Bridging a Gap in Metabarcoding Research: The ASV Table Registry

Christian Bräunig[‡], Björn Quast[‡], Peter Grobe[‡]

‡ Leibniz Institute for the Analysis of Biodiversity Change, Zoological Research Museum Koenig, Bonn, Germany

Corresponding author: Björn Quast (b.quast@leibniz-lib.de)

Abstract

Metabarcoding is a tool to routinely identify species in environmental mass-samples and thereby analyze their species composition. Using metabarcoding techniques outperforms the traditional species identification by human experts in amount, speed and quality when well curated reference data are available.

Therefore, metabarcoding can be seen as the future standard method for all biological research areas where species occurrence and distribution is in question, e.g., ecological research or monitoring projects (Porter and Hajibabaei 2018).

A common outcome of metabarcoding research are Amplicon Sequence Variant tables (ASV, Callahan et al. 2017). These tables combine the extracted sequences of all sampling plots with the occurrences of each sequence within a single plot. To identify the species, each sequence is searched in one or more reference databases that hold sequences and their known taxon identifications (e.g., Barcode Of Life Data system (BOLD) or the German Barcode of Life library (GBOL)). The sequence searches utilise tools like BLAST, BOLD identification engine, or vsearch. Found taxa and their taxonomy are added to the ASV tables as taxon assignments.

The number and precision of taxon assignments will increase with the growth of available sequences and quality of identifications in reference databases over time (Weigand et al. 2019). The introduction of new marker sequences and improvements in search tools will further enhance the taxon assignments. Thus, the taxon assignments in ASV tables are subject to change.

Projects with the aim of building up species inventories on a large scale (GBOL) or monitoring programs, like the Automated Multisensor Stations for Monitoring of BioDiversity (Wägele et al. 2022), quickly produce data sets with thousands of sequences at numerous locations.

Currently, most ASV tables are stored as supplements to publications or in private repositories. This makes analysis across multiple research projects difficult and error prone

as sequences and their taxon assignments are often not accessible. Efforts, like the European Bioinformatics Institute metagenomics with Mgnify serve the needs for uploading and annotating environmental DNA samples (Mitchell et al. 2017), but a registry for ASV tables with complete data life cycles is lacking.

To fill this gap, we develop an ASV Table Registry as part of the German Barcode of Life III - Dark Taxa project. This allows users to:

- register ASV tables and sequences
- upload and manage ASV tables with versioning
- publish ASV tables with DOIs
- search by sequences, taxa, and occurrence data
- retrieve API-based data
- assign taxonomic names with various tools and reference databases
- keep track of the applied search methods and parameters

The data life cycle of the uploaded ASV tables consists of several draft versions (each re-annotation with the identification pipeline creates a new draft version) and eventually a published version with a DOI. New draft versions can be created from the published version, then re-annotated and published again. The tracking of former taxon assignments allows researchers to re-evaluate data of former studies, compare them, and add new results. The ASV Table Registry developed here aims to make ASV tables FAIR (Findable, Accessible, Interoperable, and Reusable) and to foster the shared use in research projects.

Future development focuses on the incorporation of the MIxS standard (Yilmaz et al. 2011) and on submission of the published data to International Nucleotide Sequence Database Collaboration (INSDC) using established dataflows from the German Federation for Biological Data (GFBio) and NFDI4biodiversity.

The ASV data portal is accessible at: https://bolgermany.de/metabarcoding; the source code at: https://gitlab.leibniz-lib.de/GBOL/asv-table-registry.

Keywords

taxon annotation, environmental research, biodiversity monitoring, barcode reference databases, DNA barcodes, FAIR principles

Presenting author

Christian Bräunig

Presented at

TDWG 2022

Funding program

The ASV Table Registry is developed within the BMBF funded Project GBOL III - Dark Taxa, German Federal Ministry of Education and Research (BMBF, grant ID: 01LI1901)

Conflicts of interest

References

- Callahan BJ, McMurdie PJ, Holmes SP, et al. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11 (12): 2639-2643. https://doi.org/10.1038/ismej.2017.119
- Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S, Boland MA, Hunter F, ten Hoopen P, Alako B, Amid C, Wilkinson DJ, Curtis TP, Cochrane G, Finn RD (2017) EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Research 46 https://doi.org/10.1093/nar/gkx967
- Porter T, Hajibabaei M (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Molecular Ecology 27 (2): 313-338. https://doi.org/10.1111/mec.14478
- Wägele JW, Bodesheim P, Bourlat S, Denzler J, Diepenbroek M, Fonseca V, Frommolt K, Geiger M, Gemeinholzer B, Glöckner FO, Haucke T, Kirse AK, Kölpin A, Kostadinov I, Kühl H, Kurth F, Lasseck M, Liedke S, Losch F, Müller S, Petrovskaya N, Piotrowski K, Radig B, Scherber C, Reinhold L, Schulz J, Steinhage V, Tschan GF, Vautz W, Velotto D, Weigend M, Wildermann S (2022) Towards a multisensor station for automated biodiversity monitoring. Elsevier https://doi.org/10.15480/882.4141
- Weigand H, Beermann A, Čiampor F, Costa F, Csabai Z, Duarte S, Geiger M, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand A, Willassen E, Wyler S, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn B, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher J, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price B, Rinkevich B, Teixeira ML, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of The Total Environment 678: 499-524. https://doi.org/10.1016/j.scitotenv.2019.04.247
- Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ,

Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PSG, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, Gil IS, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spor A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glöckner FO, et al. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29 (5): 415-420. https://doi.org/10.1038/nbt.1823