Received: 20 May 2020 | Revised: 1 September 2020 | Accepted: 7 September 2020

DOI: 10.1111/1755-0998.13262

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES

WILEY

BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries

João T. Fontes[1,2] | Pedro E. Vieira[1,2] | Torbjørn Ekrem[3] | Pedro Soares[1,2] | Filipe O. Costa[1,2]

[1]Department of Biology, CBMA – Centre of Molecular and Environmental Biology, University of Minho, Braga, Portugal

[2]Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Minho, Portugal

[3]Department of Natural History, NTNU University Museum, Trondheim, Norway

Correspondence
Filipe O. Costa, Department of Biology, University of Minho, CBMA – Centre of Molecular and Environmental Biology, Campus de Gualtar, 4710-057 Braga, Portugal.
Email: fcosta@bio.uminho.pt

Funding information
European Regional Development Fund, Grant/Award Number: NORTE-01-0145-FEDER-000032; Portuguese Foundation for Science and Technology, Grant/Award Number: PTDC/BIA-BMA/29754/2017

Abstract

Biodiversity studies greatly benefit from molecular tools, such as DNA metabarcoding, which provides an effective identification tool in biomonitoring and conservation programmes. The accuracy of species-level assignment, and consequent taxonomic coverage, relies on comprehensive DNA barcode reference libraries. The role of these libraries is to support species identification, but accidental errors in the generation of the barcodes may compromise their accuracy. Here, we present an R-based application, Barcode, Audit & Grade System (BAGS) (https://github.com/tadeu95/BAGS), that performs automated auditing and annotation of cytochrome c oxidase subunit I (COI) sequences libraries, for a given taxonomic group of animals, available in the Barcode of Life Data System (BOLD). This is followed by implementing a qualitative ranking system that assigns one of five grades (A to E) to each species in the reference library, according to the attributes of the data and congruency of species names with sequences clustered in barcode index numbers (BINs). Our goal is to allow researchers to obtain the most useful and reliable data, highlighting and segregating records according to their congruency. Different tests were performed to perceive its usefulness and limitations. BAGS fulfils 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 has the potential to become a valuable addition in forthcoming DNA metabarcoding studies, in the long term contributing to globally improve the quality and reliability of the public reference libraries.

KEYWORDS

annotation, BOLD systems, DNA metabarcoding, quality control, R, reference libraries

1 | INTRODUCTION

The availability of well-curated comprehensive reference libraries is fundamental for accurate DNA barcode-based species identification (Cariani et al., 2017; Ekrem et al., 2007; Leese et al., 2016; Oliveira et al., 2016). The demand for high quality reference libraries has increased considerably since 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 from high-throughput sequencing (HTS) instruments, the required

Fontes and Vieira are joint first authors.

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 the Barcode of Life Data Systems (BOLD; Ratnasingham & Hebert, 2007) or NCBI GenBank (Sayers et al., 2019). With a few exceptions, such as R-Syst::diatom (Rimet et al., 2016), the UNITE database (Nilsson et al., 2018) or MIDORI (Machida et al., 2017), which are reference libraries compiled and curated for specific taxa, typically, there is no supervision or quality control of the reference data set. Therefore, inaccurate records in reference libraries may 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).

Errors or discordances can have operational or biological explanations. Operational errors include morphology-based misidentifications, cross-contamination of samples, mislabelling, accidental mistakes when recording data, among others (Packer et al., 2009; Pentinsaari et al., 2019; Rulik et al., 2017). Possible biological reasons for discordances include recently diverged species and incomplete lineage sorting, introgression, insufficient discrimination capacity of the barcode marker, phenotypic plasticity, among others (Costa & Antunes, 2012; Lin et al., 2018; Weber et al., 2019; Weigand et al., 2011). Although some data quality assurance and quality control (QA/QC) criteria have been implemented upstream and along the DNA barcode production workflow (e.g., Hanner, 2005), no comprehensive tool for downstream quality control of the taxonomic accuracy in DNA barcode reference libraries is available to check QA/QC in a standardized way. Some QA/QC measures are implemented in BOLD: labelling of barcode compliant records, flagging of sequences that are probably contaminations or based on misidentified specimens, flagging of sequences with stop codons (Ratnasingham & Hebert, 2007), and the possibility to run BIN-discordance reports (Ratnasingham & Hebert, 2013). However, there are several sources of potential discordance or errors that remain unscreened or unexplored through existing systems (Meiklejohn et al., 2019; Mioduchowska et al., 2018; Siddall et al., 2009; Weigand et al., 2019).

The origin of discordances and inaccuracies in DNA barcode data and DNA databases in general are well known (Harris et al., 2003; Meiklejohn et al., 2019; Mioduchowska et al., 2018; Pentinsaari et al., 2019; Siddall et al., 2009; Vilgalys, 2003), however, relatively few studies have addressed the problem of compilation, and quality control of reference libraries, particularly concerning taxonomic reliability (Leese et al., 2018; Weigand et al., 2019). For instance, CO-ARBitrator (Heller et al., 2018) detects sequences mislabelled as cytochrome c oxidase subunit I (COI), but which are originating from nonhomologous loci. The "coil" R package (Nugent et al., 2020) is also useful in detecting errors in animal barcoding and metabarcoding data by placing sequences in a reading frame and translating them to amino acids. While both packages successfully detect cases of nonhomologous barcode sequences, they do not address the issue of taxonomic congruency.

Recently, Rulik et al. (2017) proposed a preprocessing system for large data sets aiming to generate high quality DNA barcodes by verifying taxonomic consistency. However, this system requires a phylogenetic backbone for implementation, and it is meant to be used before uploading data to reference libraries. It therefore does not consider global congruence with other data already available in either BOLD or GenBank.

A large number of COI sequences are currently available in GenBank (Porter & Hajibabaei, 2018) and although a fair portion of the records may not abide to the formal barcode data standards (Ratnasingham & Hebert, 2013), they still constitute a useful resource that should not be overlooked. In fact, many metabarcoding-based studies report taxonomic assignments based on all available COI data, thereby including non-barcode compliant records. This reinforces the need for a barcode compilation, auditing and annotation system that provides an indication of the taxonomic reliability of the records for end-users of reference libraries.

Costa et al. (2012) proposed a ranking system to be implemented at the post-barcoding end of the barcode production pipeline, which considered all available sequence data for a given species (thus both barcode compliant and non-compliant). The ranking system attributes five different grades to species records (A to E), depending essentially on the level of congruency between morphospecies and the respective COI barcode clusters. Later, the system was updated to use Barcode Index Numbers (BINs; Ratnasingham & Hebert, 2013) as the reference DNA barcode clustering method (e.g., Knebelsberger et al., 2014; Oliveira et al., 2016). In global terms, the goal was to provide end-users of reference libraries with a system to sort out and annotate species that can be confidently identified with current data, from ambiguous or inaccurate records that need revision, or to flag cases of suspected hidden diversity. The implementation of this ranking system to a compilation of COI barcodes from European fish revealed that the majority of species could be confidently identified with DNA barcodes (Oliveira et al., 2016), and a number of ambiguous records could be clarified upon careful revision. However, in these implementations of the ranking system, the attribution of the grades was dependent on individual analyses of each species' data, a strategy which would be impractical for the large DNA metabarcoding reference libraries involving hundreds or thousands of species.

To address this problem, we here introduce BAGS, an R-based application for automated auditing and annotation of DNA barcode reference libraries. We adapt the proposed Oliveira et al. (2016) ranking system, essentially based on match/mismatch between BINs and morphospecies identifications. BAGS can be applied to user-provided species lists or large taxon-specific data sets 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.

2 | MATERIALS AND METHODS

2.1 | Overview of BAGS

BAGS features automated compilation of quality-filtered COI sequence data sets from BOLD, allowing for selection or exclusion of marine taxa through matching with the World Register of Marine Species (WoRMS) checklists (WoRMS Editorial Board, 2020). It delivers taxon-specific libraries annotated with qualitative grades based on BIN/morphospecies congruence and on the amount of available data for each species (A to E, see below for details), which can be downloaded whole or sorted by grade. A user-friendly interface allows for minimal operation for users nonfamiliar with R (R Development Core Team, 2019), 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 run locally, the more experienced R users have the option to make adjustments to the code. The users may then (frequently if necessary) use the annotated data sets to compile their own personalized and reviewed libraries (e.g., BOLD data sets) and use them for taxonomic assignment of HTS metabarcoding-generated reads.

BAGS is composed of four main features which are implemented in sequence (Figure 1): (a) data mining and library compilation; (b) marine taxa filter (optional); (c) library auditing and annotation; and (d) auditing output and annotation-based library sorting.

2.2 | BAGS pipeline

2.2.1 | 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. 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: 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 or 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 data set, among the four available (marine, brackish, freshwater or terrestrial).

Records are removed if at least one of the following criteria is verified: (a) records with sequences shorter than the minimum size chosen by the user (between 300 and 650 bp), 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; and (c) by default, records without information of the sampling location (either latitude or country of origin), although users can choose to include those records. Records with ambiguous expressions present in the species name (e.g., *sp.*, *complex.*, etc; see Appendix S1: https://doi.org/10.5061/dryad.2rbnzs7kx) or in the COI sequence (i.e., not IUPAC nucleotide code; see Appendix S1: https://doi.org/10.5061/dryad.2rbnzs7kx) are not removed, however, the ambiguous expression is removed.

At the end of this procedure, a filtered reference library is downloaded and available for the subsequent auditing and annotation step.

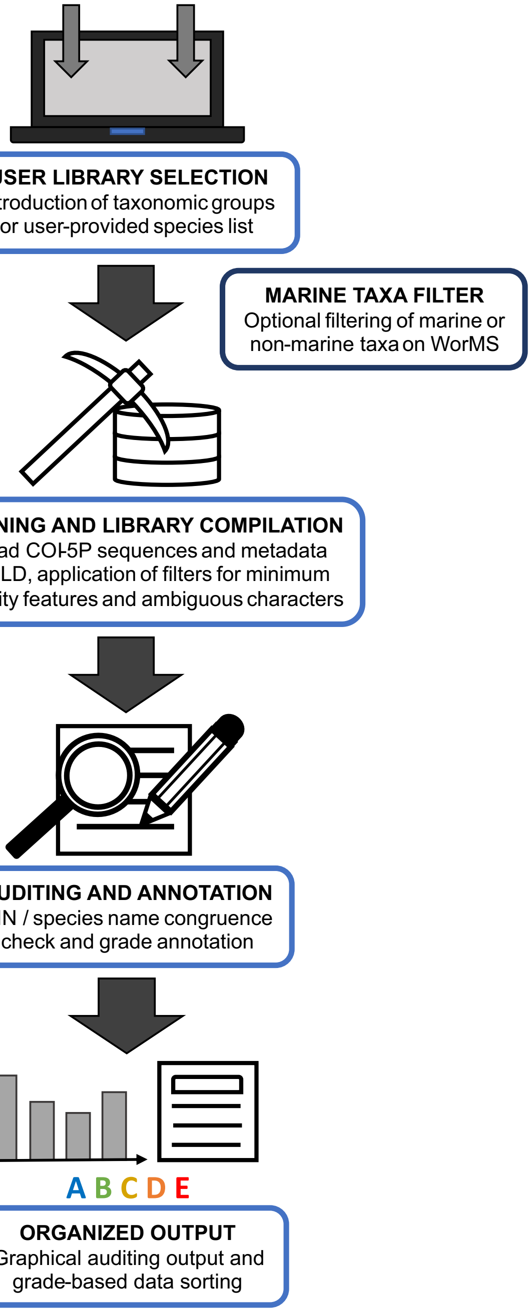
USER LIBRARY SELECTION
Introduction of taxonomic groups or user-provided species list
MARINE TAXA FILTER
Optional filtering of marine or non-marine taxa on WorMS
DATA MINING AND LIBRARY COMPILATION
Download COI-5P sequences and metadata from BOLD, application of filters for minimum size, quality features and ambiguous characters
AUDITING AND ANNOTATION
BIN / species name congruence check and grade annotation
A B C D E
ORGANIZED OUTPUT
Graphical auditing output and grade-based data sorting

FIGURE 1 Overview of BAGS' four main features and their arrangement along the informatics pipeline

2.2.2 | 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 2):

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 between three and 10 specimens in the library.

Grade C – Multiple BINs: the morphospecies is assigned to more than one BIN, and all of 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 integrates specimens from another species.

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

The BAGs auditing pipeline consists of a series of annotation steps, each comprising data checks with two possible outcomes (Figure 2). Every set of sequences for a given species entering the pipeline will be annotated with a single grade (A to E). Discordant species 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, that would limit concordance-checking to the downloaded species' data only.

BOLD (like GenBank) limits the number of searches or queries per IP/user to avoid the overload of their webservice. Therefore, to avoid blocking the access to BOLD, we periodically (approximately every two months) download the entire BOLD data set for animals and protists 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.

2.2.3 | 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

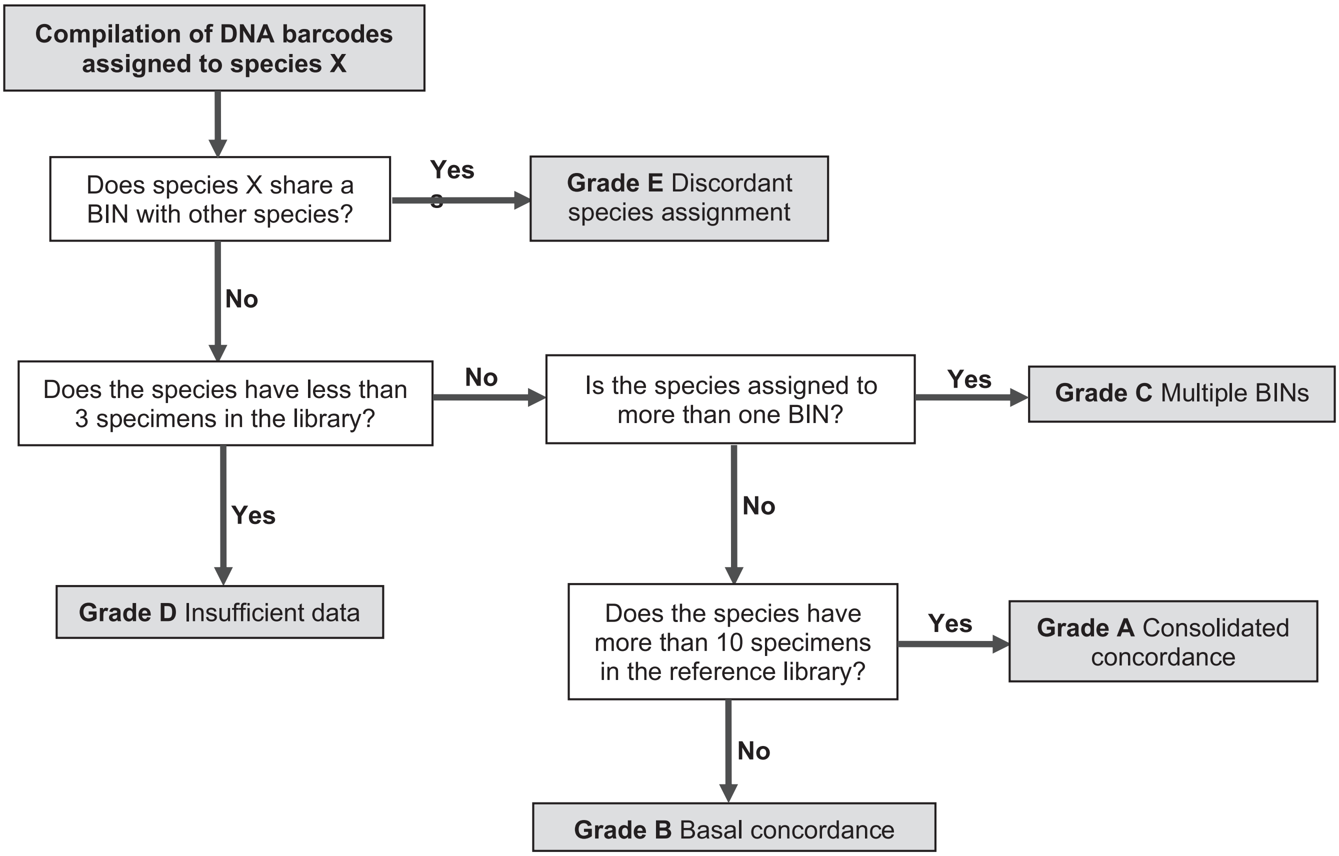
Compilation of DNA barcodes assigned to species X
Does species X share a BIN with other species?
Yes
s
Grade E Discordant species assignment
No
Does the species have less than 3 specimens in the library?
No
Is the species assigned to more than one BIN?
Yes
Grade C Multiple BINs
Yes
No
Grade D Insufficient data
Does the species have more than 10 specimens in the reference library?
Yes
Grade A Consolidated concordance
No
Grade B Basal concordance

FIGURE 2 Workflow for automated auditing and annotation of qualitative grades to each species in a BAGS-compiled reference library (adapted from Oliveira et al., 2016).

library will be created and downloaded in the form of a tabular file containing the following: species name, BIN, COI-5P 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, giving the choice of which grades to include. The fasta files can be download with all grades, combinations of different grades or separately for each grade.

Lastly, BAGs summarizes the data regarding the reference library that was created, 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 target libraries, the user must refresh the page and start over again.

2.3 | Informatic implementation

BAGS is an application written entirely in the open-source programming language R, designed using the “shiny” R package framework (Chang et al., 2019), having therefore an underlying customization with HTML and CSS. It is possible to launch BAGS locally on any environment that has R installed, as well as through any R IDE such as RStudio (RStudio Team, 2016), 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 stored at web servers and can also be used remotely, which allows its launching from any web browser (https://bags.vm.ntnu.no or https://tadeu-apps.shinyapps.io/bags; additional links are provided at https://github.com/tadeu95/BAGS). 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.

2.4 | Performance assessment

In order to test BAGS performance, two independent tests were performed. First, to understand if the marine and nonmarine taxa selection filters were functional and reliable, we downloaded three files, using the “all taxa”, the “marine taxa” and the “nonmarine” 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 three libraries previously generated.

Second, to assess the accuracy of BAGS’ auditing and grade assignment, we selected three trial reference libraries likely to display distinctive features and quality issues: marine Amphipoda (Malacostraca: Crustacea), Chironomidae (Diptera: Insecta), and marine fish (Actinopterygii, Elasmobranchii and Holocephali). These trial libraries include two key invertebrate groups in aquatic monitoring, which are likely to be relevant in metabarcoding applications, and one of the most well-represented groups of vertebrates in BOLD, thereby enabling the screening of a large and diverse number of records and species.. 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 assigned 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 used for freshwater biomonitoring under the EU Water Framework Directive (BOLD checklist DNAqua-NET: Diptera, code CL-DNADI, 584 spp. Chironomidae) in order to simplify the performance assessment. Neighbour-joining trees (Saitou & Nei, 1987) of the species assigned to grade C were created on the BOLD workbench, to evaluate the monophyly/non-monophyly of each species. 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 | Marine taxa selection 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, while the nonmarine taxa library comprised 51 species, 67 BINs and a total of 318 specimens. Comparatively, the “all taxa” option library had 123 species, 148 BINs and 1,022 specimens. The 30 species randomly sampled of the marine-filtered library were correctly assigned (i.e., all the 30 species were registered as being from marine or brackish environments when checked manually upon on WoRMS; Appendix S2: https://doi.org/10.5061/dryad.2rbnzs7kx). 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 nonmarine taxa library revealed to be all exclusive from freshwater environments (i.e. not present neither in marine or brackish waters, and therefore not present in the marine library).

3.2 | Trial data sets

The marine Amphipoda data set had a total of 6,385 specimens in the compiled library, 486 species and 736 BINs; the Chironomidae data set consisted of a total of 90,214 specimens, 1,113 species and 1,883 BINs; and the marine fishes data set comprised 107,434 specimens, 8,381 species and 9,779 BINs (Appendix S3: https://doi.org/10.5061/dryad.2rbnzs7kx). The distributions of the number of species per grade in each of the compiled reference libraries (Figure 3) show that the proportion of possible 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%). The number of species per BIN (grade E) varied between 1 and 49 for Amphipoda, 1 and 12 for Chironomidae and 1 and 88 for fish (Figure 4).

For the three groups, the 30 randomly sampled species were correctly assigned to the qualitative grades (Appendix S4: https://doi.org/10.5061/dryad.2rbnzs7kx). Grade C species (Table 1, Appendix S5: https://doi.org/10.5061/dryad.2rbnzs7kx) were mostly monophyletic: between 66% (Chironomidae) and 80% (fish). Discordances or potential errors in grade E annotations had different possible sources (Table 2). Misidentifications (between 37% and 67%) and ambiguous species names (between 10% and 33%) contributed the most to the grade E cases, while synonyms the least (overall 3.4%).

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 databases 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), our primary goal with BAGS is to facilitate the implementation of curation and quality control measures among taxonomists and biodiversity scientists.

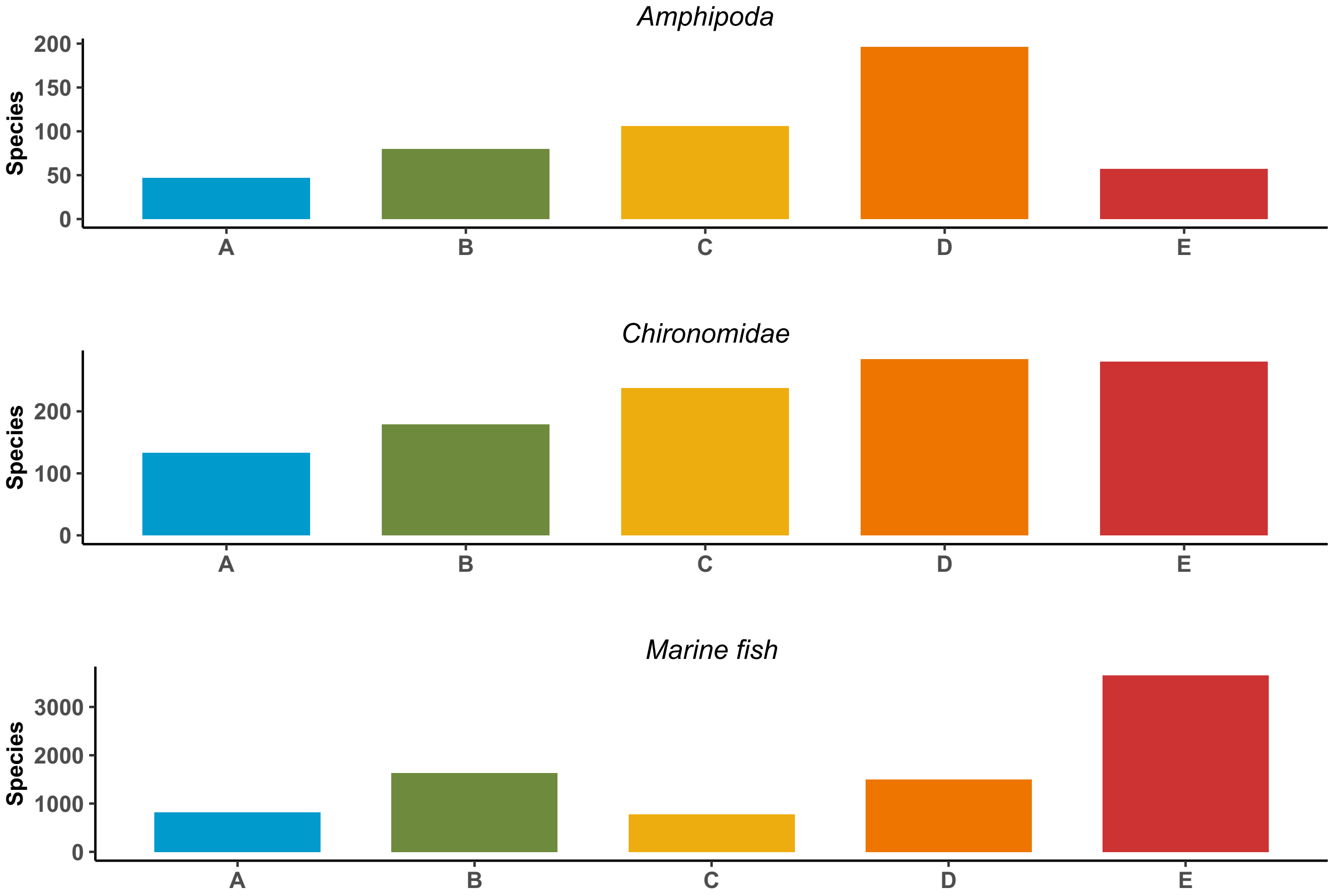

FIGURE 3 Barplots displaying the distribution of the number of species assigned to each qualitative grade for the three taxonomic groups tested. From top to bottom: marine Amphipoda, Chironomidae and marine fish (Actinopterygii, Elasmobranchii and Holocephali)

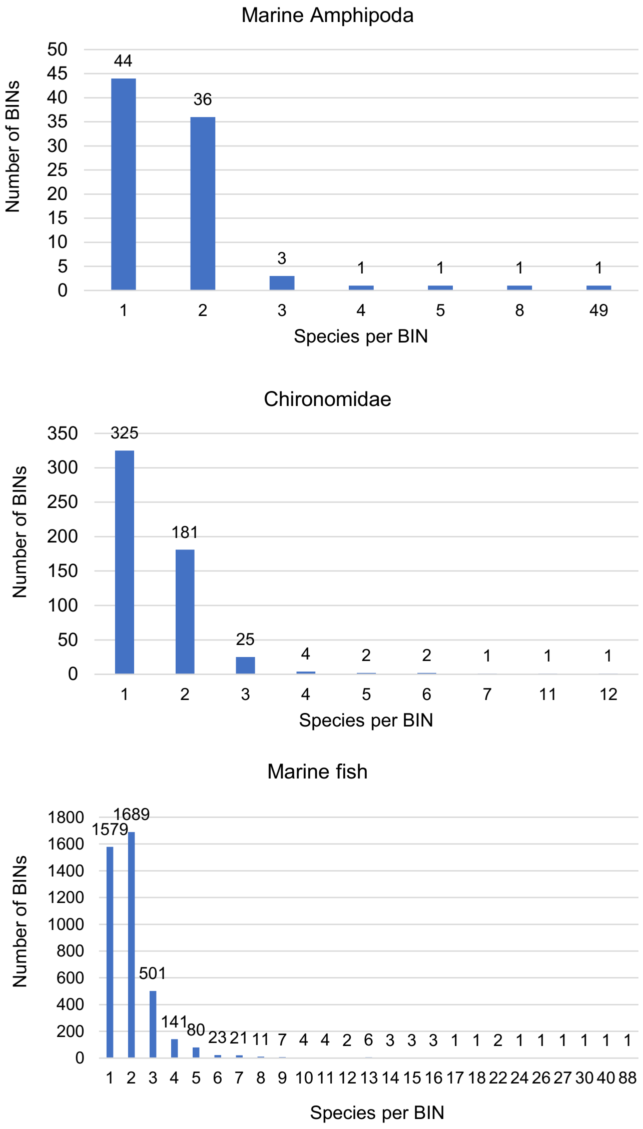

FIGURE 4 Number of species per BIN in the grade E data set generated through BAGS for each tested taxonomic group (marine Amphipoda, Chironomidae, marine fish)

TABLE 1 Percentage of monophyletic or non-monophyletic species assigned to grade C of each tested taxonomic group, according to their clustering pattern in the respective Neighbour-Joining tree

	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%

Additionally, we seek 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. First of all, 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 nondiscordant – BINS (grade C), therefore 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 (grades A, B, and D). From an auditing and taxonomic curation point of view, the morphospecies-centred approach is also more advantageous.

Ultimately, we present this application as a way to sort out taxonomic incongruencies and point out possible cases of human error during the generation of the barcodes, as well as uncovering potential cases of hidden diversity among species. Overall, our comprehensive testing of BAGS indicates that it enables researchers with a simple tool for fast screening of the quality status of massive reference libraries, thereby allowing them to sort highly robust records and pinpoint those in need of curation and revision, while unravelling the main issues that may arise during the generation of DNA barcodes for a particular group of organisms.

4.1 | BAGS performance assessment tests and some considerations

The different efficiency tests performed with BAGS (either marine/non-marine and grade annotation) allowed to verify the correct performance of this application. The different manual tests (i.e., nonautomated; Appendixes S2 and S4: https://doi.org/10.5061/dryad.2rbnzs7kx) and the ongoing tests performed by us and colleagues during beta tests, did not bring to light any errors of the application in the filtering or the auditing and annotation steps.

It is important to point out that 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* Rafinesque, 1,820). These species should not be excluded from marine reference libraries as they may be also detected in metabarcoding studies in fully marine environments. If the goal is to gather as much barcode compliant records as possible in the final data set, regardless of the habitat, it is advisable to use the “all taxa” option. However, we consider the “marine” and “non-marine” options of BAGS as a useful resource if the user wishes to use a customized and size-amenable reference library targeting preferentially only marine or nonmarine organisms.

By using three distinct taxa important in biomonitoring studies, BAGS allowed us to promptly understand the differences in the level of congruency of their available DNA barcodes and in the quality of

TABLE 2 Percentage of the different plausible origins for the assignment of grade E to species in 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%

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 (e.g., Weigand et al., 2019) or are poorly represented (high prevalence of grade D species here observed; Figure 3), reinforcing the continuous need for the completion of reference libraries.

BAGS performance tests allowed to spot a high proportion of grade C species (multiple BINs), reaching around 20% in Chironomidae and Amphipoda, but less prevalent in marine fish (Figure 3). Species with multiple BINs may occur for a number of reasons, starting with the nonoptimal BIN splitting (Ratnasingham & Hebert, 2013), *Wolbachia*-related artefacts (especially terrestrial arthropods; Smith et al., 2012), or may simply reflect phylogeographic differentiation within the same species. However, they also often suggest undescribed or cryptic diversity. 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. A number of studies have been addressing the curation of marine invertebrate's DNA barcodes, including Amphipoda (e.g. Lobo et al., 2016; Radulovici et al., 2019; Raupach et al., 2015), which may explain the lowest proportion of possible discordances (Grade E) out of the three groups analysed (Figure 3). Contrarily, the marine fishes' reference library showed a prominently high proportion (~44%) of grade E species (Figure 3), mainly due to misidentifications, consolidated morphospecies aggregated in one BIN or faulty species names lexicon (Table 2). There are some extreme cases which greatly contribute to this scenario, as for instance, BINs BOLD:AAC8034 and BOLD:AAB3926, consisting of 40 and 88 species respectively (Figure 4). In the latter case, out of 88 species, only one is spelled correctly ("*Pseudanthias squamipinnis*"), while the remaining were named "Unknown" or "*Pseudanthias* sp." followed by different alphanumeric designations. 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. Considering this and other possible grade E scenarios, we hold the view that this grade should serve as an incentive for a close examination of that particular species' records, and not as a definitive signalling of unreliability. Indeed, the detailed inspection of grade E cases after BAGS annotation revealed that most of them are probably pseudodiscordances and, if eventually clarified, could lead to an estimated overall reduction of 80% in grade E species.

4.2 | BAGS limitations

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 we identified, there are still limitations that we would like to address 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 monophyletic from non-monophyletic species, nor can it recognize synonyms and other apparent discordances, such as faulty or interim species names, in species graded E. Moreover, since BAGS implements grades which are defined based 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 versions with the introduction of customized OTU 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, this restriction would limit its efficiency. To overcome this constraint, part of the data necessary to implement the grade annotation system is regularly downloaded by us from BOLD, and used for comparison. However, since the full species name and BIN data set is locally stored for this purpose, the grade attribution can potentially change every time new barcode records and BINs are added to BOLD.

In conclusion, we can envision several prospective improvements that 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 pseudoconcordances), 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). 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, pseudodiscordances generated by synonyms or ambiguous species designations, possibly using machine learning and artificial intelligence systems. Eventually, some discordances may require individual professional judgement that cannot be accomplished with automated procedures.

It is our goal that BAGs can facilitate and stimulate the much-needed revision and curation of reference libraries. We urge all users to 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.

ACKNOWLEDGEMENTS

This study was supported by the project NextSea [NORTE-01-0145-FEDER-000032], under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). P.V. was supported through the Portuguese Foundation for Science and Technology (FCT, I.P.) in the scope of the project NIS-DNA [PTDC/BIA-BMA/29754/2017].

This manuscript contributes to the COST Action DNAqua-Net CA15219 goals, in particular Work Group 1 (WG1), and benefited from comments and suggestions over an incipient version of BAGS from participants in DNAqua-Net WG1 workshop in Limassol, Cyprus.

Finally, we would like to thank Sofia Duarte for testing BAGS, to Roger Alterskjær for all the help on the informatics and to the anonymous reviewers for their suggestions to improve the manuscript.

AUTHOR CONTRIBUTIONS

J.T.F., P.E.V., P.S., and F.O.C. designed the research plan. J.T.F., and P.E.V. wrote the BAGs script and developed the BAGS application. J.T.F., P.E.V., and T.E. performed the different assessment tests. All the authors wrote the manuscript, contributed with suggestions to the manuscript structure and reviewed the manuscript final version.

DATA AVAILABILITY STATEMENT

BAGS script and associated README file can be consulted at: https://github.com/tadeu95/BAGs. To run BAGS remotely, please use https://bags.vm.ntnu.no, https://tadeu-apps.shinyapps.io/bags or any additional link available at https://github.com/tadeu95/BAGS). All the Appendixes can be found at: https://doi.org/10.5061/dryad.2rbnzs7kx.

ORCID

João T. Fontes https://orcid.org/0000-0002-8766-4779
Pedro E. Vieira https://orcid.org/0000-0003-4880-3323
Torbjørn Ekrem https://orcid.org/0000-0003-3469-9211
Pedro Soares https://orcid.org/0000-0002-2807-690X
Filipe O. Costa https://orcid.org/0000-0001-5398-3942

REFERENCES

Bengtsson-Palme, J., Richardson, R. T., Meola, M., Wurzbacher, C., Tremblay, É. D., Thorell, K., Kanger, K., Eriksson, K. M., Bilodeau, G. J., Johnson, R. M., Hartmann, M., & Nilsson, R. H. (2018). Metaxa2 Database Builder: Enabling taxonomic identification from metagenomic or metabarcoding data using any genetic marker. *Bioinformatics*, *34*(23), 4027–4033. https://doi.org/10.1093/bioinformatics/bty482

Cariani, A., Messinetti, S., Ferrari, A., Arculeo, M., Bonello, J. J., Bonnici, L., Cannas, R., Carbonara, P., Cau, A., Charilaou, C., El Ouamari, N., Fiorentino, F., Follesa, M. C., Garofalo, G., Golani, D., Guarniero, I., Hanner, R., Hemida, F., Kada, O., … Tinti, F. (2017). Improving the conservation of mediterranean chondrichthyans: The ELASMOMED DNA barcode reference library. *PLoS One*, *12*(1), e0170244. https://doi.org/10.1371/journal.pone.0170244

Chamberlain, S. (2019). bold: Interface to Bold Systems API. R package version 0.9.0. https://cran.r-project.org/package=bold

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2019). shiny: Web Application Framework for R. R package version 1.4.0. https://cran.r-project.org/package=shiny

Costa, F. O., & Antunes, P. M. (2012). The contribution of the Barcode of Life initiative to the discovery and monitoring of biodiversity. In A. Mendonca, A. Cunha, & R. Chakrabarti (Eds.), *Natural resources, sustainability and humanity: a comprehensive view* (pp. 37–68). Springer. https://doi.org/10.1007/978-94-007-1321-5_4

Costa, F. O., Landi, M., Martins, R., Costa, M. H., Costa, M. E., Carneiro, M., Alves, M. J., Steinke, D., & 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. https://doi.org/10.1371/journal.pone.0035858

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. https://doi.org/10.1086/696613

DeSalle, R., & Goldstein, P. (2019). Review and interpretation of trends in DNA Barcoding. *Frontiers in Ecology and Evolution*, *7*, 302. https://doi.org/10.3389/fevo.2019.00302

Desiderato, D., Costa, F. O., Serejo, C., Abiatti, 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. https://doi.org/10.1111/zsc.12339

deWaard, J. R., Ratnasingham, S., Zakharov, E. V., Borisenko, A. V., Steinke, D., Telfer, A. C., Perez, K. H. J., Sones, J. E., Young, M. R., Levesque-Beaudin, V., Sobel, C. N., Abrahamyan, A., Bessonov, K., Blagoev, G., deWaard, S. L., Ho, C., Ivanova, N. V., Layton, K. K. S., Lu, L., … 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. https://doi.org/10.1038/s41597-019-0320-2

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. https://doi.org/10.1007/s13127-010-0034-y

Ekrem, T., Willassen, E., & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, *43*(2), 530–542. https://doi.org/10.1016/j.ympev.2006.11.021

Hanner, B. R. (2005). Proposed Standards for BARCODE Records in INSDC (BRIs). Technical report, Database Working Groups, Consortium for the Barcode of Life, 2009.

Harris, T. W., Lee, R., Schwarz, E., Bradnam, K., Lawson, D., Chen, W., & Stein, L. D. (2003). WormBase: A cross-species database for comparative genomics. *Nucleic Acids Research*, *31*(1), 133–137. https://doi.org/10.1093/nar/gkg053

Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data*, *5*, 180156. https://doi.org/10.1038/sdata.2018.156

Hobern, D., & Hebert, P. D. N. (2019). BIOSCAN – revealing eukaryote diversity, dynamics, and interactions. *Biodiversity Information Science and Standards*, *3*, e37333. https://doi.org/10.3897/biss.3.37333

Holstein, J. (2018). worms: Retriving Aphia Information from World Register of Marine Species. R package version 0.2.2. https://cran.r-project.org/package=worms

Hupało, K., Teixeira, M., Rewicz, T., Sezgin, M., Iannilli, V., Karaman, G. S., Grabowski, M., & 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. https://doi.org/10.1016/j.ympev.2018.11.013

Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., & Ankenbrand, M. J. (2020). BCdatabaser: On-the-fly reference database creation for (meta-)barcoding. *Bioinformatics*, *36*(8), 2630–2631. https://doi.org/10.32942/osf.io/cmfu2

Knebelsberger, T., Landi, M., Neumann, H., Kloppmann, M., Sell, A. F., Campbell, P. D., Laakmann, S., Raupach, M. J., Carvalho, G. R., & 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. https://doi.org/10.1111/1755-0998.12238

Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., Mergen, P., Pawlowski, J., Piggott, J., Rimet, F., Steinke, D., Taberlet, P., Weigand, A., Abarenkov, K., Beja, P., Bervoets, L., Björnsdóttir, S., Boets, P., Boggero, A., … Zimmermann, J. (2016). DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes*, *2*, e11321. https://doi.org/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. https://doi.org/10.1016/bs.aecr.2018.01.001

Lin, X. L., Stur, E., & Ekrem, T. (2015). Exploring genetic divergence in a species-rich insect genus using 2790 DNA barcodes. *PLoS One*, *10*(9), e0138993. https://doi.org/10.1371/journal.pone.0138993

Lin, X. L., Stur, E., & Ekrem, T. (2018). DNA barcodes and morphology reveal unrecognized species in Chironomidae (Diptera). *Insect Systematics and Evolution*, *49*(4), 329–398. https://doi.org/10.1163/1876312X-00002172

Lobo, J., Ferreira, M. S., Antunes, I. C., Teixeira, M. A. L., Borges, L. M. S., Sousa, R., Gomes, P. A., Costa, M. H., Cunha, M. R., & Costa, F. O. (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. https://doi.org/10.1139/gen-2016-0009

Machida, R., Leray, M., Ho, S., & Knowlton, N. (2017). Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, *4*, 170027. https://doi.org/10.1038/sdata.2017.27

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), e0217084. https://doi.org/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), e0199609. https://doi.org/10.1371/journal.pone.0199609

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2018). The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1111/j.1755-0998.2009.02631.x

Pennisi, E. (2019). DNA barcodes jump-start search for new species. *Science*, *364*(6444), 920–921. https://doi.org/10.1126/science.364.6444.920

Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? *PLoS One*, *15*(4), e0231814. https://doi.org/10.1371/journal.pone.0231814

Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS One*, *13*(9), e0200177. https://doi.org/10.1371/journal.pone.0200177

R Development Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Radulovici, A. E., Costa, F. O., & the Hackathon Participants (2019). New avenues for data curation: Hackathon on marine invertebrates. *Genome*, *62*(6), 422. https://doi.org/10.1139/gen-2019-0083

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. https://doi.org/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), e66213. https://doi.org/10.1371/journal.pone.0066213

Raupach, M. J., Barco, A., Steinke, D., Beermann, J., Laakmann, S., Mohrbeck, I., Neumann, H., Kihara, T. C., Pointner, K., Radulovici, A., Segelken-Voigt, A., Wesse, C., & Knebelsberger, 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. https://doi.org/10.1371/journal.pone.0139421

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A., & Bouchez, A. (2016). R-Syst:Diatom: An open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, *2016*, 1–21. https://doi.org/10.1093/database/baw016

RStudio Team (2016). *RStudio: Integrated Development for R*. RStudio Inc. 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. https://doi.org/10.1111/2041-210X.12824

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research*, *47*(D1), D94–D99. https://doi.org/10.1093/nar/gky989

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. https://doi.org/10.1093/sysbio/syp033

Smith, M. A., Bertrand, C., Crosby, K., Eveleigh, E. S., Fernandez-Triana, J., Fisher, B. L., Gibbs, J., Hajibabaei, M., Hallwachs, W., Hind, K., Hrcek, J., Huang, D.-W., Janda, M., Janzen, D. H., Li, Y., Miller, S. E., Packer, L., Quicke, D., Ratnasingham, S., … Zhou, X. (2012). *Wolbachia* and DNA barcoding insects: Patterns, potential, and problems. *PLoS One*, *7*(5), e36514. https://doi.org/10.1371/journal.pone.0036514

Vilgalys, R. (2003). Taxonomic misidentification in public DNA databases. *New Phytologist*, *160*(1), 4–5. https://doi.org/10.1046/j.1469-8137.2003.00894.x

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. https://doi.org/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. https://doi.org/10.1111/j.1755-0998.2010.02937.x

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wyler, S. A., Bouchez, A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, 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. https://doi.org/10.1016/j.scitotenv.2019.04.247

WoRMS Editorial Board (2020). World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. https://doi.org/10.14284/170

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO. BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. *Mol Ecol Resour*. 2020;00:1–11. https://doi.org/10.1111/1755-0998.13262