ARPHA Conference Abstracts 4: e65504
doi: 10.3897/aca.4.e65504

Conference Abstract

Venturing into auditing of reference libraries: from the hackathon on marine invertebrates to sorting with BAGS

Filipe O. Costa [‡, §]

‡ CBMA - Centre of Molecular and Environmental Biology, University of Minho, Braga, Portugal
§ Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Braga, Portugal

Corresponding author: Filipe O. Costa (fcosta@bio.uminho.pt)

Received: 04 Mar 2021 | Published: 04 Mar 2021

Citation: Costa FO (2021) Venturing into auditing of reference libraries: from the hackathon on marine invertebrates to sorting with BAGS. ARPHA Conference Abstracts 4: e65504. https://doi.org/10.3897/aca.4.e65504

Abstract

Reference libraries of DNA sequences are the backbone of DNA-based taxonomic identification systems. The quality and accuracy of the data in reference libraries is critical to achieve reliable identifications. Faulty or inaccurate data may have detrimental impacts in various downstream applications, perpetuating errors over long-term studies and biodiversity data repositories. This risk is particularly prevalent in metabarcoding approaches, where millions of sequences are assigned to taxa in reference libraries through automated and frequently unsupervised procedures. Although quality-compliance measures have been implemented in several stages of the DNA barcode production workflow, no systematized approach has tackled the challenges of revision, curation and annotation of reference libraries. The trend for increasing detection of cryptic diversity further complicates this task.

Here we outline the conclusions of the application of two distinct approaches to audit and annotate reference libraries: the hackathon on marine invertebrates hosted by the 8th IBOL conference, and the bioinformatics application "Barcode, Audit & Grade System" (BAGS; Fontes et al. 2021). The former consisted on the assembly of 18 researchers involved in marine barcoding, aiming to audit and annotate a very large number DNA barcode records available in BOLD from major marine invertebrate taxa, including all or selected groups of Annelida, Crustacea, Echinodermata and Mollusca. Discordant Barcode Index Numbers (BINs), that is, BINs including more than one species, were reviewed individually, and the respective records annotated with one of the 4 following tags: MIS-ID (misidentification); AMBIG (ambiguous, unable to resolve); COMPLEX (multiple BINs); SHARE (barcodes shared among species in the same BIN). This effort resulted in the processing of >80.000 barcodes, corresponding to >7.500 species, of which 7% were tagged MIS-ID, 17% AMBIG, 13% COMPLEX and 1% SHARE, with Gastropoda displaying particularly high levels of ambiguity. The sizeable portion of MIS-ID and AMBIG tags raises concern. Yet, part of the AMBIG tags merely reflect underlying uncertainty in species taxonomic status, rather than the deposition of erroneous data in BOLD. Hence, in addition to auditing and annotation, extensive effort should continue to be allocated to the underpinning alpha taxonomy of reference libraries.

The second approach here described is BAGS, which consists on an R-based application that provides an user-friendly platform for automated auditing of user-selected metazoan cytochrome oxidase I (COI) reference libraries. BAGS sorts BOLD's records and species into 5 grades, depending on whether they display BIN concordance (A, B) multiple BINs (C), less than two records (D) or discordant BINs (E). A WoRMS-linked filter allows to select or exclude marine taxa, and a reporting component provides a graphical overview and FASTA files assorted in different combinations of grades. Therefore, BAGS can provide a quick appraisal of the status of an user-defined reference library, allowing simultaneously to recognize the most reliable records, the incidence of cases high intraspecific divergence, gaps in representativeness, and inaccuracies of potential concern. A pilot assessment of BAGS performance in three datasets comprising marine fish, Chironomidae (Insecta) and marine Amphipoda (Crustacea) highlighted the differences in the congruence status of the respective reference libraries.

In conclusion, the hackathon had and expressive contribution to the revision and annotation of a very large number of marine invertebrate records lodged in BOLD. Human-mediated revision is highly-reliable and consequential, however, it constituted a massive undertaking that can hardly be repeated without a previous refinement and substantial reduction of the datasets to be revised. This could be achieved resorting to automated revision systems, among which BAGS constitutes a first step. We intend to progress with the expansion and improvement of BAGS, namely by introducing further refinements in the analyses of grade E data, in order to automatically discard simple cases of discordance, thereby reducing the amount of data needing human-mediated revision. Recognition of the need for automated reference library auditing and curation systems is essential to raise confidence of researchers, environmental managers and governmental agencies for the adoption and implementation of DNA-based approaches in aquatic biomonitoring.

Keywords

DNA (meta)barcoding, BOLD, curation, COI, Barcode Index Number

Presenting author

Filipe O. Costa

Presented at

1st DNAQUA International Conference (March 9-11, 2021)

Acknowledgements

The hackathon was organized with financial support from DNAqua-Net in the scope of the 8th International Barcode of Life Conference in Trondheim, Norway in June 2019. We thank DNAqua-Net for the funding provided and the local conference organizers for all the logistical support received. We are grateful to the BOLD team for their help with data queries and analytics. We also thank the hackathon participants for vibrant discussions during and after the event.

BAGS development was supported by the project NextSea [NORTE-01-0145-FEDER-000032], under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). It 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.

References

- Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO (2021) BAGS: an automated Barcode, Audit & Grade System for DNA barcode reference libraries. Molecular Ecology Resources 21: 573-583. https://doi.org/10.1111/1755-0998.13262