ARPHA Preprints

PENSOFT

PREPRINT

Author-formatted, not peer-reviewed document posted on 17/04/2024

DOI: https://doi.org/10.3897/arphapreprints.e125475

The ASV registry: a place for ASVs to be

Christian Bräunig, Sandra Meid, Björn Quast, Vera Rduch, Peter Grobe

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 17/04/2024* DOI: https://doi.org/10.3897/arphapreprints.e125475

The ASV registry: a place for ASVs to be

Christian Bräunig, Sandra Meid, Björn Quast, Vera Rduch, Peter Grobe

Abstract

Despite the effectiveness of DNA metabarcoding for gaining insights into biodiversity and environmental species composition, a centralized management and storage option including easy accessibility of already published data is lacking. Since most data is published as supplementary material or in private repositories, DNA metabarcoding has a huge untapped potential to be used for analysis across multiple taxa, sample locations or multiple research projects. We developed a platform to register, manage and identify amplicon sequence variants (ASVs) or zero-radius OTUs (ZOTUs), respectively, against several barcode reference datasets. Moreover, ASV tables can be uploaded, managed, versioned, and published with DOIs thus contributing to the full research Data Life Cycle.

Keywords

amplicon sequence variant, barcode reference databases, biodiversity monitoring, data storage, DNA metabarcoding, operational taxonomic unit, taxon annotation

Introduction

DNA metabarcoding allows for the routine identification of species and analysis of species composition in environmental mass samples. If well curated reference data are available, this method can yield comparable results or even outperform conventional species identification by human experts in quantity, quality and speed (Elbrecht et al. 2017; Remmel et al. 2024). This is also why DNA metabarcoding is not only an important tool for science, but is also increasingly used by environmental authorities (Leese et al. 2023). Hence, it is reasonable to consider it a future standard method in any biological research where species occurrence and distribution are of interest, such as ecological research and biomonitoring (Porter and Hajibabaei 2018).

The sequencing reads produced by metabarcoding methods usually have two possible fates: individual reads may be clustered based on sequence similarity, thereby generating operational taxonomic units (OTUs; Sokal 1963). Alternatively, single sequences are extracted and quality assured from multiple reads by bioinformatics tools and become amplicon sequence variants (ASVs; Callahan et al. 2017). As single DNA sequences, they offer a finer resolution and a higher degree of comparability across datasets (Callahan et al. 2017).

In case multiple mass samples were analyzed simultaneously, organizing the found sequences in so-called ASV tables is reasonable (Callahan et al. 2017). These tables combine the extracted sequences of all sampling plots with the number of occurrences of each sequence within the respective plots. They can be enriched by taxonomic

identification of the individual sequences, often including a taxonomy of the identified taxon. The production of these taxonomic assignments is the step at which the species composition of the samples is revealed. This can be done using tools like BLAST, the Barcode Of Life Data (BOLD; Ratnasingham and Hebert 2007) system identification engine, or vsearch against DNA barcode reference databases like the BOLD database or the German Barcode of Life library (GBOL; German Barcode of Life Consortium 2011). Some of the analyzed sequences will have no species or genus name associated, either because they stem from DNA of specimens that belong to species not contained in the reference database, or they belong to distinct populations that are genetically different to the reference material of the same species in the database. The resolution of species detection hinges on the quality and coverage of the used reference libraries (Macher et al. 2021). Since the number of available sequences and the quality of identifications in reference databases are growing over time (Weigand et al. 2019), the number and precision of taxon assignments will increase continuously. The introduction of new marker sequences and improvements in search tools will further lead to improved taxon assignments. Thus, the taxon assignments in ASV tables are subject to constant change. Repeated taxon identification of a given set of ASV sequences is hence not only worthwhile to make use of this ongoing growth and improvement, but necessary to keep the identifications up to date and as informative as possible.

While data acquisition with metabarcoding is boosted, most ASV tables are still stored as supplements to publications or in private repositories. This greatly impairs access and reuse of these data. As a result, analyses across multiple research projects are difficult, error prone and often not up-to-date. Efforts, like the EBI metagenomics with MGnify (Richardson et al. 2023) serve the needs for uploading and annotating environmental DNA samples (Mitchell et al. 2017), but there is no registry for ASV tables in which they can be sustainably stored, managed, versioned and published in a well-documented manner.

The ASV Registry presented here contributes to the completion of the Research Data Life Cycle (Rüegg et al. 2014; Griffin et al. 2017) and makes ASV data available to open science under the FAIR (Wilkinson et al. 2016) criteria.

Implementation

The ASV registry

To address the need for centralized archival and management of ASV data, we have developed the ASV registry. This platform allows users to archive, analyze, manage and retrieve ASV data in consideration of FAIR data principles (Wilkinson et al. 2016). We have designed the ASV registry as a full-fledged web application to impose minimal software requirements on users. It is reachable via any browser at https://asv.bolgermany.de/metabarcoding and full usage only requires creating an account. The services provided are entirely free of charge.

Technically, the web platform is based on the Pyramid web framework (https://trypyramid.com/) and sits atop a MySQL database (https://www.mysql.com/). An ElasticSearch indexer (https://www.elastic.co/elasticsearch) manages search and retrieval processes.

Data life cycle and key features

The ASV registry supports a specific life cycle for ASV tables to facilitate effective and efficient usage of the contained data. Before the upload of any data, users must first create or be added to a so-called Project container (Fig. 1). Analogous to real projects, these act as organizational units to group uploaded data and support collaborative work, while also handling access management. Only members of a given project can upload, analyze, delete and publish the data within it. Project creators can dynamically add and remove members.

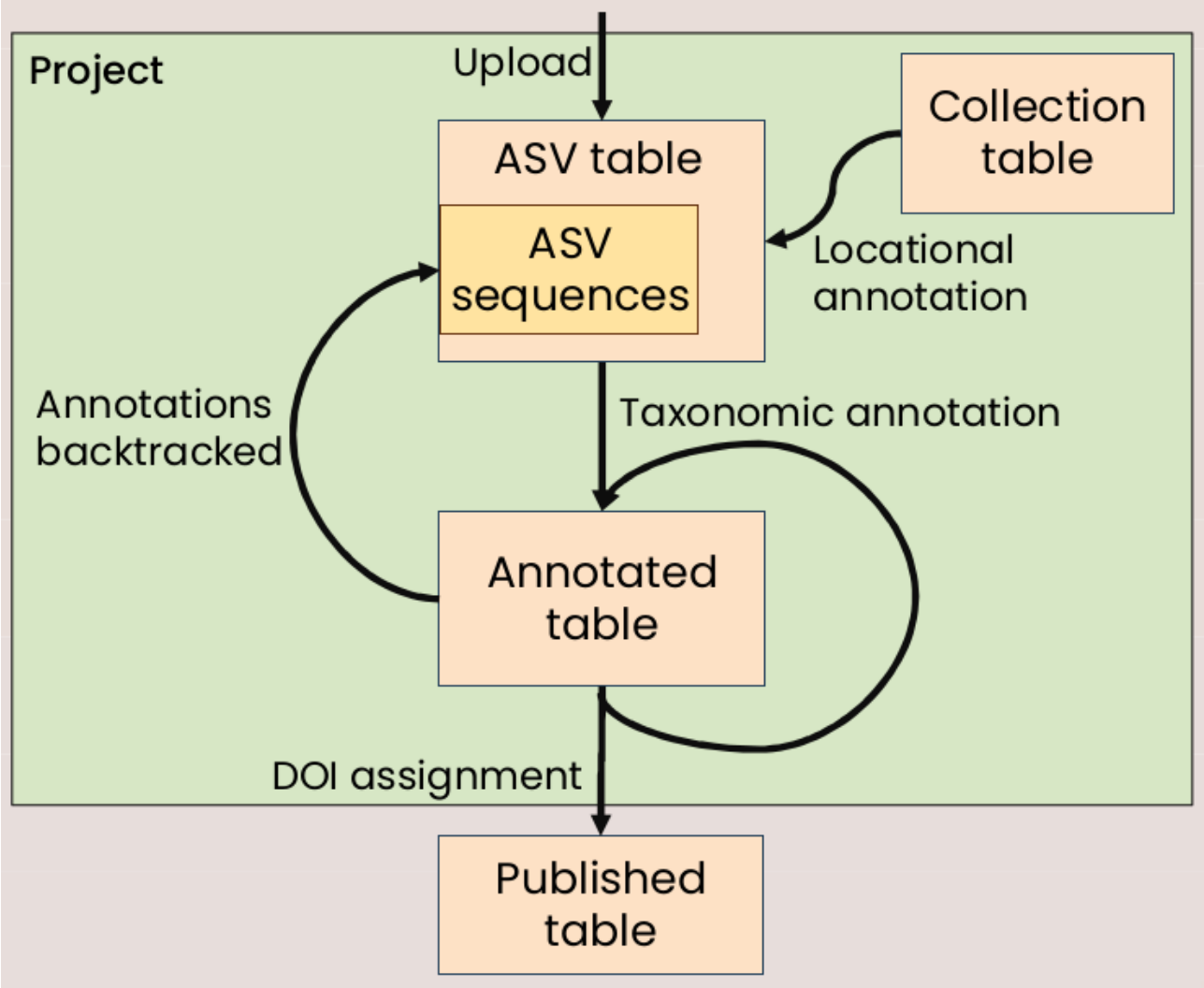

Figure 1. ASV table specific data life cycle. The Project container (green) acts as collaborative work space and access management. Publication of a table via DOI assignment moves the table out of this container and makes it publicly available.

For upload, ASV tables must adhere to a basic format (Fig. 2). Further information can be enclosed in the table, such as collection metadata, but these are not considered during initial upload.

esv_id	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152440	1	3	1	0	1	0	TCTATCTTCATACTTATTTCATTCTTCTCCATCAA
uniq152443	1	0	3	0	1	1	TTTATCATCTAGAATTGCTCATGGAGGAGCTTCC
uniq152454	3	1	0	0	2	0	ACTTTCATCTAATATTGCCCATGGAGGTAGTTCT
uniq152461	2	0	3	0	0	1	ACTATCAAATAATATATCCCATGCAGGAGCATCA
uniq152464	1	3	2	0	0	0	ACTTGCAGGAGCTATTGCTCATGGTGGAGGATC
uniq152471	1	1	3	1	0	0	ATTGTCGAACAATATATTTCACAGAGGTGCCTCA
uniq152480	0	3	1	0	1	1	ACTATCTGCTAATATTGCTCATAGAGGAGCATCA
uniq152481	0	3	1	1	1	0	TCTATCCTCAAATATTGCACATAACGGTTCATCT
uniq152482	0	5	0	0	1	0	ACTTGCAGGAGCTATTGCTCATGGTGGGTGATC
uniq152483	0	2	2	1	1	0	CCTATCTTCAGGAATTGCACATGGAGGTGCTTCT

Figure 2. Initial data format. ASV tables that are to be uploaded must at the very least contain the following information: the DNA sequence of each ASV (green) and the integer-based read count of each ASV per sampling plot (blue), where each column corresponds to one sample. User-defined identifiers for the ASVs (orange) can also be included. These will be saved in the database, but will not be used as identifiers during later processing.

Upon upload of an ASV table, it receives a unique persistent identifier, in the form of a simple integer, by which it will be findable at any later point. The original file will be saved, but also all individual data points are written to the database. Each individual sequence within the table will also be saved to the database as part of a so-called set. A set consists of the sequence and the sequencing primer pair selected during the table upload, and is assigned a unique identifier in the format

`GBOL_ASV_ID_X`

where X corresponds to a continuously incremented integer (Fig. 3). The combination of sequence data and corresponding sequencing primers allows for more precise data points while limiting redundancy.

esv_id	ASV_Set_IDs	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152440	GBOL_ASV_ID_97	1	3	1	0	1	0	TCTATCTTCATACTTATTT
uniq152443	GBOL_ASV_ID_41385	1	0	3	0	1	1	TTTATCATCTAGAATTGCT
uniq152454	GBOL_ASV_ID_41390	3	1	0	0	2	0	ACTTTCATCTAATATTGCC
uniq152461	GBOL_ASV_ID_41391	2	0	3	0	0	1	ACTATCAAATAATATATCC
uniq152464	GBOL_ASV_ID_41392	1	3	2	0	0	0	ACTTGCAGGAGCTATTGC
uniq152471	GBOL_ASV_ID_41388	1	1	3	1	0	0	ATTGTCGAACAATATATTT
uniq152480	GBOL_ASV_ID_41389	0	3	1	0	1	1	ACTATCTGCTAATATTGCT
uniq152481	GBOL_ASV_ID_41384	0	3	1	1	1	0	TCTATCCTCAAATATTGCA
uniq152482	GBOL_ASV_ID_41387	0	5	0	0	1	0	ACTTGCAGGAGCTATTGC
uniq152483	GBOL_ASV_ID_41386	0	2	2	1	1	0	CCTATCTTCAGGAATTGC

Figure 3. Assignment of persistent identifiers during upload. Each individual ASV uploaded as part of an ASV table receives a unique, persistent identifier (yellow; see Figure 2 for the other color codes). Notably, the unique combination of sequence and primers selected during table upload constitute a new ASV. This approach to ID assignment greatly limits redundancy within the database and helps to make recurring amplicons evident.

After the initial upload, the table enters the dedicated ASV Table life cycle and becomes enrichable in various ways: Firstly, and most importantly, the sequences in the table can receive taxonomic assignments. At the moment, the registry provides thirteen reference databases for this purpose, including multiple SILVA, NCBI and PLANiTS datasets respectively. Users can set various parameters for this process and also select whether the chosen databases should be queried in parallel or sequentially. Numerous assignment data are routinely stored after each annotation run: found taxon, taxonomy of the found taxon, specimen id and link specific to the used reference database, E-value and percent

identity. All of these data are backtracked to the individual ASV sequences. As the assignments are subject to change over time, repeated taxonomic annotation is supported by the registry, thereby allowing users to keep their assignments up to date. Any assignments found will be attached to both tables and individual sequences within the database. Methods and parameters of the respective assignment runs are tracked. Secondly, by way of uploading so-called collection tables, users can add metadata regarding the collecting conditions for each sampling plot in the table (Fig. 1). Metadata may include but is not limited to coordinates, locality descriptions and temperature during collection.

At any point after the upload, but most sensibly after any number of taxonomic annotation runs, a table can be downloaded. Users can tailor which annotation data to include and where, and exclude data of no interest (Fig. 4). This way, taxonomy data from several runs and over time can be readily compared in the downloaded file (Fig. 5). Users can also choose in which format to download the table to accommodate potential downstream applications.

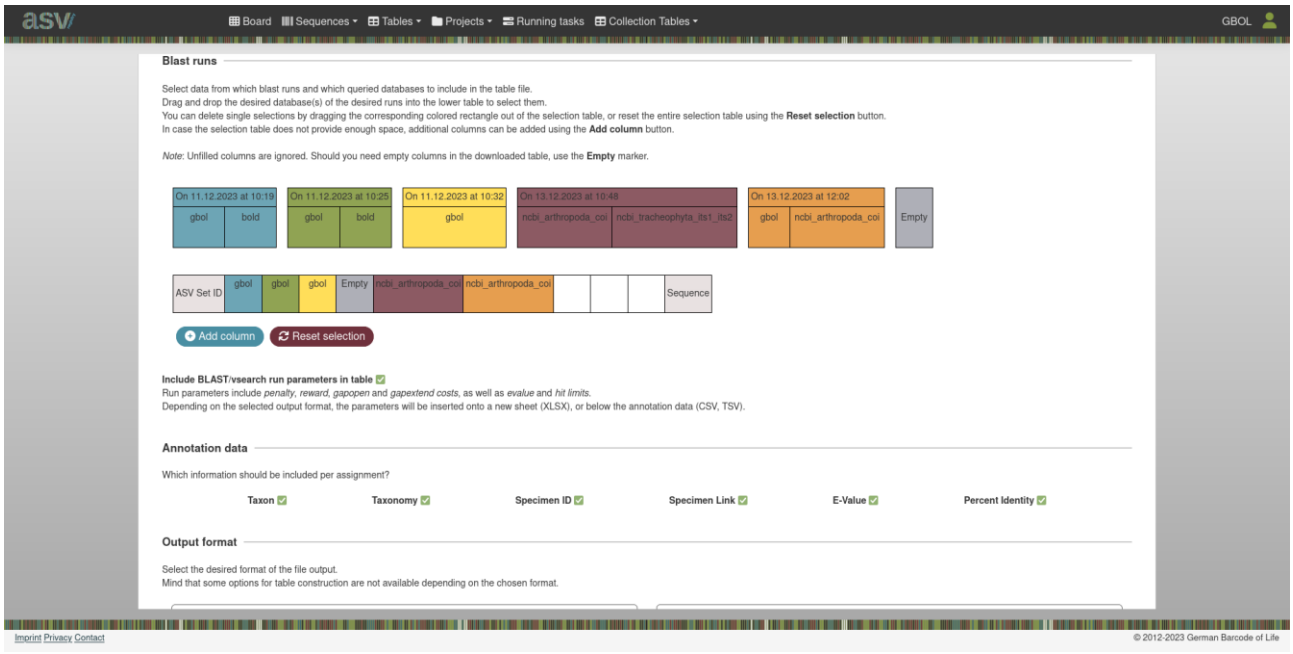

Figure 4. Table download options. Users have numerous options for tailoring downloaded tables to their needs. It is possible to choose queried databases from selected annotation runs, respectively. The order in which annotation runs appear and the data points to be shown for each annotation is selectable, too. The file format for the output file can also be chosen.

		Blast run 69 from 20240124-12h51m GBOL		Blast run 69 from 20240124-12h51m BOLD							
esv_id	ASV_Set_IDs	Taxon Name	Percent Identity	Taxon Name	Percent Identity	PCRPCR1C	PCRPCR1a	PCRPCR1b	PCRPCR2C	PCRPCR2a	PCRPCR2b	Seq
uniq152440	GBOL_ASV_ID_97	Bombus lapidarius	100	Bombus lapidarius atlanticus	100	1	3	1	0	1	0	TCTATC
uniq152443	GBOL_ASV_ID_41385	Empis tessellata	100	Empis tessellata	100	1	0	3	0	1	1	TTTATC
uniq152454	GBOL_ASV_ID_41390	Axylia putris	100	Axylia putris	100	3	1	0	0	2	0	ACTTTC
uniq152461	GBOL_ASV_ID_41391	Macrophya diversipes	100	Macrophya diversipes	100	2	0	3	0	0	1	ACTATC
uniq152464	GBOL_ASV_ID_41392			Omocestus petraeus	99	1	3	2	0	0	0	ACTTGC
uniq152471	GBOL_ASV_ID_41388	Gastrodes abietum	100			1	1	3	1	0	0	ATTGTC
uniq152480	GBOL_ASV_ID_41389	Himacerus apterus	100	Himacerus apterus	99	0	3	1	0	1	1	ACTATC
uniq152481	GBOL_ASV_ID_41384	Attelabus nitens	100			0	3	1	1	1	0	TCTATC
uniq152482	GBOL_ASV_ID_41387					0	5	0	0	1	0	ACTTGC
uniq152483	GBOL_ASV_ID_41386	Nemopoda nitidula	100	Nemopoda nitidula	100	0	2	2	1	1	0	CCTATC

Figure 5. Taxonomic annotation. During the taxonomic annotation process various assignment data are recorded and saved. Users can then decide which of these data to include, in which order to arrange data from individual annotation processes and where within the final table to include these data. This allows the final ASV table to be maximally informative without becoming chaotic and hard to read. In this case, two separate databases have been queried (pink, purple), and in each case the found taxa and percentages of identity have been added to the downloaded table file. For the other color codings see Figure 2.

Once satisfied with the state of a table, users can then choose to publish their table. First, a publication draft is created for such a table and a preliminary DOI is assigned that can already be used to point to the table within the registry. After adding publication-relevant metadata, the table can be published, whereby the preliminary DOI is made final. This moves it out of the project container and access is thus no longer limited to project members (Fig. 1). Instead, it becomes available to all users – logged in or not – to view, download and use.

The ASV registry provides a number of methods to search through and retrieve project-internal and published data. These include, but are not limited to, filtering sequences by taxon assignment, tables and sequences by project, sequences by location in case collection tables were uploaded (Fig. 6).

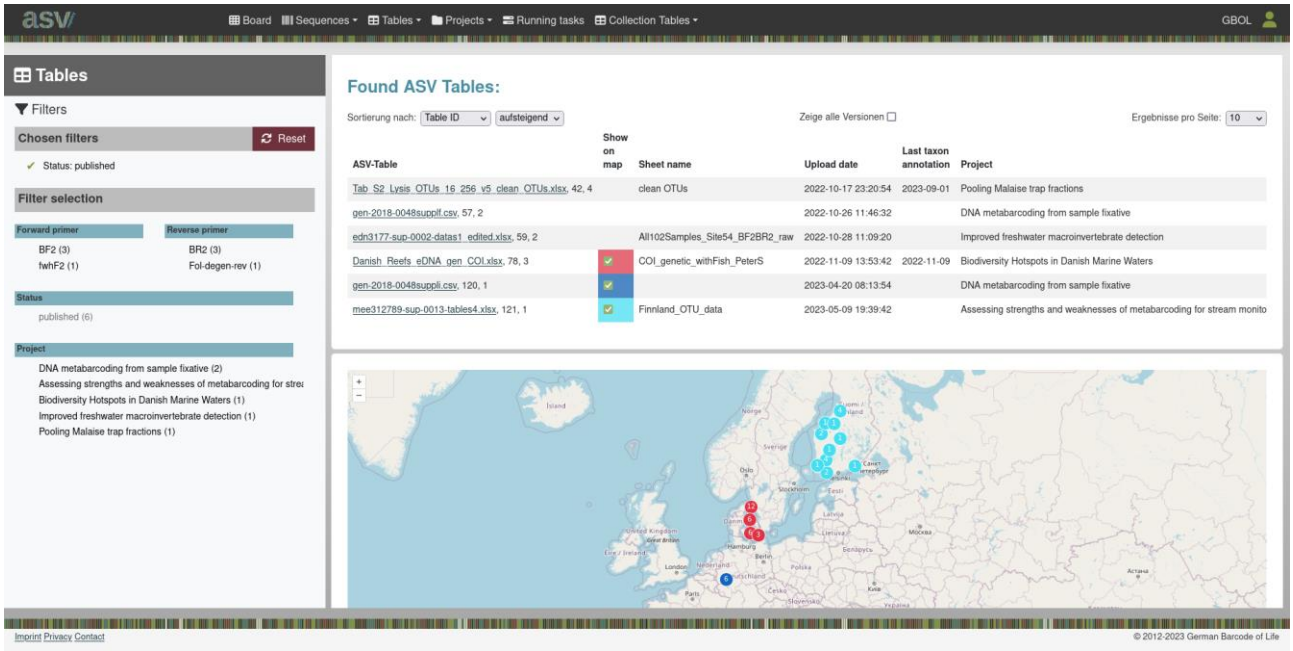

Figure 6. Table search page. Published tables and not yet published tables from the signed-in user's projects can be easily searched and retrieved on the dedicated web page. Aside from filters and a map, an option for keyword search featuring different modes and input suggestions also exists.

Conclusions

The ASV registry, the web application we present here, aims to pave the way for efficient and effective use of ASV data, which is rarely possible at the moment. By providing a platform for researchers to archive, analyze and ultimately publish their ASV data, we hope to foster comparative and collaborative DNA metabarcoding research and biodiversity monitoring in the future. Easier access to published data and structured publication of new data will certainly boost the collective usefulness of DNA metabarcoding.

Acknowledgements

This research was funded by the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung, BMBF) as part of the project “GBOL III: Dark Taxa” as Research for Sustainable Development (Forschung für Nachhaltige Entwicklung, FONA; www.fona.de) under the funding reference 16LI1901A.

Further, we are grateful to Dr. Ralph Peters and all other members of the GBOL III: Dark Taxa project at the Leibniz-Institute for the Analysis of Biodiversity Change for their ongoing support during code development and testing.

We would also like to thank Dr. Tamara Hartke, Dr. Vera Zizka, Vera Prenzel and Hadeel Ragab for kindly allowing us to use data from their projects for illustrative purposes.

Project description

The presented software has been developed within a dedicated work package of the GBOL III: Dark Taxa project. The registry builds upon and makes use of the work done in the previous two funding phases of the German Barcode of Life (GBOL) project to establish a barcode library for the flora, fauna and fungi of Germany, while supporting the current research to investigate so-called dark taxa – species rich but so far understudied groups of insects (Hausmann et al. 2020; Rduch and Peters 2020).

Study area description: Metabarcoding, metagenomics and bioinformatics

For further information, see https://gbol.bolgermany.de/gbol3/de/gbol-dark-taxa/.

Web location (URIs)

The ASV data portal is available at https://asv.bolgermany.de/metabarcoding.

The source code is available at https://gitlab.leibniz-lib.de/GBOL/asv-registry.

Technical specification

Platform: web application

Programming language: Python, Javascript, MySQL

Operational system: various

Usage rights

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

Elbrecht V, Vamos EE, Meissner K, Aroviita J and Leese F (2017), Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods Ecol Evol, 8: 1265-1275. https://doi.org/10.1111/2041-210X.12789

German Barcode of Life Consortium (Wägele W, Haszprunar G, Eder J, Xylander W, Borsch T, Quandt D, Grobe P, Pietsch S, Geiger M, Astrin J, Rulik B, Hausmann A, Moriniere J, Holstein J, Krogmann L, Monje C, Traunspurger W, Hohberg K, Lehmitz R, Müller K, Nebel M and Hand R) (2011), GBOL Webportal at https://www.bolgermany.de. [Dataset]. Version: 20170316. Data Publisher: Zoological Research Museum Koenig - Leibniz Institute for Animal Biodiversity.

Griffin PC, Khadake J, LeMay KS, Lewis SE, Orchard S, Pask A, Pope B, Roessner U, Russell K, Seemann T, Treloar A, Tyagi S, Christiansen JH, Dayalan S, Gladman S, Hangartner SB, Hayden HL, Ho WWH, Keeble-Gagnère G, Korhonen PK, Neish P, Prestes PR, Richardson MF, Watson-Haigh NS, Wyres KL, Young ND and Schneider MV (2017), Best practice data life cycle approaches for the life sciences. *F1000Research*, *6*, 1618. https://doi.org/10.12688/f1000research.12344.2

Hausmann A, Krogmann L, Peters RS, Rduch V and Schmidt S (2020), GBOL III: Dark Taxa. iBOL Barcode Bulletin 10 (1). https://doi.org/10.21083/ibol.v10i1.6242

Leese F, Woppowa L, Bálint M, Höss S, Krehenwinkel H, Lötters S, Meissner K, Nowak C, Rausch P, Rduch V, Rulik B, Weigand AM, Zimmermann J, Koschorrek J and Zughart W (2023), DNA-basierte Biodiversitätsanalysen im Natur- und Umweltschutz: Welche Optionen haben wir für eine Standardisierung? Eine Handlungsempfehlung aus Forschung & Praxis. BfN Schriften, 666. https://doi.org/10.19217/skr666

Macher TH, Beermann AJ and Leese F (2021), TaxonTableTools: A comprehensive, platform-independent graphical user interface software to explore and visualise DNA

metabarcoding data. Mol Ecol Resour, 21: 1705-1714. https://doi.org/10.1111/1755-0998.13358

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 and 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 and 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

Ratnasingham S and Hebert PDN (2007), BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes, 7: 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rduch V and Peters RS (2020), GBOL III: Dark Taxa – die dritte Phase der German Barcode of Life Initiative hat begonnen. Koenigiana 14 (2): 91-107.

Remmel N, Buchner D, Enss J, Hartung V, Leese F, Welti EAR et al. (2024) DNA metabarcoding and morphological identification reveal similar richness, taxonomic composition and body size patterns among flying insect communities. *Insect Conservation and Diversity*, 1–15. Available from: https://doi.org/10.1111/icad.12710

Richardson LJ, Allen B, Baldi G, Beracochea M, Bileschi M, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell L, Curtis T, Escobar-Zepeda A, Gurbich T, Kale V, Korobeynikov A, Raj S, Rogers AB, Sakharova E, Sanchez S, Wilkinson D and Finn RD (2023), MGnify: the microbiome sequence data analysis resource. Nucleic Acids Research 51: D753–D759. https://doi.org/10.1093/nar/gkac1080

Rüegg J, Gries C, Bond-Lamberty B, Bowen GJ, Felzer BS, McIntyre NE, Soranno PA, Vanderbilt KL and Weathers KC (2014), Completing the data life cycle: using information management in macrosystems ecology research. Frontiers in Ecology and the Environment, 12: 24-30. https://doi.org/10.1890/120375

Sokal RR (1963), The Principles and Practice of Numerical Taxonomy. *Taxon*, *12*(5): 190–199. https://doi.org/10.2307/1217562

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 and 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

Wilkinson M, Dumontier M, Aalbersberg I, et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Supplementary material

None