Third-party Annotations: Linking PlutoF platform and the ELIXIR Contextual Data ClearingHouse for the reporting of source material annotation gaps and inaccuracies

Kessy Abarenkov[‡], Allan Zirk[‡], Kadri Põldmaa[§], Timo Piirmann[‡], Raivo Pöhönen[‡], Filipp Ivanov[‡], Kristjan Adojaan[§], Urmas Kõljalg[§]

‡ University of Tartu Natural History Museum, Tartu, Estonia
§ University of Tartu, Tartu, Estonia

Corresponding author: Kessy Abarenkov (kessy.abarenkov@ut.ee)

Abstract

Third-party annotations are a valuable resource to improve the quality of public DNA sequences. For example, sequences in International Nucleotide Sequence Databases Collaboration (INSDC) often lack important features like taxon interactions, species level identification, information associated with habitat, locality, country, coordinates, etc. Therefore, initiatives to mine additional information from publications and link to the public DNA sequences have become common practice (e.g. Tedersoo et al. 2011, Nilsson et al. 2014, Groom et al. 2021). However, third-party annotations have their own specific challenges. For example, annotations can be inaccurate and therefore must be open for permanent data management. Further, every DNA sequence (except sequences from type material) can carry different species names, which must be databased as equal scientific hypotheses. PlutoF platform provides such data management services for third-party annotations.

PlutoF is an online data management platform and computing service provider for biology and related disciplines. Registered users can enter and manage a wide range of data, e.g., taxon occurrences, metabarcoding data, taxon classifications, traits, and lab data. It also features an annotation module where third-party annotations (on material source, geolocation and habitat, taxonomic identifications, interacting taxa, etc.) can be added to any collection specimen, living culture or DNA sequence record. The UNITE Community is using these services to annotate and improve the quality of INSDC rDNA Internal Transcribed Spacer (ITS) sequence datasets. The National Center for Biotechnology Information (NCBI) is linking its ITS sequences with their annotations in PlutoF. However, there is still missing an automated solution for linking annotations in PlutoF with any sequence and sample record stored in INSDC databases. One of the ambitions of the BiCI

KL Project is to solve this through operating the ELIXIR Contextual Data ClearingHouse (CDCH). CDCH offers a light and simple RESTful Application Programming Interface (API) to enable extension, correction and improvement of publicly available annotations on sample and sequence records available in ELIXIR data resources. It facilitates feeding improved or corrected annotations from databases (such as secondary databases, e.g., PlutoF, which consume and curate data from repositories) back to primary repositories (databases of the three INSDC collaborative partners).

In the Biodiversity Community Integrated Knowledge Library (BiCIKL) Project, the University of Tartu Natural History Museum is leading the task of linking the two components—the web interface provided by the PlutoF platform and CDCH APIs—to allow user-friendly and effortless reporting of errors and gaps in sequenced material source annotations. The API and web interface will be promoted to those communities (such as taxonomists, those abstracting from the literature, and those already using the community curated data) with the appropriate knowledge and tools who will be encouraged to report their enhanced annotations back to primary repositories.

Keywords

annotating DNA sequences, data management, linking data, BiCIKL

Presenting author

Kessy Abarenkov

Presented at

TDWG 2021

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

References

- Groom QJ, Dillen M, Huybrechts P, Johaadien R, Kyriakopoulou N, Fernandez FJQ, Trekels M, Wong WY (2021) Connecting molecular sequences to their voucher specimens. BioHackrXiv Preprints https://doi.org/10.37044/osf.io/93qf4
- Nilsson RH, Hyde K, Pawłowska J, Ryberg M, Tedersoo L, Aas AB, Alias S, Alves A, Anderson CL, Antonelli A, Arnold AE, Bahnmann B, Bahram M, Bengtsson-Palme J, Berlin A, Branco S, Chomnunti P, Dissanayake A, Drenkhan R, Friberg H, Frøslev TG, Halwachs B, Hartmann M, Henricot B, Jayawardena R, Jumpponen A, Kauserud H, Koskela S, Kulik T, Liimatainen K, Lindahl B, Lindner D, Liu J, Maharachchikumbura S, Manamgoda D, Martinsson S, Neves MA, Niskanen T, Nylinder S, Pereira OL, Pinho DB, Porter T, Queloz V, Riit T, Sánchez-García M, de Sousa F, Stefańczyk E, Tadych M, Takamatsu S, Tian Q, Udayanga D, Unterseher M, Wang Z, Wikee S, Yan J, Larsson E, Larsson K, Kõljalg U, Abarenkov K (2014) Improving ITS sequence data for identification of plant pathogenic fungi. Fungal Diversity 67 (1): 11-19. https://doi.org/10.1007/s13225-014-0291-8
- Tedersoo L, Abarenkov K, Nilsson RH, Schüssler A, Grelet G, Kohout P, Oja J, Bonito G, Veldre V, Jairus T, Ryberg M, Larsson K, Kõljalg U (2011) Tidying Up International Nucleotide Sequence Databases: Ecological, Geographical and Sequence Quality Annotation of ITS Sequences of Mycorrhizal Fungi. PLoS ONE 6 (9). https://doi.org/10.1371/journal.pone.0024940