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

João Tadeu Fontes[1], Pedro Vieira[1,1], Torbjørn Ekrem[2], Pedro Soares[1], and Filipe Costa[1]

[1]University of Minho
[2]Norwegian University of Science and Technology

June 5, 2020

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, BAGS (Barcode, Audit & Grade System), 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 ultimate 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 have 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.

Hosted file

BAGS_main document.pdf available at https://authorea.com/users/317355/articles/457103-bags-an-automated-barcode-audit-grade-system-for-dna-barcode-reference-libraries

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

2

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

3

Amphipoda

Chironomidae

Marine fish

Marine Amphipoda

Number of BINs — Species per BIN

44, 36, 3, 1, 1, 1, 1
(x-axis: 1, 2, 3, 4, 5, 8, 49)

Chironomidae

Number of BINs — Species per BIN

325, 181, 25, 4, 2, 2, 1, 1, 1
(x-axis: 1, 2, 3, 4, 5, 6, 7, 11, 12)

Marine fish

Number of BINs — Species per BIN

1579, 1689, 501, 141, 80, 23, 21, 11, 7, 4, 4, 2, 6, 3, 3, 3, 1, 1, 2, 1, 1, 1, 1, 1, 1
(x-axis: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 22, 24, 26, 27, 30, 40, 88)