Assembling and auditing a comprehensive DNA barcode reference library for
European marine fishes

L. M. Oliveira¹, T. Knebelsberger², M. Landi¹, P. Soares¹, M. J. Raupach², F. O. Costa¹*

¹Center of Molecular and Environmental Biology (CBMA), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
²Senckenberg am Meer, German Center for Marine Biodiversity Research (DZMB), Suedstrand 44, 26382 Wilhelmshaven, Germany

Running headline: DNA barcode library for European marine fish

*Author to whom correspondence should be addressed:
Telephone: +351 253601527
Facsimile: +351 253678980
Email address: fcosta@bio.uminho.pt
ABSTRACT

A large-scale comprehensive reference library of DNA barcodes for European marine fishes was assembled, allowing the evaluation of taxonomic uncertainties and species genetic diversity, which were otherwise hidden in geographically restricted studies. A total of 4,118 DNA barcodes were assigned to 358 species generating 366 BINs (Barcode Index Number). Initial examination revealed as much as 141 BIN discordances (more than one species in each BIN). After implementing an auditing and 5-grade (A to E) annotation protocol, the number of discordant species BINs was reduced to 44 (13% / grade E), while concordant species BINs amounted to 271 (78% / grades A and B), and 14 other had insufficient data (grade D). Fifteen species displayed comparatively high intraspecific divergences ranging from 2.0% to 18.5% (grade C), which is biologically paramount information to be considered in fish species monitoring and stock assessment. On balance, this compilation contributed to the detection of 59 European fish species in likely need of taxonomic clarification or re-evaluation. The generalized implementation of an auditing and annotation protocol for reference libraries of DNA barcodes is recommended.

Key words

Marine fishes; DNA barcode; reference library; taxonomic reliability grade; Barcode Index Number; hidden diversity
INTRODUCTION

DNA barcoding, especially the partial sequencing of cytochrome c oxidase subunit I (COI), has been successfully employed as a molecular tool for the identification and discrimination of fish species in the past (Knebelsberger et al., 2014). Nevertheless, given the increasing number of publications involving DNA barcodes of European marine fish, a global synthesis of these data, including the compilation and annotation of a reference library, is still lacking. Despite the frequently large distance separating samples, previous studies showed the reliability of DNA barcoding for marine fish identification independently of geographic distance (Ward et al., 2008; Zemlak et al., 2009).

Apart from the compilation, the main objective is to analyze the consistency of DNA barcodes obtained by independent research groups. Public databases, namely GenBank and BOLD (Barcode of Life Data System; Ratnasingham & Hebert, 2007), are susceptible to operational errors, including inaccurate taxonomic identification of the original specimens and insufficient quality of the molecular data and metadata (Knebelsberger et al., 2014). Methodological control measures are imperative, including species identification by expert taxonomists and submission of compliant data according to the requirements of the Barcode Data Standards (Walters & Hanner, 2006).

Post-barcoding annotation tools for libraries are vital to maintain the quality standards of the compiled data, as for example the assignment of categories of taxonomic reliability of DNA barcodes (Costa et al., 2012). Such approaches combined with automated analysis tools secure the quality of the library and allows the user, either skilled or not, to use it confidently with high reliability. A reference library, in addition to its use as a robust tool for the identification of sequences from unknown organisms.
(Costa et al., 2012), is also essential for applications involving authentication of fishery products (Hanner et al., 2011), either fresh or processed (Carvalho et al., 2015), and detection of illegal use of protected species for biosecurity (Armstrong & Ball, 2005; Rasmussen & Morrissey, 2008).

In the specific case of European species, such reference library is valuable to assist the identification and management of fish stocks, frequently shared between the member states (Landi et al., 2014), either through the detection of mixed fisheries containing mislabeled species, or through the assessment of regional biodiversity of a given species or by enabling tools for authenticity of fish stocks (Mariani et al., 2015). The objective of this work is to assemble for the first time a large-scale comprehensive reference library of DNA barcodes for European marine fishes, based on all publicly available DNA barcodes, in order to examine and annotate the consistency and reliability of records obtained independently from multiple regions and studies.

MATERIAL AND METHODS

DATA GENERATION AND COMPILATION

A dataset (DS-EUROFISH, doi: dx.doi.org/10.5883/DS-EUROFISH) was created on BOLD, including samples previously generated by the research groups authoring the current manuscript, encompassing samples from the Atlantic, Mediterranean Sea, North Sea and Baltic Sea, as well as sequences obtained from BOLD projects, and GenBank sequences associated with publications, focusing on European marine fishes. The compilation effort followed the previously suggested
quality criteria for COI sequences (Walters & Hanner, 2006). In addition, new COI barcode sequences were obtained from specimens collected on the Portuguese coast and in the Baltic Sea, following published protocols (Costa et al., 2012). The sequences were submitted to GenBank (Accessions KX586190-KX586232) and added to DS-EUROFISH, where the respective metadata can be consulted. The final dataset is summarized in Table I.

DATA ANALYSES AND ANNOTATION

All sequences listed in Table I were aligned using MAFFT version 7 (Katoh & Standley, 2013). Bayesian Inference (BI) was used to create a phylogenetic tree in order to visualize the sequence clustering pattern. The software MrBayes, version 3.2 (Ronquist et al., 2012) was used to produce the BI tree, using the best fit substitution model GTR + G + I, which was determined using IQ-TREE, version 1.3.0 (Nguyen et al., 2014). The analysis was run for 2 million iterations in two parallel runs with 4 chains each, and with tree sampling every 500 iterations (4000 trees sampled). A burn-in of 25% was used, discarding the first 1000 sampled trees.

The Barcode Index Number (BIN) system (Ratnasingham and Hebert, 2013) was used for the assignment of molecular operational taxonomic units (MOTUs). BINs were examined for the whole DS-EUROFISH library using the ‘BIN Discordance Report’ analysis tool available on BOLD. Average pairwise distances between BINs were estimated using the Kimura 2-parameter (K2P) model (Kimura, 1980), implemented in the “Distance summary” tool in BOLD. This model was selected because of its generalized use in the barcoding literature, therefore facilitating comparison of reported distances between studies.
In order to assess the level of taxonomic reliability in the library, species-specific DNA barcode subsets were ranked from Grade A to E as described before (Costa et al., 2012; Borges et al., 2016). The basis of such rating systems is that taxonomic reliability is greater if barcode sequences from independent researchers cluster unambiguously and consistently for a given species. Following the procedure illustrated in Figure 1, species-specific DNA barcodes were ranked as:

Grade A: External concordance: unambiguous BIN match between specimens of the same morphospecies from independent BOLD projects or published sequences.

Grade B: Internal concordance: species’ BIN congruent within one dataset, with at least 3 specimens of the same species examined but no matching sequences found from independent studies.

Grade C: Suboptimal concordance (putative intraspecific genetic structure): at least 3 specimens of the same morphospecies are available within the library and split among more than one nearest neighbouring BIN.

Grade D: Insufficient Data: low number of specimens analysed (1 or 2 individuals) and no matching sequences available in BOLD.

Grade E: Discordant species assignments: sequences for a given species in our data set did not match with the BIN (or BINs) for the same species in BOLD. The specimen may match with a BIN of a different species or was assigned to a separate non-neighbouring BIN.

The auditing procedure here followed, assumes that automated BIN attribution and discordance flagging cannot account for all potential flaws in the DNA barcode pipeline, requiring a detailed inspection and judgment for each individual case. BINs discordances can be attributed fundamentally to 3 sets of reasons: either morphology or
molecular-based evidence do not reflect accurately the species boundaries, or a set of diverse operational failures, inaccuracies or limitations along the DNA barcoding pipeline produce misleading discordances. The latter include, among other, inaccurate morphological identifications, synonyms and misapplied species names, mislabeled specimens, cross-contamination during DNA extraction or amplification procedures, or eventually, failure of the BIN clustering algorithm to discriminate species with very low interspecific distances. The discordant BIN revision step introduced in the auditing and annotation protocol (Fig. 1), provides an opportunity for a personal evaluation by a skilled auditor in order to discard possible operational artefacts. Some artefacts were straightforwardly spotted, as in the case of synonyms or misapplied names, using FishBase (Froese & Pauly, 2015) as a reference for accepted species names. Other types of artefacts, such as contamination or mislabelling were screened out applying the majority rule (Ratnasingham and Hebert, 2013), in the cases where within a BIN with a large majority of congruent DNA barcodes, generated from various independent sources, there was one or few outstanding accompanying sequences from a taxonomically distant species originated from a single source. When BINs discordances could not be confidently ruled out, the grade E was attributed as a precautionary measure, until further evidence can help clarify the nature of the data disagreement.

For annotation purposes the “extra info” field implemented in BOLD was used to inscribe the attributed grade, followed by the auditor’s initials and date. BOLD also allows to complement this procedure with pre-established tags that can be associated with the specimen record (e.g. “contamination” or “misidentification”), or new ones that may be created at the user’s discretion.
RESULTS

A total number of 4,118 DNA barcodes distributed over 358 species of 34 orders were compiled, mined from 18 BOLD projects and 13 publications (Table I), four of them with no project associated on BOLD, only GenBank accessions (Moura et al., 2008; Straube et al., 2010; Serra-Pereira et al., 2011; Ardura et al., 2013). All of the specimens were identified down to the species level and 43 sequences are originally published under this study. The DS-EUROFISH library contains three fish classes with more than three quarters of the species belonging to the class Actinopterygii (bony fishes), followed by the Elasmobranchii class (cartilaginous fish) and the Holocephali with only two species. The distribution of samples follows similar patterns with bony fishes represented by more than 3,000 sequences and the remaining mostly from the Elasmobranchii class.

DNA barcodes were assigned to 366 distinct BINs corresponding to the before-mentioned 358 species in the library. A total of 213 concordant BINs (58%), basically indicating BINs containing records from only one species, were found, whereas 141 (39%) were discordant, displaying at least two different species within a single BIN. Furthermore, 12 BINs that include only one single sequence were also detected.

Subsequent inspection of the BIN composition revealed potential artifacts (i.e. synonyms, misidentifications) that led to an overestimation and unrealistic percentage of discordant BINs. A total of 97 in 141 discordant BINs, more than half of the putative discordant BINs, displayed further concordance following a careful inspection of the entries in the database (see auditing procedure in Fig. 1). This reveals that the discordance was due to either misidentified records or from records with incomplete taxonomy. These cases are characterized by BINs displaying a high level of taxonomic
and a wide geographic range. One example of such case is the species *Boops boops* (L. 1758) which is validated by 87 records containing entries by different researchers; however, the BIN also contains two entries of *Oblada melanura* (L. 1758), a species that is found in a separate BIN with 21 concordant entries. It is very likely that records of *O. melanura* found in the *B. boops* BIN cluster are caused by misidentification. In addition, cases of incomplete taxonomy were relatively common along BINs. For example the BIN containing *Gadus morhua* L. 1758 contained 22 entries identified only to the class Actinopterygii, resulting in an erroneous classification as discordant. A few cases also included a discordant classification due to the use of synonymous and unaccepted taxonomic names, as in the case of the BIN cluster of *Chelidonichthys lucerna* L. 1758. Subsequently to the inspection of the BINs for artifacts, the number of discordant ones decreased to 44.

Following the ranking system for taxonomic reliability (Costa *et al.*, 2012), a total of 242 species (70% of a total of 344 morphospecies with attributed BIN) can be classified with the highest level of reliability (Grade A), meaning that each species was allocated consistently with a single BIN providing for the user an unequivocal identification of a given species. Grade B was assigned to 29 species (8%) with concordant BINs but limited to a single study and no matching sequences in BOLD, whereas 15 species (4%) showed suboptimal concordance and were graded C. Their divergence into neighbouring BINs was mostly associated with geographical clustering. Fourteen of the species examined (4%) had a low number of sequences available (<3), and therefore were assigned to grade D. A considerable percentage of species in the reference library – 13% (44 species) – showed taxonomic ambiguity. This includes also economically important species, which were allocated into BINs containing several
species but from the same genus. Table II lists the species, or groups of species, which were attributed grade E, together with an annotation about the reasons for discordance and possible justification.

The 15 species graded C showed distances between BINs higher than 2% and reaching 18.5% in one case. These 15 species were assigned to a total of 36 BINs, from 2 to 4 BINs per species. Results are displayed in Table III. In most cases the records of a species were assigned in two different BINs, and the specimens were sorted among BINs according to their geographical origin. The most common geographic splits were obtained between the Atlantic and the East Mediterranean (4 species), between the Atlantic and the North Sea (2 species) and between the west and east Mediterranean (2 species). Examples of intraspecific structure and geographically sorted monophyletic clusters in three of the C-graded fish species can be visualized in a section of the BI phylogenetic tree displayed in Fig. 2 (full tree available in Fig. S1). Three species within the Atlantic Ocean were divided into 2 BINs, independently of no geographical separation. The remaining species contained two or more BINs where specimens were not geographically sorted. Further investigation on the status of these species as a unit is warranted.

DISCUSSION

The relevance of the implementation of a post-barcode auditing and annotation procedure to the European fish reference library was illustrated in the present paper by the significant reduction of discordant BINs reported after individual inspection and judgment (from 141 to 44). In addition to the examples of BIN discordance artifacts provided in the methods and results, there were examples of the occasional inadequacy
of the BIN clustering algorithm to discriminate species with very low interspecific
distances. Such is the case of the genus *Trachurus*, namely *Trachurus mediterraneus*
(Steindachner 1868), *Trachurus picturatus* (Bowdich 1825), and *Trachurus trachurus*
(L. 1758), three well-established species, each one holding its exclusive set of DNA
barcode haplotypes and forming neighboring monophyletic clusters. Yet, species which
were finally attributed with grade E, still represent a fair proportion of the total (13%).
Although there will be cases of species which cannot be resolved with DNA barcodes,
as for example the shad species *Alosa alosa* and *Alosa fallax* due to mtDNA
introgression (Alexandrino *et al.*, 2006; Faria *et al.*, 2012), the status of other grade E
species may be eventually clarified as additional data become available and detailed
studies are performed (e.g. gobies, Knebelsberger & Thiel, 2014).

The library compilation and auditing procedure here followed was also crucial in
the detection of some species exhibiting comparatively high levels of intraspecific
genetic distances (grade C species). Extensive data on COI barcode variation in
thousands of fish species shows that the vast majority of well-established species have
average intraspecific COI distances below 2% (Ward *et al.*, 2009; Ward, 2012). The 15
cases listed in Table III, therefore require additional investigation and verification of
their species status, ideally entailing a morphological and multi loci revision of
specimens from populations across the distribution range. Independently of the
conclusions of such revisions, the occurrence of highly divergent and geographically
segregated intraspecific mitochondrial lineages is a strong indication of population
isolation that should be considered for stock management and conservation purposes.
An annotated DNA barcode library can be of great utility to help mapping such lineages
in greater detail, and to provide a basis for lineage (or eventually stock) identification in
fisheries landings and, consequently, improving lineage or stock-specific catch statistics.

Overall, the annotation of the reference library of European marine fish produced a clear majority of species with a high level of data congruence and taxonomic reliability (70% and 8% A and B grades, respectively), meaning that DNA barcode-based identifications of those species are very robust. Furthermore, attribution of grade C to a species does not preclude its robust DNA barcode-based identification, but, on the contrary, may enable gathering more detailed geographic or stock-specific data on that species. As new DNA barcode data are generated and made available for more species and populations, additional auditing and annotation must be carried out regularly. Through such regular reviewing, grades may be changed and, by means of an iterative process, the expected trend is that species move progressively to upper grades due to the continuous refinement of the data and the auditing process: grade D and E species will tend to be re-assigned to upper grades, and grade B species will be re-assessed in light of new data from independent sources confirming or refuting initial congruence. Grade A species are also subject to re-assignment, but much less likely to change.

A global appraisal of the completeness of the reference library for European marine fish, reveals that the available COI barcode data only covers a small fraction of the reported ichthyofauna, notably only about 28% of the species reported for the Portuguese EEZ and extended continental platform area (Carneiro et al., 2014), or even a lower proportion (26.5 %) considering all ichthyofauna listed for Europe in the European Register of Marine Species (Costello et al., 2006). Hence, substantial research commitment is still required to complete the reference library for European marine fish, although the existing core library already covers the majority of the most abundant and
commercially relevant species. The availability of a comprehensive reference library, dully audited and annotated, for European ichthyofauna provides a crucial framework for a DNA-based identification system of fish species, with far-reaching applications and benefits for fish biology, ecology, fisheries and fisheries products quality control (Costa & Carvalho, 2007). The emergence of second and third generation sequencing technologies further expanded the potential of DNA-based identification systems, particularly by enabling species identification from community or environmental samples, rather than from individual specimens sequentially (Bohmann et al., 2014; Creer et al., 2016). Supported by this technology, ecosystem-based approaches to ichthyofaunal ecology and fisheries can be applied which incorporate analysis of different trophic levels and biotic interactions. Among other applications, it can be used for high-throughput species identification in ichthyoplankton surveys (Bucklin et al., 2016), gut content analyses and trophic web research (Leray et al., 2013; Leray et al., 2015), facilitation of species identification in processed food, commercial markets and food industry (Shokralla et al., 2016) as well as for non-invasive monitoring of fish species in environmental DNA (eDNA) obtained from seawater (Bohmann et al., 2014; Thomsen & Willerslev, 2015).

This study clearly demonstrates that only by integrating data from multiple sources it is possible to unravel pertinent cases of taxonomic uncertainties and hidden species diversity that otherwise would have remain unnoticed. The cases of deep within-species divergences detected constitute biologically meaningful information that should be considered in fish species monitoring and stock assessment. The geographically focused assembly and auditing of DNA barcodes is therefore essential to assure the robustness and consistency of the reference libraries. To this end, it is strongly recommended that an auditing an annotation framework, such as the one here applied, is
adopted by the research community to fully substantiate the potential of the reference
libraries, and to improve their accuracy and utility to the various end-users.

The authors would like to thank the three independent reviewers for their helpful
reviews of this manuscript. This work was supported by national funds through the FCT
I.P. and by the ERDF through the COMPETE2020 – Program Operacional
Competitividade e Internacionalização (POCI) in the scope of the strategic programme
UID/BIA/04050/2013 (POCI-01-0145-FEDER-007569). PS is supported by FCT, ESF,
POPH and the FCT Investigator Programme (IF/01641/2013). ML was supported by an
FCT fellowship with reference SFRH/BPD/45246/2008. We thank R. Martins and M.
Carneiro (IPMA, Portugal), as well as A. Kijewska and R. Wenne (Institute of
Oceanology, Polish Academy of Sciences, Poland) for providing specimens from
Portugal and from the Baltic Sea, respectively, used to generate the new sequences here
reported.

Supporting Information

Supporting Information may be found in the online version of this paper:

Fig. S1. Bayesian inference (BI) tree constructed using COI barcode sequence data from
4118 sequences assigned to 358 marine fish species. A best-fit substitution model
(GTR+I+G) was applied.
References


Electronic References