Improving FAIRness of eDNA and Metabarcoding Data: Standards and tools for European Nucleotide Archive data deposition

Joana Paupério[‡], Vikas Gupta[‡], Josephine Burgin[‡], Suran Jayathilaka[‡], Jerry Lanfear[§], Kessy Abarenkov[|], Urmas Kõljalg[|], Lyubomir Penev[¶,#], Guy Cochrane[‡]

‡ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom
§ ELIXIR, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom
| University of Tartu Natural History Museum, Tartu, Estonia
¶ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
# Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria

Corresponding author: Joana Paupério (joanap@ebi.ac.uk)

Abstract

The advancements in sequencing technologies have promoted the generation of molecular data for cataloguing and describing biodiversity. The analysis of environmental DNA (eDNA) through the application of metabarcoding techniques enables comprehensive descriptions of communities and their function, being fundamental for understanding and preserving biodiversity. Metabarcoding is becoming widely used and standard methods are being generated for a growing range of applications with high scalability. The generated data can be made available in its unprocessed form, as raw data (the sequenced reads) or as interpreted data, including sets of sequences derived after bioinformatics processing (Amplicon Sequence Variants (ASVs) or Operational Taxonomic Units (OTUs)) and occurrence tables (tables that describe the occurrences and abundances of species or OTUs/ASVs). However, for this data to be Findable, Accessible, Interoperable and Reusable (FAIR), and therefore fully available for meaningful interpretation, it needs to be deposited in public repositories together with enriched sample metadata, protocols and analysis workflows (ten Hoopen et al. 2017).

Metabarcoding raw data and associated sample metadata is often stored and made available through the International Nucleotide Sequence Database Collaboration (INSDC) archives (Arita et al. 2020), of which the European Nucleotide Archive (ENA, Burgin et al. 2022) is its European database, but it is often deposited with minimal information, which hinders data reusability.

Within the scope of the Horizon 2020 project, Biodiversity Community Integrated Knowledge Library (BiCIKL), which is building a community of interconnected data for biodiversity research (Penev et al. 2022), we are working towards improving the

standards for molecular ecology data sharing, developing tools to facilitate data deposition and retrieval, and linking between data types.

Here we will present the ENA data model, showcasing how metabarcoding data can be shared, while providing enriched metadata, and how this data is linked with existing data in other research infrastructures in the biodiversity domain, such as the Global Biodiversity Information Facility (GBIF), where data is deposited following the guidelines published in Abarenkov et al. (2023). We will also present the results of our recent discussions on standards for this data type and discuss future plans towards continuing to improve data sharing and interoperability for molecular ecology.

Keywords

biodiversity, sequence data, metadata, deposition and retrieval, linked data

Presenting author

Joana Paupério

Presented at

TDWG 2023

Funding program

The BiCIKL project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Grant title

BiCIKL - Biodiversity Community Integrated Knowledge Library

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Abarenkov K, Andersson A, Bissett A, Fossøy F, Grosjean M, Hope M, Jeppesen TS, Kõljalg U, Lundin D, Nilsson H, Prager M, Schigel D, Suominen S, Svenningsen C,

Frøslev TG (2023) Publishing DNA-derived data through biodiversity data platforms. GBIF Secretariat v1.3 https://doi.org/10.35035/doc-vf1a-nr22

- Arita M, Karsch-Mizrachi I, Cochrane G (2020) The international nucleotide sequence database collaboration. Nucleic Acids Research 49: D121-D124. https://doi.org/10.1093/nar/gkaa967
- Burgin J, Ahamed A, Cummins C, Devraj R, Gueye K, Gupta D, Gupta V, Haseeb M, Ihsan M, Ivanov E, Jayathilaka S, Balavenkataraman Kadhirvelu V, Kumar M, Lathi A, Leinonen R, Mansurova M, McKinnon J, O'Cathail C, Paupério J, Pesant S, Rahman N, Rinck G, Selvakumar S, Suman S, Vijayaraja S, Waheed Z, Woollard P, Yuan D, Zyoud A, Burdett T, Cochrane G (2022) The European Nucleotide Archive in 2022. Nucleic Acids Research 51: D121-D125. https://doi.org/10.1093/nar/gkac1051
- Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin C, Barov B, Demirova I, Hristova K (2022) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8: e81136. https://doi.org/10.3897/rio.8.e81136
- ten Hoopen P, Finn R, Bongo LA, Corre E, Fosso B, Meyer F, Mitchell A, Pelletier E, Pesole G, Santamaria M, Willassen NP, Cochrane G (2017) The metagenomic data life-cycle: standards and best practices. GigaScience 6 (8): 1-11. https://doi.org/10.1093/gigascience/gix047