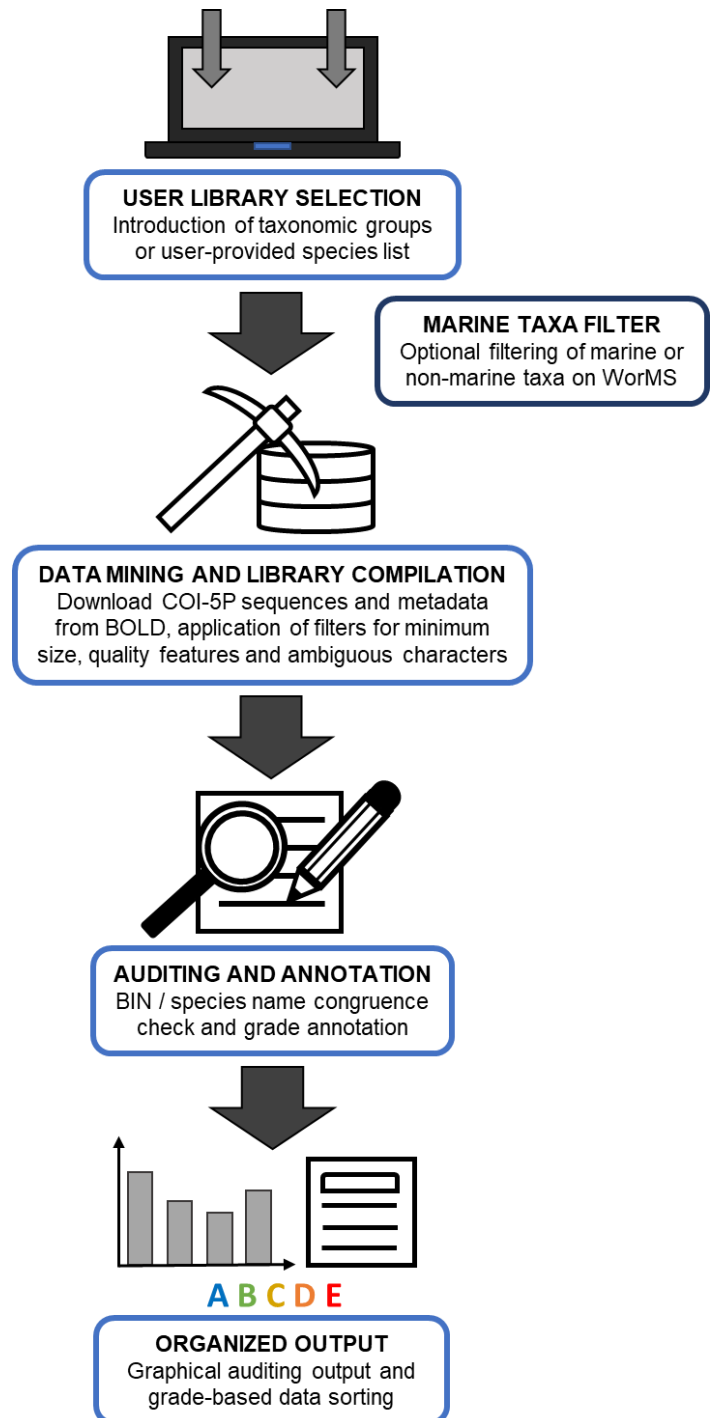


Figure 5 - Overview of BAGS' four main features and their arrangement along the informatics pipeline.

BAGS is composed of four main features which are implemented in sequence (Figure 5): a) data mining and library compilation, b) marine taxa selection/exclusion filter (optional), c) library auditing and annotation and d) auditing output and annotation-based library sorting.

2.1.2. BAGS pipeline

2.1.2.1. Data mining and library compilation

BAGS offers the option for library compilation based on a choice of taxa or through a user-provided species list (view of website's tabs in Annex 1). Records matching the selected taxa or species list will be retrieved and then filtered. All the data is retrieved from BOLD (www.boldsystems.org), using the “bold” R package (Chamberlain, 2019). Therefore, the taxa introduced by the user must be present in BOLD at the time of use. Any taxonomic rank from species to phylum belonging to the kingdom Animalia can be submitted, but it should be noted that some ranks, particularly intermediate ranks, are not implemented in BOLD or may not be available for some species.

The mining of the target taxa can be achieved through three options (view of website's tabs in Annex 2): download all the records available (all taxa), download only records of species occurring in marine habitats (which may include any taxa present in brackish waters) or download the non-marine species' records (i.e. not present in neither marine nor brackish water habitats). This marine species selection or exclusion filter is accomplished resorting to the “worms” R package (Holstein, 2018), which checks the habitat type(s) assigned in WoRMS to each species in a query dataset, among four available (marine, brackish, freshwater or terrestrial).

After the data mining, records are removed if at least one of the following criteria is verified: a) records with sequences shorter than 500 base pairs, or with sequences that have more than 1% ambiguous base calls (Ns); b) records without species name (this includes records identified only by genus or any higher taxonomic rank), or without BIN; c) records without information of the sampling location (either latitude or country of origin). Records with ambiguous expressions present in the species name (e.g. *sp.*, *complex.*, *etc.*; see Annex 3) or in the COI sequence (i.e. not IUPAC nucleotide code; see Annex 4) are not removed, however, the ambiguous expression is removed.

2.1.2.2. Auditing and annotation

Following the initial quality-filtering steps, the BAGS pipeline subsequently proceeds to the implementation of the auditing and annotation system adapted with modifications from Oliveira *et al.*

(2016). The five annotation grades attributed to each species in a compiled library are defined as follows (Figure 6):

Grade A – Consolidated concordance: the morphospecies is assigned to a single BIN, which integrates only members of that species. Additionally, the species is represented by more than 10 specimens in the library.

Grade B – Basal concordance: the morphospecies is assigned to a single BIN, which integrates only members of that species, but there are 10 or less specimens in the library.

Grade C – Multiple BINs: the morphospecies is assigned to more than one BIN, and all those BINs integrate only members of that species.

Grade D – Insufficient data: the species has less than three specimens available in the library and none of the BINs assigned to the species integrate specimens from another species.

Grade E – Discordant species assignment: more than one species is assigned to a single BIN. All the records belonging to that species will be assigned to grade E.

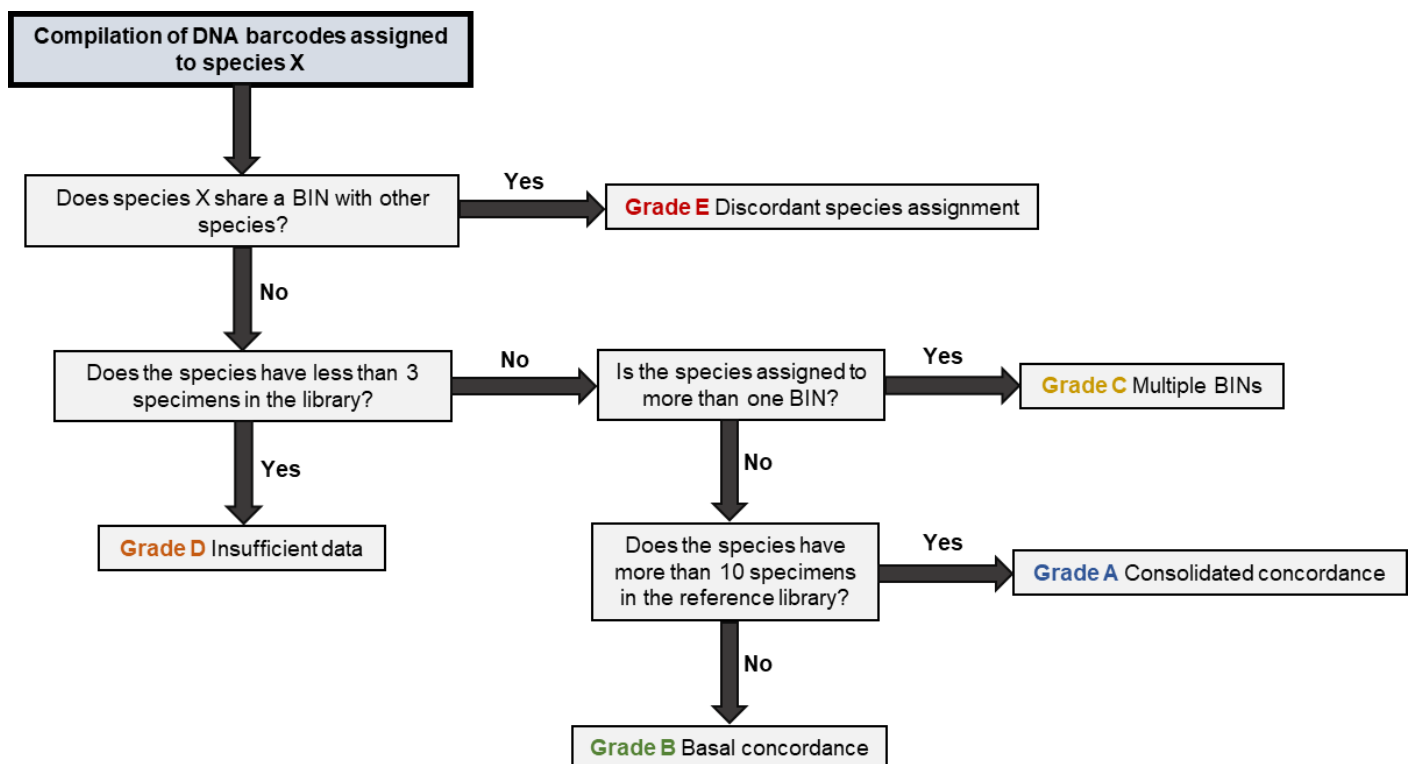


Figure 6 - Workflow for automated auditing and annotation of qualitative grades to each species in a BAGS compiled reference library (adapted from Oliveira et al. 2016).

The BAGS auditing pipeline consists of a series of annotation steps, each comprising data checks with two possible outcomes (Figure 6). Each set of sequences for a given species entering the pipeline

will be annotated with a single grade (A to E). Discordant assignments (grade E) are immediately screened at the front end of the pipeline, followed by records with insufficient data (grade D), then grade C. Grades A or B are attributed last, if the records were not retained in the previous screens. The screening steps involve checking against the full BOLD database, thus not exclusively considering the reference library being downloaded at the time of the annotation, which would limit concordance-checking to the download species' data only.

BOLD (similarly to GenBank and other biological data bases) limits the number of searches or queries per IP/user to avoid the overload of their webservice. Therefore, to avoid blocking the access when querying for large reference libraries, the entire BOLD dataset for animals is downloaded periodically (every two months) in order to calculate the number of BINs for each species, as well as the number of species for each BIN. With this solution, BAGS can work faster and without the computational limitations of real-time query searches on BOLD.

2.1.2.3. Output and annotation-based file sorting

The auditing system proceeds then to the annotation of the records with the pre-defined grades to each species in the reference library, following the pipeline described before. In due course, the reference library will be created and downloaded directly to the web browser transfers folder in the form of a tabular file containing the following: species name, BIN, COI sequence, country or region of origin, the grade that was attributed to the species, number of base pairs in the sequence, family, order, class, sample ID, process ID, latitude, longitude and in the case of marine taxa libraries, an additional column with the valid species name according to WoRMS. The user has also the option to download the reference library in *fasta* format (text-based format specifically for the representation of nucleotide or amino acid sequences), giving the choice of which grades to include. The *fasta* files can be downloaded with all grades, combinations of different grades or separately for each grade (Annex 5).

Lastly, BAGS summarizes the data regarding the reference library that was created (Annex 6), in the form of a text report plus two bar plots: one displaying the number of specimens for each attributed grade and another displaying the number of species for each attributed grade. In order to repeat the process for additional targets, the user must refresh the page and start over again.

2.1.2.4. Informatic implementation

Initially, a script written entirely in the open-source programming language R was created in order to implement the data mining, auditing, annotation, output and file sorting in a semi-automated way. The R script was then converted into a web application, BAGS, designed using the “shiny” R package

framework (Chang, Cheng, Allaire & Yihui, 2019), having therefore an underlying customization with the HTML and CSS marking languages. In addition to the R packages previously mentioned, in order to code the entire application, the following R packages were utilized: “seqRFLP” (Ding & Zhang, 2012), “data.table” (Dowle & Srinivasan, 2019), “stringr” (Wickham, 2019), “readr” (Wickham, Hester & Francois, 2018), “fingerprint” (Guha, 2018), “dplyr” (Wickham, François, Henry & Müller, 2020), “ggplot2” (Wickham, 2016), “shinyWidgets” (Perrier, Meyer & Granjon, 2020) and “snakecase” (Grosser, 2019).

It is possible to launch BAGS locally on any environment that has R installed, as well as through any R Integrated Development Environment (IDE) such as RStudio (Rstudio Team, 2020), where it can fully operate as long as there is a stable internet connection and the databases BOLD and WoRMS are functional. The application can be used without any prior knowledge of the R programming language, and the instructions for launching it can be consulted in the “README” file. BAGS is also hosted at the web server shinyapps.io, which allows its launching from any web browser (web link: <https://tadeu-apps.shinyapps.io/bags>), and at the following web link <https://bags.lsd.di.uminho.pt/>. The script that allows the application to be run locally without constraints in R, as well as a “README” file, are currently stored at GitHub: <https://github.com/tadeu95/BAGS>.

BAGS’ graphical-user interface (GUI) consists of a website organized by tabs and sub-tabs (Figure 7). It includes a “home” tab with three sub-tabs to explain the motivation behind the development of the application, as well as a brief explanation of the workflow and pipeline that is being implemented. The auditing and annotation pipeline is then executed in the following “taxa for auditing”, “download grades libraries” and “auditing report” tabs (see Annexes 1, 2, 5 and 6). The final “contact and resources” tab is meant to provide some useful hyper-links of various institutions and research groups related to BAGS, as well as a way to cite the application.

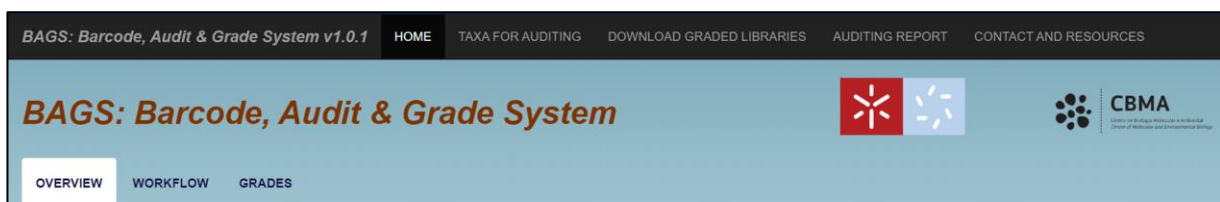


Figure 7 - Print screen of BAGS home page.

2.2. Performance assessment

2.2.1. Marine taxa selection or exclusion filters test

In order to test BAGS performance, two independent tests were performed. First, to understand whether the marine taxa selection or exclusion filters were functional and reliable, three reference libraries

were downloaded using the “all taxa”, the “marine taxa” and the “non-marine” taxa options for a family of shrimps, Palaemonidae, which comprises species from various aquatic habitats. This was followed by checking the report generated by BAGS and manually checking 30 random species from each of the 3 libraries previously generated and verifying if their registered habitats at WoRMS correspond to the habitat selected in BAGS.

2.2.2. Grade assignment test

To understand the effectiveness of BAGS regarding the auditing and grade assignment, three groups of organisms likely to display distinctive compositional features and quality issues in their reference libraries were selected: marine Amphipoda (Malacostraca: Crustacea), Chironomidae (Diptera: Insecta) from all habitats and marine fish (Actinopterygii, Elasmobranchii and Holocephali). Three reference libraries were downloaded using as input “Amphipoda” (within the marine taxa filter option), “Chironomidae” (all taxa option), and “Actinopterygii,Elasmobranchii,Holocephali” also within the marine taxa filter option. Then, the grade assignment was checked by randomly sampling 30 species from each grade, from each compiled library and checking the data manually to assess if the grades were correctly assigned to their specimens. Due to the massive amount of data available for Chironomidae (more than 400,000 sequences accessible on BOLD), the species in the compiled library were matched against a list of European species for the group, obtained through the BOLD workbench (BOLD checklist DNAqua-NET: Diptera, code CL-DNADI), in order to retain only species from Europe. Subsequently, Neighbour-Joining trees were created on the BOLD workbench for the grade C species of each tested taxonomic group, to understand whether the checked species were monophyletic or non-monophyletic. Within grade E, different plausible origins for the discordance were scored for the following categories: synonym; faulty or ambiguous species names; consolidated morphospecies grouped in one BIN, probable misidentification and inconclusive origin.

3. RESULTS

3.1. Performance tests

3.1.1. Marine taxa selection or exclusion filter

Using the input “Palaemonidae” within the marine filter, the marine taxa library comprised 60 species assigned to 73 BINs, and a total of 577 specimens (Table 1), while the non-marine taxa library comprised 51 species, 67 BINs and a total of 318 specimens (Table 1). Comparatively, the “all taxa” option library had 123 species, 148 BINs and 1,022 specimens (Table 1). The 30 species randomly sampled of the marine taxa-filtered library were correctly assigned (i.e. all the 30 species were registered as being from marine or brackish environments when manually checked on WoRMS). Nonetheless, this included species which were registered simultaneously as occurring in both marine and freshwater habitats. On the other hand, the 30 species manually checked from the non-marine taxa library revealed to be all exclusively from freshwater environments (i.e. not present neither in marine nor brackish waters, and therefore not present in the marine library). Lastly, the 30 species manually checked from the “all taxa” library belonged to all habitats where members of the Palaemonidae family can be found (marine habitats, fresh water and brackish habitats).

Table 1 - Number of specimens, species and BINs for each of the three libraries created with BAGS for the family Palaemonidae.

Library	Specimens	Species	BINs
All-taxa	1,022	123	148
Marine	577	60	73
Non-marine	318	51	67

3.1.2. Trial datasets

The marine Amphipoda dataset comprised a total of 6,385 specimens in the compiled library, 486 species and 736 BINs; the Chironomidae dataset consisted of a total of 90,214 specimens, 1,113 species and 1,883 BINs; and the marine fishes dataset comprised 107,434 specimens, 8,381 species and 9,779 BINs (Table 2).

Table 2 - Number of specimens, species and BINs for each of the three libraries used for the grade assignment test (Marine Amphipoda, Chironomidae, Marine fish).

Library	Specimens	Species	BINs
Marine Amphipoda	6,385	486	736
Chironomidae	90,214	1,113	1,883
Marine fish	107,434	8,381	9,779

The distributions of the number of species per grade in each of the compiled reference libraries (Figure 8) show that the proportion of cases of hidden diversity (grade C) is higher in the two invertebrate libraries (Amphipoda and Chironomidae; around 20%) compared with the marine fish library (less than 10%). Cases of insufficient records, which consist of species with less than three specimens in the BAGS-compiled library (Grade D), are also less prevalent in the marine fishes (~18%) when compared to both invertebrate libraries (40% and 26% for Amphipoda and Chironomidae respectively). On the other hand, cases of apparent discordance (Grade E) are considerably less prevalent in the Amphipoda library (only 12% of the cases) and much more frequent in the marine fish library (44%). For the grade E cases (i.e. multiple morphospecies in the same BIN), the number of species per each given BIN varied between 1 and 49 for Amphipoda, 1 and 12 for Chironomidae and 1 and 88 for fish (Figure 9).

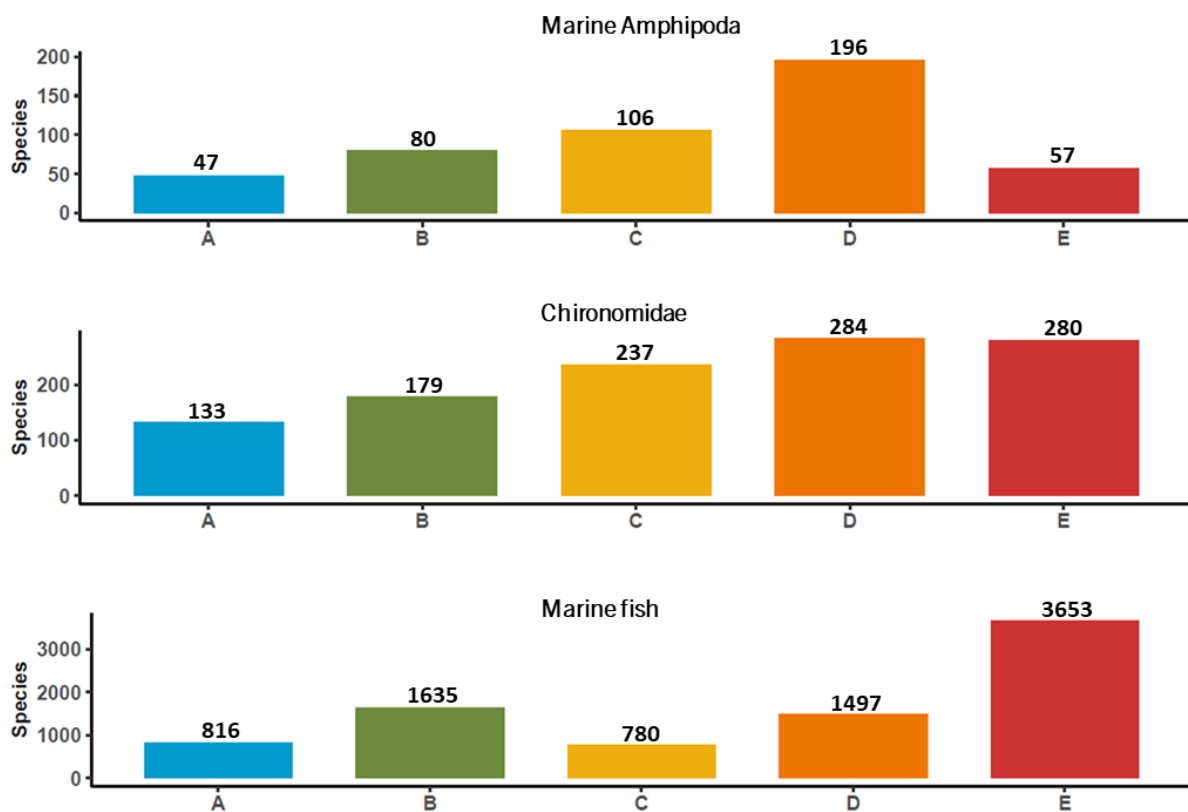


Figure 8 - Barplots displaying the distribution of the number of species assigned to each qualitative grade for the three taxonomic groups tested. From top to bottom: marine Amphipoda, Chironomidae and marine fish (*Actinopterygii*, *Elasmobranchii* and *Holocephali*).

For the three groups, all randomly sampled species were correctly assigned to the qualitative grades, according to their definition and the data present at BOLD at the time of the test. Grade C species (Table 3) were mostly monophyletic, between 66.7% (Chironomidae) and 80% (fish). Discordances or potential errors in grade E annotations had different possible sources (Table 4). Misidentifications (between 37% and 67%) and ambiguous species names (between 10% and 33%) contributed the most to

the grade E cases, followed by consolidated morphospecies aggregated in one BIN (between 0% and 26%), while synonyms contributed the least (overall 3.4%).

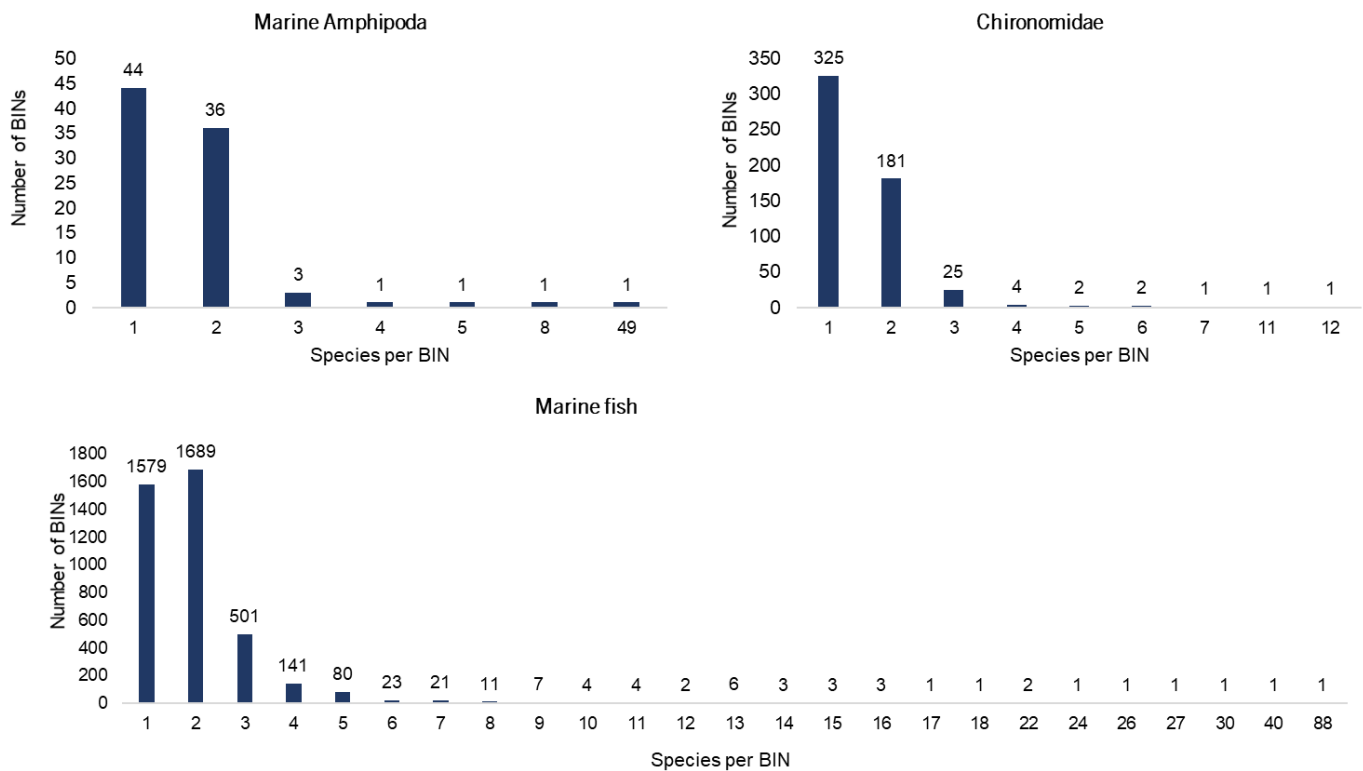


Figure 9 - Number of species per BIN in the grade E dataset generated through BAGS for each tested taxonomic group (marine Amphipoda, Chironomidae and marine fish).

Table 3 - Percentage of monophyletic or non-monophyletic tested species assigned to grade C of each tested taxonomic group, according to their position in the Neighbour-Joining trees constructed.

	Monophyletic	Non-monophyletic
Marine Amphipoda	76.7%	23.3%
Chironomidae	66.7%	33.3%
Marine fish	80.0%	20.0%
Overall	74.4%	25.6%

Table 4 - Percentage of the different plausible origins for the assignment of grade E to species for each tested taxonomic group.

	Synonym	Ambiguous species names	Consolidated morphospecies aggregated in one BIN	Misidentification	Inconclusive
Marine Amphipoda	0.0%	30.0%	0.0%	66.7%	3.3%
Chironomidae	0.0%	33.3%	10.0%	50.0%	6.7%
Marine fish	10.0%	10.0%	26.6%	36.7%	16.7%
Overall	3.4%	24.4%	12.2%	51.1%	8.9%

4. DISCUSSION

While molecular and computational tools have been increasingly providing taxonomists with large volumes of data to analyse, the need for systems which classify and audit that data is now more relevant than ever. This is especially the case when dealing with publicly available DNA barcodes, which can be freely submitted to biological data bases and subsequently used by researchers anywhere, at any time (Curry *et al.*, 2018; Meiklejohn *et al.*, 2019). Moreover, given the establishment of DNA barcoding as one of the primary drivers behind the recent scientific efforts in uncovering and explaining biodiversity (DeSalle & Goldstein, 2019; Pennisi, 2019), BAGS' main goal is to facilitate the implementation of curation and quality control measures among taxonomists and molecular ecologists. Additionally, another important goal is to do this through a user-friendly and automated platform, removing any need for programming skills in order to audit and annotate a reference library.

BAGS differs from the “BIN discordance report” available at BOLD, within the sequence analysis tools. Firstly, whereas the BOLD tool is BIN-centred, our approach is morphospecies-centred. This fundamental difference has a number of consequences. While BOLD reports discordant BINs, BAGS reports on discordant morphospecies, meaning that a morphospecies displaying even a single record in a discordant BIN is classified as grade E. The morphospecies-centred approach also enables BAGS to report on species occurring in multiple – but non-discordant – BINs (grade C), thus serving as a barometer of suspected hidden diversity in reference libraries. Finally, BAGS also takes into consideration the amount of sequences available in the database, providing a grasp of gaps in comprehensiveness of coverage for morphospecies in the reference libraries (grade A, B and D). From an auditing and taxonomic curation viewpoint, the morphospecies-centred approach is also more advantageous.

4.1. BAGS performance assessment tests

The different efficiency tests performed with BAGS (either marine taxa selection/exclusion or grade annotation) allowed to verify the correct performance of this application. The different manual tests and the ongoing tests performed during beta tests, did not bring to light any errors of the application in the filtering or the auditing and annotation steps.

4.1.1. Marine taxa selection or exclusion filter test

BAGS' marine taxa selection or exclusion filtering options proved to be successful at selecting species based on the habitat data recorded in WoRMS. However, it is important to point out certain details of this tool. For instance, some transitional marine species (i.e. present in estuaries) are registered in WoRMS as being from brackish habitats, which can include both typical marine or freshwater species

(e.g. *Phoxinus* or *Palaemonetes*). These species should not be excluded from marine reference libraries as they may also be detected in metabarcoding studies in fully marine environments. Moreover, records retrieved from BOLD containing ambiguous species names will not be included in marine or non-marine libraries, even if they are meant to represent a marine genus (e.g. "*Hippocampus* sp. *FLWL06*" or "*Pseudanthias* sp. *KSA_1880*"). These ambiguous species names will not be recognized by WoRMS, therefore BAGS is not able to allocate these records to their corresponding habitat. If the goal is to gather as much barcode compliant records as possible in the final dataset, regardless of the habitat, it is advisable to use the "all taxa" option. However, the "marine" and "non-marine" options of BAGS are a useful resource if the user wishes to use a customized and size-amenable reference library targeting preferentially only marine or non-marine organisms.

4.1.2. Trial datasets

The choice of taxonomic groups for testing BAGS included two key invertebrate groups in aquatic monitoring (marine Amphipoda and Chironomidae), likely to be relevant in metabarcoding applications, and a well-represented group of vertebrates in BOLD with a large number of species (marine fish). By using three distinct taxonomic groups important in biomonitoring studies, BAGS allowed to promptly understand the differences in the level of congruency of their available DNA barcodes and in the quality of their respective reference libraries.

Recent initiatives (e.g. deWaard *et al.*, 2019; Hobern & Hebert, 2019; Leese *et al.*, 2016) have been striving to increase the taxonomic coverage of universal databases, however, DNA barcodes are still missing for many species (Weigand *et al.*, 2019) or are poorly represented (high prevalence of grade D species here observed; Figure 8), reinforcing the continuous need for the completion of reference libraries. Ultimately, grade D is meant to single out species with few barcode standard abiding records. Nonetheless, if a species contains less than 3 specimens in a reference library, while containing even one single discordant record (i.e. a specimen which shares a BIN with at least one specimen of a different species), it will be classified by BAGS as grade E and not as grade D.

4.1.2.1. Cases of possible hidden diversity (Grade C)

BAGS performance tests allowed to spot a high proportion of possible cases of hidden diversity, reaching around 20% in Chironomidae and Amphipoda, but less prevalent in marine fish (Figure 8). Indeed, a fair amount of cases of cryptic diversity have been reported in the literature for marine amphipods (e.g. Hyalidae, Desiderato *et al.*, 2019; Gammaridae, Hupało *et al.*, 2019), while the family Chironomidae belongs to an order (Diptera) notorious for incorporating large numbers of hidden species

(Ekrem, *et al.*, 2010; Lin *et al.*, 2015). In marine fish, on the other hand, detection of cryptic species has been less reported (Knebelsberger *et al.*, 2014; Oliveira *et al.*, 2016), maybe due to the fact that their taxonomy is possibly more updated, morphological differentiation is more rigorously established for most species, or the fact that their high mobility may reduce the likelihood of genetic divergence between populations over larger distances. Most species assigned grade C were monophyletic, as determined through the examination of the phylogenetic trees, reaching a global percentage of 74.4% (Table 3).

However, the fewer cases of non-monophyletic species, can potentially have originated from some operational error or misidentification. For instance, the Chironomidae species *Eukiefferiella claripennis* and *Synorthocladius semivirens* each had a total of 133 and 38 specimens in the neighbour-joining tree, respectively. Although the majority of both species' sequences were clustered monophyletically in the tree, there were exceptions which indicate the possibility of errors or misidentifications. The non-monophyletic records were two *Synorthocladius semivirens* specimens and one *Eukiefferiella claripennis* specimen, which were placed closer to specimens of the *Cricotopus* genus (Figure 10), but still assigned to their own BINs. One hypothesis to explain these three grade C records is that in reality, the specimens belonged to *Cricotopus* species, and were misidentified as *Synorthocladius semivirens* and *Eukiefferiella claripennis*. In the case of *Eukiefferiella claripennis*, the single misplaced sequence was imported from GenBank to BOLD, underlining the importance of auditing and curating not only the data that is directly submitted to BOLD, but also the data imported from other biological data bases. Alternatively, the disparate records may conceal hidden diversity in need of proper diagnosis and taxonomic placement.

Another case of non-monophyly is the marine fish species *Ostorhinchus doederleini*, which comprised a total of 15 specimens in the grade C tree, 14 of which were monophyletic, with a single exception, as displayed in Figure 11. Since all the neighbouring specimens belong to the *Ostorhinchus* genus, it's conceivable that a single *Ostorhinchus doederleini* was misidentified and most likely belonged to a different *Ostorhinchus* species. Additionally, the only misplaced sequence is assigned to a different BIN (BOLD:AAC5895) from the two BINs (BOLD:AAC5894, BOLD:AAC5893) which were assigned to the remaining 14 monophyletic records.

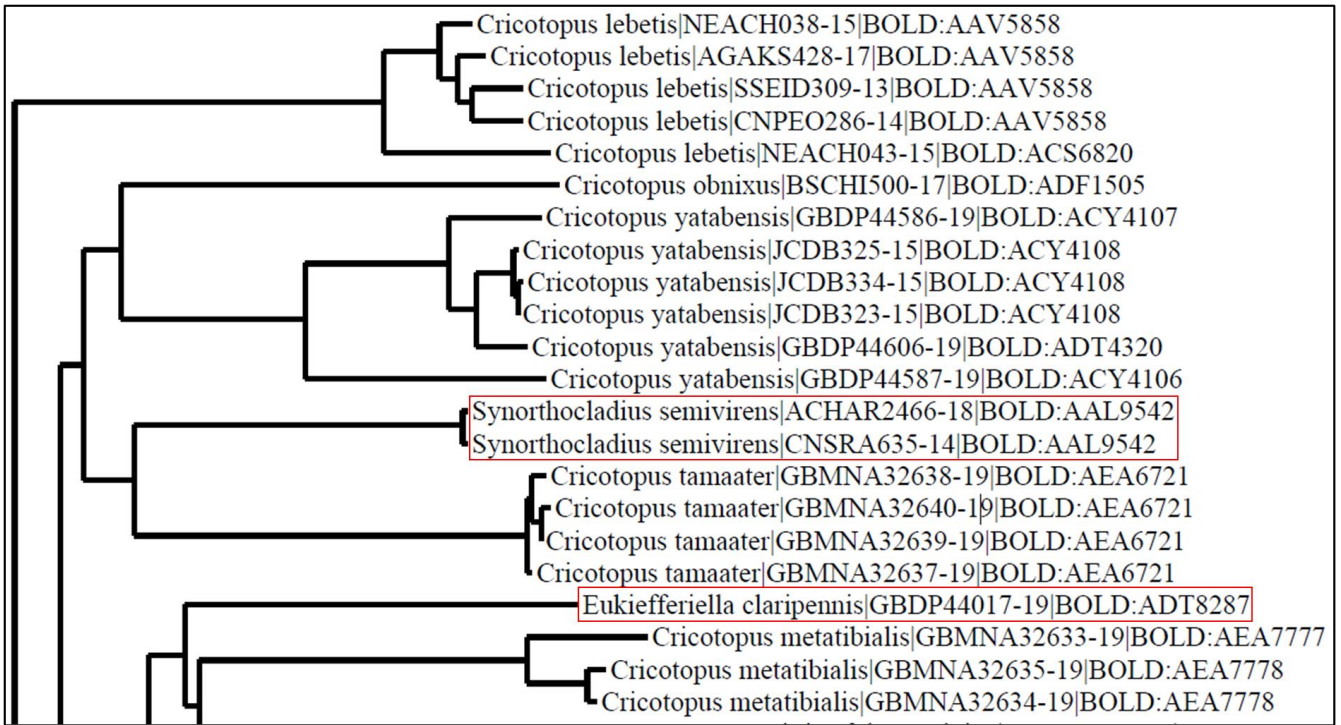


Figure 10 - Subset of the neighbour-joining tree created for the grade C species belonging to the Chironomidae reference library.

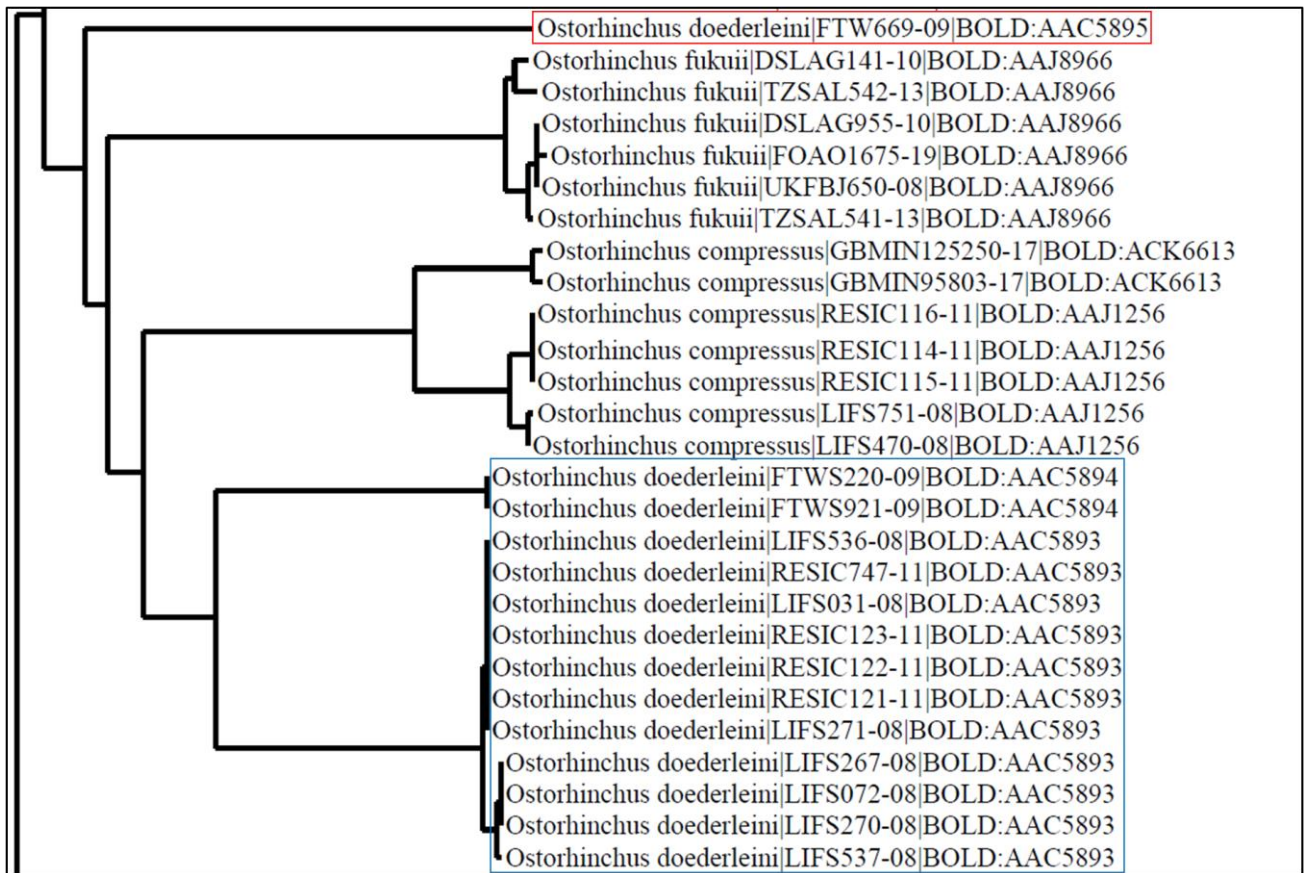


Figure 11 - Subset of the neighbour-joining tree created for the grade C species belonging to the marine fish reference library.

A similar situation to that of *Ostorhinchus doederleini* was observed in the marine Amphipoda *Ampithoe dalli*, which comprised a total of 17 specimens in the grade C tree, being that 16 of them were placed monophyletically and just one was non-monophyletic (Figure 12). Observing the subset of the tree, it can be postulated that perhaps the misplaced sequence was misidentified as *Ampithoe dalli*, but in reality, the specimen belonged to a distinct species of *Ampithoe* without matching specimens in BOLD, or may even constitute an undescribed species, although other possibilities cannot be discarded.

These examples corroborate the importance of the distinction between monophyletic and non-monophyletic grade C species, since they seem to represent two different possible scenarios. Monophyletic species tend to have a higher probability of holding hidden diversity since they are distributed by more than one BIN exclusive to said species. Alternatively, non-monophyletic species are more likely to display some sort of incongruency since they are placed closer to records belonging to different species.

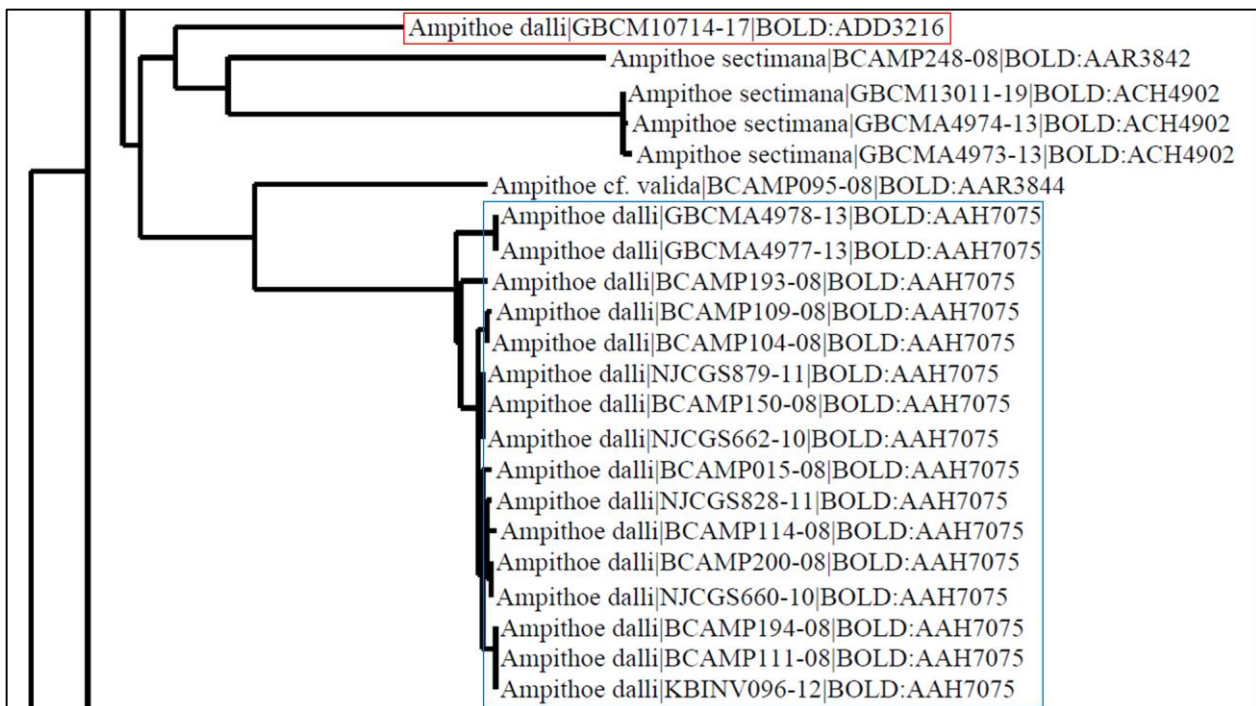


Figure 12 - Subset of the neighbour-joining tree created for the grade C species belonging to the marine Amphipoda reference library.

4.1.2.2. Cases of discordance (Grade E)

A number of studies have been addressing the curation of marine invertebrate's DNA barcodes, including Amphipoda (e.g. Lobo *et al.*, 2017; Raupach *et al.*, 2015), which may explain the lowest proportion of possible discordances (Grade E) out of the three groups analysed (Figure 8). Nonetheless, while manually checking the grade E amphipod species, several probable misidentifications were found, as well as pseudo-discordances in the form of ambiguous species names (Table 4). For instance,

Ampelisca diadema specimens were distributed through five distinct BINs, with two of them (BOLD:ABW2163 and BOLD:ACH8191) including more than one different species. BIN BOLD:ABW2163 included eight *Ampelisca diadema* sequences and one *Ampelisca spinipes* sequence, which was probably misidentified, given the close morphological resemblance between the two species. BIN BOLD:ACH8191 included three *Ampelisca ledoyeri*, two *Ampelisca tenuicornis* and one single *Ampelisca diadema* sequence, meaning that one or even two of these species could also have been potentially misidentified. On the other hand, other grade E species such as *Austrochiltonia subtenuis*, only had one BIN assigned to all of its 49 sequences (BOLD:ACH6392). However, this BIN also included sequences with ambiguous species names (e.g. “*Austrochiltonia aff.subtenuis Amp1*”, “*Austrochiltonia sp. Hap76*”, “*Austrochiltonia aff.subtenuis MOH16Amp1*” and “*Austrochiltonia aff. subtenuis 2amp1*”). This kind of nomenclature is frequently used to refer to different lineages of a single species, although BAGS still classifies it as discordance.

The marine fishes’ reference library showed a prominently high proportion (~44%) of grade E species (Figure 8), mainly due to misidentifications, consolidated morphospecies aggregated in one BIN or faulty species names lexicon (Table 4). There are some extreme cases which greatly contribute to these scenarios, as for instance, BINs BOLD:AAC8034 and BOLD:AAB3926, consisting of 40 and 88 species respectively (Figure 8). In the latter case, out of 88 specimens, only one is spelled correctly (“*Pseudanthias squamipinnis*”), while the remaining were named “Unknown” or “*Pseudanthias sp.*” followed by different alphanumeric designations. Similarly to the previously mentioned grade E species, since these ambiguous species names, possibly interim names, are not properly standardized, BAGS considers them different species for the purpose of comparison against the BOLD database and grade assignment, even though it does remove the ambiguous expressions and specimens assigned only to genus, in the compiled libraries.

Out of the three libraries, the marine fishes’ also had the higher percentage of discordances caused by several consolidated morphospecies being assigned to a single BIN (26.6%). As an example, BIN BOLD:ABZ0850 holds specimens named with four different non-ambiguous species names (*Carcharhinus obscurus*, *Carcharhinus galapagensis*, *Carcharhinus plumbeus* and *Carcharhinus longimanus*), which points to the possibility of low-resolution of the BIN system algorithms in terms of molecularly differentiating these species.

Synonym species names such as *Yongeichthys nebulosus*, were also found to cause the attribution of grade E. *Yongeichthys nebulosus* is not an accepted species name at WoRMS, in detriment

of the accepted name *Acentrogobius nebulosus*. BAGS considers synonyms as discordances because within BOLD, there are BINs which hold specimens named with both versions (e.g. BOLD:AAC3655 and BOLD:ABY4787). Nevertheless, as previously mentioned, the reference libraries created with the marine and non-marine options, contain a column specifically for species name accepted by WoRMS, helping in the signalling of the cases of synonym species names.

Considering these and other possible grade E scenarios, the user of this system should keep in mind that this grade should serve as an incentive for a close examination of that particular species' records, and not as definitive signalling of unreliability. Indeed, the detailed inspection of grade E cases after BAGS annotation revealed that most of them are likely pseudo-discordances and, if eventually clarified, could lead to an estimated overall reduction of 80% in grade E species.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

5.1. Conclusions and BAGS limitations

Generally, the grade assignment implemented in this study proves itself to be very helpful by providing a solid grasp of the taxonomic congruency between the morphological identification and the algorithmic identification (through the BIN system) of a given species, as it has been previously demonstrated (Costa *et al.*, 2012; Knebelsberger *et al.*, 2014; Oliveira *et al.*, 2016). Its usefulness is especially enhanced with BAGs, where the implementation of the grades can be achieved with minimal effort by the user in an automated way. Nonetheless, the user of this application should be mindful that the complexity of taxonomy and the complexity of its assessment through molecular and computational tools, require a critical look at the results of the auditing of a reference library.

Typically, grade A and B usually point in the direction that the barcodes existent for the species are so far congruent, although it sometimes may be the case that all records for a species originate from the same source, which can potentially make this congruency misleading, or at least, temporary. Grade C seems to be especially good at detecting possible cryptic species or hidden diversity, given that it is more common in groups of animals where their morphology is more complex or difficult to discriminate. Grade D is the less problematic since it exists solely to signal species with a limited number of records. Lastly, Grade E can be considered the most problematic grade, since it can be the result of a myriad of different errors and mistakes, or simply the result of the inherent complexity of biological data.

Although this current version of BAGS has its own merits and stands on its own as a complete tool, filling a gap in the current DNA barcoding research landscape that was identified, there are still

limitations which should be addressed in future versions. Currently, BAGS does not have the ability to flag gross sequence mismatches, such as bacterial sequences mistakenly assigned to animals, as it has been previously reported (Siddall *et al.*, 2009). Although these might be rare events, it would be useful to fully discriminate these cases so that the congruency of the reference library is increased, and more errors are subsequently flagged. Additionally, in its current version, BAGS cannot distinguish grade C's monophyletic from non-monophyletic species, nor can it recognize synonyms and apparent discordances, such as faulty or interim species names, in species graded E. Moreover, since BAGS implements grades which are defined on the BIN/morphospecies matches, the limitations associated with the accuracy of the BIN clustering algorithm may emerge in some results or particular groups of organisms. This could be possibly improved in future version with the introduction of customized OUT clustering algorithms that may be useful to complement the BIN-based auditing, opening possibilities for its application beyond COI sequences and the BOLD database.

Many databases (e.g. BOLD, GenBank, WoRMS) have systems that detect excessive calls by the same user (i.e. too many searches or queries) that might overload their webservice and therefore, they either limit the number of calls or block the user's IP address for a period of time. Since BAGS relies on multiple searches on BOLD (especially in the case of species lists uploaded by the user), this restriction would limit its efficiency. To overcome this constraint, as previously mentioned, part of the data necessary to implement the grade annotation system is regularly downloaded from BOLD, and used for comparison each time a user downloads a new reference library. However, since the full species name and BIN dataset is locally stored for this purpose, the grade attribution can potentially change every time new barcode records and BINs are added to BOLD.

5.2. Future perspectives

Several prospective improvements may be considered in future versions of BAGS. One such key improvement would be to introduce the capability to detect cases of deep discordance which may in fact appear concordant (hence pseudo-concordances) such as the cases of bacterial DNA inadvertently amplified from metazoan DNA during PCR, further included in public genetic repositories assigned to metazoan species (Siddall *et al.*, 2009). The introduction of a phylogenetic placement auditing tool would constitute a possible solution to detect such events and it would also be essential to discriminate cases of monophyly and non-monophyly in grade C-assigned species. Additional improvements to BAGS may include implementation of alternative clustering algorithms and customized filtering thresholds, making it prone for future implementations using other DNA-barcode sequence systems and databases. Finally, the

inclusion of a subsidiary tool to perform a detailed revision of grade E records, in order to signal, for example, pseudo-discordances generated by synonyms or ambiguous species designations, possibly using machine learning and artificial intelligence algorithms. Eventually, some discordances may require individual professional judgement that cannot be accomplished with automated procedures.

The ultimate goal is that BAGS can facilitate and stimulate the much-needed revision and curation of reference libraries. It is urged that all users contribute to this critical task for the sake of the quality of the libraries and ultimately the soundness of the research that depends on it.

6. BIBLIOGRAPHY

- Ball-Damerow, J. E., Brenskelle, L., Barve, N., Soltis, P. S., Sierwald, P., Bieler, R., LaFrance, R., Ariño, A. H., & Guralnick, R. P. (2019). Research applications of primary biodiversity databases in the digital age. *PLoS one*, *14*(9), e0215794. doi: 10.1371/journal.pone.0215794
- Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology letters*, *15*(4), 365-377. doi: 10.1111/j.1461-0248.2011.01736.x
- Bengtsson-Palme, J., Richardson, R. T., Meola, M., Wurzbacher, C., Tremblay, É. D., Thorell, K., ... Nilsson, R. H. (2018). MetaxA2 database builder: Enabling taxonomic identification from metagenomic or metabarcoding data using any genetic marker. *Bioinformatics*, *34*(23), 4027-4033. doi: 10.1093/bioinformatics/bty482
- Blair, J., Gwiazdowski, R., Borrelli, A., Hotchkiss, M., Park, C., Perrett, G., & Hanner, R. (2020). Towards a catalogue of biodiversity databases: An ontological case study. *Biodiversity data journal*, *8*, e32765. <https://doi.org/10.3897/BDJ.8.e32765>
- Brown, D. S., Jarman, S. N., & Symondson, W. O. C. (2012). Pyrosequencing of prey DNA in reptile faeces: Analysis of earthworm consumption by slow worms. *Molecular Ecology Resources*, *12*(2), 259-266. doi: 10.1111/j.1755-0998.2011.03098.x
- Cariani, A., Messinetti, S., Ferrari, A., Arculeo, M., Bonello, J. J., Bonnici, L., ... Tinti, F. (2017). Improving the conservation of mediterranean chondrichthyans: The ELASMOMED DNA barcode reference library. *PLoS ONE*, *12*(1). doi: 10.1371/journal.pone.0170244
- Carvalho, D. C., Palhares, R. M., Drummond, M. G., & Frigo, T. B. (2015). DNA Barcoding identification of commercialized seafood in South Brazil: A governmental regulatory forensic program. *Food Control*, *50*, 784-788. doi: 10.1016/j.foodcont.2014.10.025
- Chamberlain, S. (2019). *bold: Interface to Bold Systems API. R package version 0.9.0*. Retrieved from <https://cran.r-project.org/package=bold>
- Costa, F. O., & Carvalho, G. R. (2010). New insights into molecular evolution: prospects from the Barcode of Life Initiative (BOLI). *Theory in Biosciences*, *129*(2-3), 149-157. doi: 10.1007/s12064-010-0091-y
- Costa, F. O., Landi, M., Martins, R., Costa, M. H., Costa, M. E., Carneiro, M., ... Carvalho, G. R. (2012). A ranking system for reference libraries of DNA barcodes: application to marine fish species from Portugal. *PLoS One*, *7*(4), 1-9. doi: 10.1371/journal.pone.0035858
- Costa, F. O., & Antunes, P. M. (2012). The contribution of the barcode of life initiative to the discovery and monitoring of biodiversity. In *Natural Resources, Sustainability and Humanity*. Springer, Dordrecht, 37-68. doi: 10.1007/978-94-007-1321-5_4
- Costello, M. J., Wilson, S., & Houlding, B. (2012). Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology*, *61*(5), 871. doi: 10.1093/sysbio/syr080
- Curry, C. J., Gibson, J. F., Shokralla, S., Hajibabaei, M., & Baird, D. J. (2018). Identifying north American freshwater invertebrates using DNA barcodes: Are existing COI sequence libraries fit for purpose? *Freshwater Science*, *37*(1), 178-189. doi: 10.1086/696613

- DeSalle, R., & Goldstein, P. (2019). Review and Interpretation of Trends in DNA Barcoding. *Frontiers in Ecology and Evolution*, 7, 302. doi: 10.3389/fevo.2019.00302
- Desiderato, A., Costa, F. O., Serejo, C. S., Abbiati, M., Queiroga, H., & Vieira, P. E. (2019). Macaronesian islands as promoters of diversification in amphipods: The remarkable case of the family Hyalidae (Crustacea, Amphipoda). *Zoologica Scripta*, 48(3), 359-375. doi: 10.1111/zsc.12339
- deWaard, J. R., Ratnasingham, S., Zakharov, E. V., Borisenko, A. V., Steinke, D., Telfer, A. C., ... Hebert, P. D. N. (2019). A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Scientific Data*, 6, 308. doi: 10.1038/s41597-019-0320-2
- Ding, Q., & Zhang, J. (2012). *seqRFLP: Simulation and visualization of restriction enzyme cutting pattern from DNA sequences*. Retrieved from <https://cran.r-project.org/package=seqRFLP>
- Djurhuus, A., Pitz, K., Sawaya, N. A., Rojas-Márquez, J., Michaud, B., Montes, E., ... Breitbart, M. (2018). Evaluation of marine zooplankton community structure through environmental DNA metabarcoding. *Limnology and Oceanography: Methods*, 16(4), 209–221. doi: 10.1002/lom3.10237
- Dowle, M., & Srinivasan, A. (2019). *data.table: Extension of 'data.frame'*. Retrieved from <https://cran.r-project.org/package=data.table>
- Duarte S, Vieira PE, Lavrador AS, & Costa F. O. (2020). Status and prospects of marine NIS detection and monitoring through (e)DNA metabarcoding. *BioRxiv*, <https://doi.org/10.1101/2020.05.25.114280>.
- Durand, J. D., & Borsa, P. (2015). Mitochondrial phylogeny of grey mullets (Acanthopterygii: Mugilidae) suggests high proportion of cryptic species. *Comptes Rendus - Biologies*, 338(4), 266–277. doi: 10.1016/j.crv.2015.01.007
- Ekrem, T., Stur, E., & Hebert, P. D. N. (2010). Females do count: Documenting Chironomidae (Diptera) species diversity using DNA barcoding. *Organisms Diversity and Evolution*, 10(5), 397–408. doi: 10.1007/s13127-010-0034-y
- Fennessy, J., Bidon, T., Reuss, F., Kumar, V., Elkan, P., Nilsson, M. A., ... Janke, A. (2016). Multi-locus Analyses Reveal Four Giraffe Species Instead of One. *Current Biology*, 26(18), 2543–2549. doi: 10.1016/j.cub.2016.07.036
- GBIF Secretariat: GBIF Backbone Taxonomy. Accessed on the 16th of July, 2020. doi: 10.15468/39omei
- Guha, R. (2018). *fingerprint: Functions to Operate on Binary Fingerprint Data*. Retrieved from <https://cran.r-project.org/package=fingerprint>
- Grealy, A. C., McDowell, M. C., Scofield, P., Murray, D. C., Fusco, D. A., Haile, J., ... Bunce, M. (2015). A critical evaluation of how ancient DNA bulk bone metabarcoding complements traditional morphological analysis of fossil assemblages. *Quaternary Science Reviews*, 128, 37-47. doi: 10.1016/j.quascirev.2015.09.014
- Grosser, M. (2019). *snakecase: Convert Strings into any Case*. Retrieved from <https://cran.r-project.org/package=snakecase>
- Hanner, B. R. (2005). Proposed Standards for BARCODE Records in INSDC (BRIs). Technical report, Database Working Groups, Consortium for the Barcode of Life, 2009.
- Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*,

- 270(1512), 313–321. doi: 10.1098/rspb.2002.2218
- Hebert, P. D. N., Hollingsworth, P. M., & Hajibabaei, M. (2016). From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 371(1702). doi: 10.1098/rstb.2015.0321
- Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). Data Descriptor: A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data*, 5, 180156. doi: 10.1038/sdata.2018.156
- Hiiesalu, I., Öpik, M., Metsis, M., Lilje, L., Davison, J., Vasar, M., ... Pärtel, M. (2012). Plant species richness belowground: Higher richness and new patterns revealed by next-generation sequencing. *Molecular Ecology*, 21(8), 2004–2016. doi: 10.1111/j.1365-294X.2011.05390.x
- Hobern, D., & Hebert, P. (2019). BIOSCAN - Revealing Eukaryote Diversity, Dynamics, and Interactions. *Biodiversity Information Science and Standards*, 3, e37333. doi: 10.3897/biss.3.37333
- Holstein, J. (2018). *worms: Retriving Aphia Information from World Register of Marine Species. R package version 0.2.2*. Retrieved from <https://cran.r-project.org/package=worms>
- Hortal J, Lobo JM, Jiménez-Valverde A. (2007). Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conserv Biol*, 21(3), 853-863. doi:10.1111/j.1523-1739.2007.00686.x
- Hupało, K., Teixeira, M. A. L., Rewicz, T., Sezgin, M., Iannilli, V., Karaman, G. S., ... Costa, F. O. (2019). Persistence of phylogeographic footprints helps to understand cryptic diversity detected in two marine amphipods widespread in the Mediterranean basin. *Molecular Phylogenetics and Evolution*, 132, 53-66. doi: 10.1016/j.ympev.2018.11.013
- Janzen, D. H., Burns, J. M., Cong, Q., Hallwachs, W., Dapkey, T., Manjunath, R., ... Grishin, N. V. (2017). Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proceedings of the National Academy of Sciences*, 114(31), 8313–8318. doi: 10.1073/pnas.1621504114
- Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., & Ankenbrand, M. J. (2019). BCdatabaser: on-the-fly reference database creation for (meta-)barcoding. *Bioinformatics*, 36(8), 2630-2631. <https://doi.org/10.32942/osf.io/cmfu2>
- Kneibelsberger, T., Landi, M., Neumann, H., Kloppmann, M., Sell, A. F., Campbell, P. D., ... Costa, F. O. (2014). A reliable DNA barcode reference library for the identification of the North European shelf fish fauna. *Molecular Ecology Resources*, 14(5), 1060–1071. doi: 10.1111/1755-0998.12238
- Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution*, 30(1), 25–35. doi: 10.1016/j.tree.2014.10.008
- Larsen, B. B., Miller, E. C., Rhodes, M. K., & Wiens, J. J. (2017). Inordinate fondness multiplied and redistributed: the number of species on earth and the new pie of life. *The Quarterly Review of Biology*, 92(3), 229-265. doi: 10.1086/693564
- Lamkin, M., & Miller, A. I. (2016). On the challenge of comparing contemporary and deep-time biological-extinction rates. *BioScience*, 66(9), 785-789. doi: 10.1093/biosci/biw088
- Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., ... Zimmermann, J. (2016).

- DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes*, 2, e11321. doi: 10.3897/rio.2.e11321
- Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., ... Weigand, A. M. (2018). Why We Need Sustainable Networks Bridging Countries, Disciplines, Cultures and Generations for Aquatic Biomonitoring 2.0: A Perspective Derived From the DNAqua-Net COST Action. *Advances in Ecological Research*, 58, 63–99. doi: 10.1016/bs.aecr.2018.01.001
- Leite, B. R., Vieira, P. E., Teixeira, M. A. L., Lobo-Arteaga, J., Hollatz, C., Borges, L. M. S., ... & Costa, F. O. (2020). Gap-analysis and annotated reference library for supporting macroinvertebrate metabarcoding in Atlantic Iberia. *Regional Studies in Marine Science*, 101307. doi: 10.1016/j.rsma.2020.101307
- Lin, X., Stur, E., & Ekrem, T. (2015). Exploring genetic divergence in a species-rich insect genus using 2790 DNA barcodes. *PLoS ONE*, 10(9), e0138993. doi: 10.1371/journal.pone.0138993
- Liu, Z. F., Ci, X. Q., Li, L., Li, H. W., Conran, J. G., & Li, J. (2017). DNA barcoding evaluation and implications for phylogenetic relationships in Lauraceae from China. *PLoS ONE*, 12(4), 1–20. doi: 10.1371/journal.pone.0175788
- Lobo, J., Ferreira, M. S., Antunes, I. C., Teixeira, M. A. L., Borges, L. M. S., Sousa, R., ... Hogg, I. (2017). Contrasting morphological and DNA barcode-suggested species boundaries among shallow-water amphipod fauna from the southern European Atlantic coast. *Genome*, 60(2), 147-157. doi: 10.1139/gen-2016-0009
- Maiti, S. K., & Chowdhury, A. (2013). Effects of anthropogenic pollution on mangrove biodiversity: a review. *Journal of Environmental Protection*, 2013. doi: 10.4236/jep.2013.412163
- Mardulyn, P., & Whitfield, J. B. (1999). Phylogenetic Signal in the COI, 16S, and 28S Genes for Inferring Relationships among Genera of Microgastrinae (Hymenoptera; Braconidae): Evidence of a High Diversification Rate in This Group of Parasitoids. *Molecular Phylogenetics and Evolution*, 12(3), 282–294.
- Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. *PLoS ONE*, 14(6), 1–14. doi: 10.1371/journal.pone.0217084
- Mioduchowska, M., Czyz, M. J., Gołdyn, B., Kur, J., & Sell, J. (2018). Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too “universal”? *PLoS ONE*, 13(6), 1–16. doi: 10.1371/journal.pone.0199609
- Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How Many Species Are There on Earth and in the Ocean? *PLoS Biol*, 9(8): e1001127. doi:10.1371/journal.pbio.1001127
- Moudrý, V., & Devillers, R. (2020). Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecological Informatics*, 56, 101051. doi: 10.1016/j.ecoinf.2020.101051
- Nugent, C. M., Elliott, T. A., Ratnasingham, S., & Adamowicz, S. J. (2020). Coil: An R package for cytochrome c oxidase I (COI) DNA barcode data cleaning, translation, and error evaluation. *Genome*, 63(6), 291-305. doi: 10.1139/gen-2019-0206
- Oliveira, L. M., Knebelsberger, T., Landi, M., Soares, P., Raupach, M. J., & Costa, F. O. (2016). Assembling and auditing a comprehensive DNA barcode reference library for European marine

- fishes. *Journal of Fish Biology*, 89(6), 2741–2754. doi: 10.1111/jfb.13169
- Packer, L., Gibbs, J., Sheffield, C., & Hanner, R. (2009). DNA barcoding and the mediocrity of morphology. *Molecular Ecology Resources*, 9(Suppl s1), 42–50. doi: 10.1111/j.1755-0998.2009.02631.x
- Pegg, G. G., Sinclair, B., Briskey, L., & Aspden, W. J. (2008). MtDNA barcode identification of fish larvae in the southern Great Barrier Reef – Australia. *Scientia Marina*, 70(S2), 7–12. doi: 10.3989/scimar.2006.70s27
- Pennisi, E. (2019). DNA barcodes jump-start search for new species. *Science*, 364(6444), 920–921. doi: 10.1126/science.364.6444.920
- Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2019). Forensics and DNA Barcodes – Do Identification Errors Arise in the Lab or in the Sequence Libraries? *BioRxiv*, 738138. doi: 10.1101/738138
- Perrier, V., Meyer, F., & Granjon, D. (2020). *shinyWidgets: Custom Inputs Widgets for Shiny*. Retrieved from <https://cran.r-project.org/package=shinyWidgets>
- Porazinska, D. L., Giblin-Davis, R. M., Esquivel, A., Powers, T. O., Sung, W., & Thomas, W. K. (2010). Ecometagenetics confirm high tropical rainforest nematode diversity. *Molecular Ecology*, 19(24), 5521–5530. doi: 10.1111/j.1365-294X.2010.04891.x
- Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS ONE*, 13(9), 1–16. doi: 10.1371/journal.pone.0200177
- R Development Core Team. (2017). R: A language and environment for statistical computing. *Vienna, Austria*. doi: R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rasmussen, R. S., & Morrissey, M. T. (2008). DNA-based methods for the identification of commercial fish and seafood species. *Comprehensive Reviews in Food Science and Food Safety*, 7(3), 280–295. doi: 10.1111/j.1541-4337.2008.00046.x
- Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular Ecology Notes*, 7(3), 355–364. doi: 10.1111/j.1471-8286.2006.01678.x
- Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE*, 8(7). doi: 10.1371/journal.pone.0066213
- Raupach, M. J., Barco, A., Steinke, D., Beermann, J., Laakmann, S., Mohrbeck, I., ... Kneibelsberger, T. (2015). The application of DNA barcodes for the identification of marine crustaceans from the North Sea and adjacent regions. *PLoS ONE*, 10(9), e0139421. doi: 10.1371/journal.pone.0139421
- Rayé, G., Miquel, C., Coissac, E., Redjadj, C., Loison, A., & Taberlet, P. (2011). New insights on diet variability revealed by DNA barcoding and high-throughput pyrosequencing: Chamois diet in autumn as a case study. *Ecological Research*, 26(2), 265–276. doi: 10.1007/s11284-010-0780-5
- Rimet, F., Abarca, N., Bouchez, A., Kusber, W. H., Jahn, R., Kahlert, M., ... Zimmermann, J. (2018). The potential of High-Throughput Sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. *Fottea*, 18(1), 37–54. doi: 10.5507/fot.2017.013

- Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., ... Bouchez, A. (2016). R-Syst::diatom: An open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, 2016, 1-21. doi: 10.1093/database/baw016
- Roskov Y., Ower G., Orrell T., Nicolson D., Bailly N., Kirk P.M., Bourgoin T., DeWalt R.E., Decock W., Nieukerken E. van, Zarucchi J., Penev L., eds. (2019). Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist. Digital resource at www.catalogueoflife.org/annual-checklist/2019. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-884X.
- Rstudio Team. (2020). Rstudio: Integrated Development for R. *RStudio, PBC, Boston, MA*. Retrieved from <http://www.rstudio.com/>
- Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., ... Ahrens, D. (2017). Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes. *Methods in Ecology and Evolution*, 8(12), 1878–1887. doi: 10.1111/2041-210X.12824
- Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17, e00547. doi: 10.1016/j.gecco.2019.e00547
- Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., & Reyes, A. (1999). Evolutionary genomics in Metazoa: The mitochondrial DNA as a model system. *Gene*, 238(1), 195–209. doi: 10.1016/S0378-1119(99)00270-X
- Saitoh, T., Sugita, N., Someya, S., Iwami, Y., Kobayashi, S., Kamigaichi, H., ... Nishiumi, I. (2015). DNA barcoding reveals 24 distinct lineages as cryptic bird species candidates in and around the Japanese Archipelago. *Molecular Ecology Resources*, 15(1), 177–186. doi: 10.1111/1755-0998.12282
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–5467. doi: 10.1073/pnas.74.12.5463
- Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research*, 47(D1), D94-D99. doi: 10.1093/nar/gky989
- Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., 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(5), 892-901. doi: 10.1111/1755-0998.12236
- Siddall, M. E., Fontanella, F. M., Watson, S. C., Kvist, S., & Erséus, C. (2009). Barcoding bamboozled by bacteria: Convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, 58(4), 445–451. doi: 10.1093/sysbio/syp033
- Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., ... Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere.” *Proceedings of the National Academy of Sciences of the United States of America*, 103(32), 12115-12120. doi: 10.1073/pnas.0605127103
- Stampar, S. N., Maronna, M. M., Kitahara, M. V, Angelis, S. A. De, Lopes, C. S. S., & Reimer, J. D. (2017). DNA barcodes unlocking the phenotypic plasticity in adult and larvae: a case study in Ceriantharia (Cnidaria, Anthozoa). *Genome* 60(11), 998.

- Subbotin, S. A., Toumi, F., Elekçioğlu, I. H., Waeyenberge, L., & Tanha Maafi, Z. (2018). DNA barcoding, phylogeny and phylogeography of the cyst nematode species of the Avenae group from the genus *Heterodera* (Tylenchida: Heteroderidae). In *Nematology*, 20(7), 671-702. doi: 10.1163/15685411-00003170
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045-2050. doi: 10.1111/j.1365-294X.2012.05470.x
- Teixeira M. A. L., Vieira P. E., Pleijel F., Sampieri B., Ravara A., Costa F. O., Nygren A (2020). Molecular and morphometrics combination enriches a large *Eumida* (Annelida, Polychaeta) species complex in European waters. *Zoologica Scripta* 49, 222–235. doi: 10.1111/zsc.12397.
- Trivedi, S., Aloufi, A. A., Ansari, A. A., & Ghosh, S. K. (2016). Role of DNA barcoding in marine biodiversity assessment and conservation: An update. *Saudi Journal of Biological Sciences*, 23(2), 161–171. doi: 10.1016/j.sjbs.2015.01.001
- Trivedi, S., Aloufi, A. A., Rehman, H., Saggu, S., & Ghosh, S. K. (2016). DNA barcoding: tool for assessing species identification in Reptilia. *J Entomol Zool Stud*, 4(1), 332-337.
- Vences, M., Thomas, M., Bonett, R. M., & Vieites, D. R. (2005). Deciphering amphibian diversity through DNA barcoding: Chances and challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462), 1859–1868. doi: 10.1098/rstb.2005.1717
- Vieira, P. E., Desiderato, A., Holdich, D. M., Soares, P., Creer, S., Carvalho, G. R., ... Queiroga, H. (2019). Deep segregation in the open ocean: Macaronesia as an evolutionary hotspot for low dispersal marine invertebrates. *Molecular Ecology*, 28(7), 1784–1800. doi: 10.1111/mec.15052
- Weber, A. A. T., Stöhr, S., & Chenuil, A. (2019). Species delimitation in the presence of strong incomplete lineage sorting and hybridization: Lessons from *Ophioderma* (Ophiuroidea: Echinodermata). *Molecular Phylogenetics and Evolution*, 131, 138–148. doi: 10.1016/j.ympev.2018.11.014
- Weigand, A. M., Jochum, A., Pfenninger, M., Steinke, D., & Klussmann-Kolb, A. (2011). A new approach to an old conundrum-DNA barcoding sheds new light on phenotypic plasticity and morphological stasis in microsnails (Gastropoda, Pulmonata, Carychiidae). *Molecular Ecology Resources*, 11(2), 255–265. doi: 10.1111/j.1755-0998.2010.02937.x
- Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., ... Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment*, 678, 499–524. doi: 10.1016/j.scitotenv.2019.04.247
- Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. Retrieved from <https://cran.r-project.org/package=dplyr>
- Wickham, H., Hester, J., & François, R. (2018). *readr: Read Rectangular Text Data*. Retrieved from <https://cran.r-project.org/package=readr>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Retrieved from <https://ggplot2.tidyverse.org>
- Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. Retrieved from <https://cran.r-project.org/package=stringr>

- Chang, W., Cheng, J., Allaire, J.J., Yihui, J. M. (2019). *shiny: Web Application Framework for R. R package version 1.4.0*. Retrieved from <https://cran.r-project.org/package=shiny>
- Wilson, E. O. (2017). Biodiversity research requires more boots on the ground. *Nature Ecology & Evolution*, *1*(11), 1590-1591. doi: 10.1038/s41559-017-0360-y
- Wong, L. L., Peatman, E., Lu, J., Kucuktas, H., He, S., Zhou, C., ... Liu, Z. (2011). DNA barcoding of catfish: Species authentication and phylogenetic assessment. *PLoS ONE*, *6*(3), 1–7. doi: 10.1371/journal.pone.0017812
- WoRMS. (2019). World Register of Marine Species. Available from <Http://Www.Marinespecies.Org>. doi: 10.14284/170

7. ANNEXES

TAXA SELECTION **UPLOAD SPECIES LIST**

A

Any biome Marine Taxa Only Excluding Marine Taxa

Download, audit and annotate library for all species

Enter the name of the taxonomic group or groups separated by commas, without spaces:

Q Example: Carnivora,Ursidae,Artiodactyla,Soricomorpha

Download

NOTE: Since the download process includes the auditing and annotation of the library, the report is ready once the download is concluded. Make sure to refresh the page every time you are about to download a new library.

TAXA SELECTION **UPLOAD SPECIES LIST**

B

Download, audit and annotate library for species list

Upload a txt or tsv file comprising a list of species:

Browse for a file No file selected

NOTE: Untick the header checkbox if your file does not have a header for the species column.

Header

Display

Head

All

Download

NOTE: Since the download process includes the auditing and annotation of the library, the report is ready once the download is concluded. Make sure to refresh the page every time you are about to download a new library.

Annex 1 - Print screen of the BAGS application displaying the two main options for reference libraries generation. A - Download of a reference library giving as input a list of taxonomic groups separated by commas without spaces; B - Download a reference library giving as input list of species.

TAXA SELECTION

UPLOAD SPECIES LIST

A

Any biome

Marine Taxa Only

Excluding Marine Taxa

Download, audit and annotate library for all species

Enter the name of the taxonomic group or groups separated by commas, without spaces:

Download

NOTE: Since the download process includes the auditing and annotation of the library, the report is ready once the download is concluded.
Make sure to refresh the page every time you are about to download a new library.

TAXA SELECTION

UPLOAD SPECIES LIST

B

Any biome

Marine Taxa Only

Excluding Marine Taxa

Download, audit and annotate library for marine species

Enter the name of the taxonomic group or groups separated by commas, without spaces:

Download

NOTE: This option selects only species that are considered as marine and/or brackish at WoRMS.

NOTE: Since the download process includes the auditing and annotation of the library, the report is ready once the download is concluded.
Make sure to refresh the page every time you are about to download a new library.

TAXA SELECTION

UPLOAD SPECIES LIST

C

Any biome

Marine Taxa Only

Excluding Marine Taxa

Download, audit and annotate library for non-marine species

Enter the name of the taxonomic group or groups separated by commas, without spaces:

Download

NOTE: This option excludes all species which are assigned exclusively to marine and/or brackish at WoRMS.

NOTE: Since the download process includes the auditing and annotation of the library, the report is ready once the download is concluded.
Make sure to refresh the page every time you are about to download a new library.

Annex 2 - Print screen of the BAGS application displaying the three habitat filtering options for reference libraries generation. A – Download a reference library including all species belonging to the chosen taxonomic group; B – Download a reference library including only species occurring in marine or brackish environments; C – Download a reference library excluding species occurring in marine or brackish environments. The habitat of the species is defined according to the information present at WoRMS.

Annex 3 - Ambiguous expressions removed from the species names during the BAGS auditing pipeline.

Removed expressions
sp.
sp. nov
complex.
f.
nr.
s.l.
grp.
type
group
cmplx.
Any digit from 0-9

Annex 4 - List of IUPAC code bases which are not removed by the BAGS auditing pipeline. Every other alphanumerical character is removed from the COI sequences along the BAGS pipeline.

IUPAC Code	Base
A	Adenine
C	Cytosine
T	Thymine
G	Guanine
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	Any base
-	gap

Download graded libraries in fasta format

Choose which grades to include in your library:

NOTE: Make sure the data set download is already completed

Graded library including only species with **grade A**

 Download A library

Graded library including only species with **grade B**

 Download B library

Graded library including only species with **grade C**

 Download C library

Graded library including only species with **grade D**

 Download D library

Graded library including only species with **grade E**

 Download E library

Download graded libraries in fasta format

Choose which grades to include in your library:

NOTE: Make sure the data set download is already completed


Graded library including only species with **grades A and B**

 Download AB library


Graded library including only species with **grades A, B and C**

 Download ABC library

Graded library including only species with **grades A, B, C and D**

 Download ABCD library

Graded library including species with **all grades assigned**

 Download ABCDE library

Annex 5 - Print screen of the BAGS application displaying the options available to download fasta files comprising the COI sequences present in a previously downloaded reference library. A – Download fasta files for individual grades; B – Download fasta files for grouped grades.

Library auditing report

A

REPORT

Taxa name:

Marine taxa name: Octopoda,Cetacea,Hydrozoa,Hippocampus,Valvatida

Non-marine taxa name:

Number of species: 417

Number of BINs: 517

Total Number of specimens in reference library: 5016

Specimens

Number of specimens with A grade: 1257 | Percentage: 25.1 %

Number of specimens with B grade: 329 | Percentage: 6.56 %

Number of specimens with C grade: 800 | Percentage: 15.9 %

Number of specimens with D grade: 222 | Percentage: 4.43 %

Number of specimens with E grade: 2576 | Percentage: 51.4 %

Species

Number of species with A grade: 27 | Percentage: 6.47 %

Number of species with B grade: 59 | Percentage: 14.1 %

Number of species with C grade: 37 | Percentage: 8.87 %

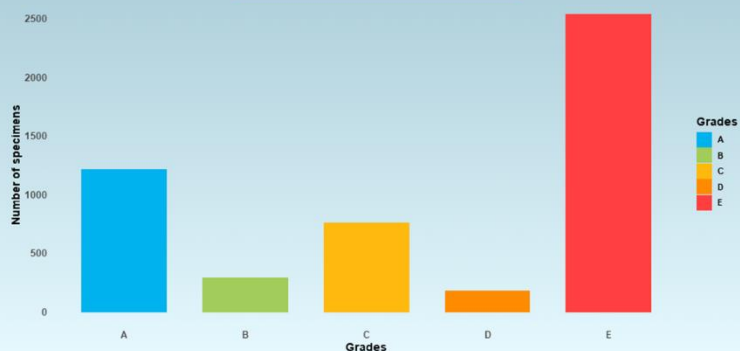
Number of species with D grade: 138 | Percentage: 33.1 %

Number of species with E grade: 149 | Percentage: 35.7 %

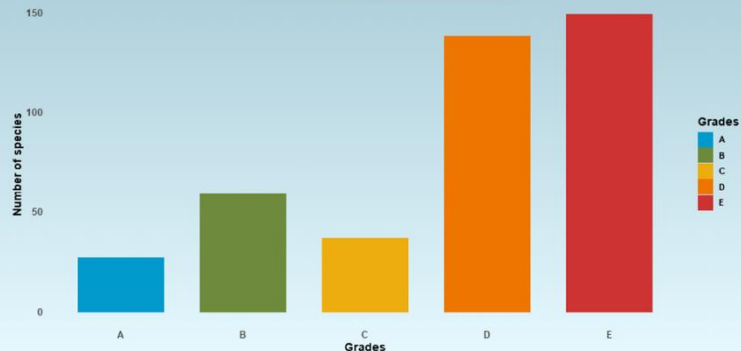
Barplots display

B

NUMBER OF SPECIMENS PER GRADE



NUMBER OF SPECIES PER GRADE



Annex 6 - Print screen of the BAGS application displaying options to summarize the auditing and annotation of a reference library, exemplified by a reference library for the taxonomic groups: Octopoda, Cetacea, Hydrozoa, Hippocampus, Valvatida. A – Text report; B – Graphical report in the form of two barplots (left: specimens per grades, right: species per grade).