ARPHA Conference Abstracts 4: e64671
doi: 10.3897/aca.4.e64671

Conference Abstract

Assessing accuracy and completeness of GenBank for eDNA metabarcoding: towards a reliable marine fish reference database

Cristina Claver[‡], Oriol Canals[‡], Naiara Rodriguez-Ezpeleta[‡]

‡ AZTI,Marine Research, Basque Research and Technology Alliance, Sukarrieta, Spain

Corresponding author: Cristina Claver (cristinaclaver123@gmail.com)

Received: 19 Feb 2021 | Published: 04 Mar 2021

Citation: Claver C, Canals O, Rodriguez-Ezpeleta N (2021) Assessing accuracy and completeness of GenBank for eDNA metabarcoding: towards a reliable marine fish reference database. ARPHA Conference Abstracts 4: e64671. https://doi.org/10.3897/aca.4.e64671

Abstract

Environmental DNA (eDNA) metabarcoding, the process of sequencing DNA collected from the environment for producing biodiversity inventories, is increasingly being applied to assess fish diversity and distribution in marine environments. Yet, the successful application of this technique deeply relies on accurate and complete reference databases used for taxonomic assignment. The most used markers for fish eDNA metabarcoding studies are the cytochrome C oxidase subunit 1 (COI), 16S ribosomal RNA (16S), the 12S ribosomal RNA (12S) and cytochrome b (cyt b) genes, whose sequences are usually retrieved from GenBank, the largest DNA sequence database that represents a worldwide public resource for genetic studies. Thus, the completeness and accuracy of GenBank is critical to derive reliable estimations from fish eDNA metabarcoding data. Here, we have i) compiled the checklist of European marine fishes, ii) performed a gap analysis of the four genes and, within COI and 12S, also of the most used barcodes for fish, and iii) developed a workflow to detect potentially incorrect records in GenBank. We found that from the 1965 species in the checklist (1761 Actinopterygii, 189 Elasmobranchii, 9 Holocephali, 4 Petromyzonti and 2 Myxini), about 70% have sequences for COI, whereas less have sequences for 12S, 16S and cyt b (45-55%). Among the species for which COI ad 12S sequences are available, about 60% and 40% have sequences covering the most used barcodes respectively. The analysis of pairwise distances between sequences revealed pairs belonging to the same species with significantly low similarity and pairs belonging to different high level taxonomic groups (class, order) with significantly large similarity. In light of this further confirmation of presence of a substantial number of incorrect records in GenBank, we propose a method for identifying and removing spurious sequences to create reliable and accurate reference databases for eDNA metabarcoding.

Keywords

Environmental DNA, metabarcoding, DNA barcode reference, barcoding gap, European fish checklist

Presenting author

Cristina Claver

Presented at

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