BOLDistilled: Automated construction of comprehensive but compact DNA barcode reference libraries

Prosser SWJ*, Floyd RM, Thompson KA, and Hebert PDN

Abstract

Advances in DNA sequencing technology have stimulated the rapid uptake of protocols—such as eDNA analysis and metabarcoding—that infer the species composition of environmental samples from DNA sequences. DNA barcode reference libraries play a critical role in the interpretation of sequences gathered through such protocols, but many lack adequate taxonomic curation, include redundant records, do not support end-user analytical pipelines, and are not permanently archived in repositories. Furthermore, because DNA sequencers are outpacing Moore's Law and reference libraries are rapidly expanding, the computational power required to assign sequences to source taxa increases yearly. To address these limitations while also providing access to anonymized private data from the Barcode of Life Data System (BOLD), we introduce an algorithmic approach to construct DNA barcode reference libraries that overcome the above issues. Hosted online, 'BOLDistilled' libraries are comprehensive but compact, because the algorithm distills genetic variation into a minimal set of records. We generated a BOLDistilled library for the barcode region of the cytochrome *c* oxidase 1 gene (COI) based on all data in BOLD. This library contains 1.2M records versus 17.5M in the complete library, a compression which reduced the time required for sequence analysis of metabarcoded samples by ≥98% with no reduction in the accuracy of taxonomic placements. BOLDistilled libraries will be updated routinely, with the current version and all previous versions available at boldsystems.org/BOLDistilled. By providing access to persistent, comprehensive, and high-quality reference data, BOLDistilled libraries will strengthen the capacity of DNA-based identification systems to advance biodiversity science.

*sprosser@uoguelph.ca

Introduction

DNA-based specimen identifications are now the global standard. Both basic science (Pringle et al., 2019) and applied projects (Aylagas et al., 2016) rely on the ability to infer taxonomy from short DNA sequences. Three underpinning analytical protocols—DNA barcoding, metabarcoding, and eDNA analysis—all rely on the PCR amplification and sequencing of targeted genetic markers. Resulting sequences are identified by comparison with a reference library comprised of sequences with associated taxonomic assignments. Therefore, well-validated reference libraries are critical for DNA-based species identification (Ahmed et al., 2019; Liu et al., 2020; Rimet et al., 2021).

The Barcode of Life Data System (BOLD) (Ratnasingham et al., 2024; Ratnasingham & Hebert, 2007) is the global repository for DNA barcode sequences (Hebert et al., 2003). It currently contains 21.8 million DNA sequences, which have received considerable taxonomic curation. Nevertheless, because many species remain undescribed, some barcode records cannot gain a taxonomic assignment below the family, subfamily, or genus level. Moreover, even known species often require attention from expert taxonomists to associate a sequence with a species. Barcode Index Numbers (BINs)—clusters of barcode sequences assigned to a unique alphanumeric identifier—offer a solution (Ratnasingham & Hebert, 2013). Every month, sequences new to BOLD are either assigned to an existing BIN or found new ones.

The computational requirements for these taxonomic assignments are certain to grow for two reasons: 1) DNA barcode reference libraries are expanding at an ever-accelerating pace due to decreasing analytical costs and increasing uptake; and 2) the data files generated by DNA metabarcoding and DNA protocols are surging because of massive increases in sequence generation. Jointly these factors drive the Carlson curve (Carlson, 2003), which shows that the pace of biotechnological development scales in a similar fashion to computational advances (Moore, 1965). Indeed, DNA sequencing technology is advancing faster than Moore's Law (Muers, 2011), leading to an ever-increasing gap between the supply and demand of computational resources. The solution lies in accelerating the computational capacity for analysis and/or reducing the computational demand by reducing the size of reference libraries without compromising their ability to deliver taxonomic inferences.

Here, we describe an approach that generates comprehensive DNA barcode reference libraries distilled from data on BOLD. Each 'BOLDistilled' library captures a snapshot of the genetic diversity and associated taxonomic information on BOLD in a minimalist dataset. These libraries are versioned, publicly available, and archived to allow the reproducibility of

analyses conducted on a particular sequence array. The distillation process employs an open-source algorithm to capture the full range of genetic variation while retaining the fewest records for each BIN (or operational taxonomic unit [OTU] for taxa outside the animal kingdom). We also resolve each BIN (or OTU) to a single consensus taxonomy to reduce ambiguities which would otherwise be introduced during the assignment of sequences to their source taxa. These libraries, which include both public and anonymized private sequences, will initially be published quarterly through a doi at boldsystems.org/BOLDistilled with past versions archived and accessible. We present the first BOLDistilled library for the barcode region of cytochrome *c* oxidase I (COI), the most heavily represented locus on BOLD (96% of records), and the only locus for which BINs are generated.

Materials and Methods

Unless otherwise stated, all analyses were performed on a custom-built computer equipped with an AMD Ryzen ThreadRipper 7980Xs 128-core CPU, 128 GB RAM, and an NVIDIA GeForce RTX 4090 GPU. The operating system was Ubuntu 24.04.1 LTS.

All 21.8M records on BOLD were downloaded on March 10, 2025. Among them, 17.3M with BIN assignments were retained, leading to coverage for 1.2M BINs. Another 220K records belonging to prokaryote (e.g., aerobic bacteria) and other eukaryote (Fungi, Protista) lineages were also retained. Although they lack BIN assignments, retention of these taxa is important because sequences deriving from them are often present among sequence arrays recovered from barcoding, metabarcoding, and eDNA analyses (Hallam et al., 2021; Young & Hebert, 2022). The latter records were clustered into OTUs using VSEARCH (Rognes et al., 2016) to create BIN analogues for the purpose of data distillation (see below). The final BOLDistilled library contained approximately 1.2M records with BIN assignments and another 23K records with OTU assignments (hereafter, for simplicity, we refer to both BINs and OTUs as ‘BINs’).

The BOLDistilling process (Fig. 1) employs an algorithm that acts on each BIN to select the minimal number of sequences that effectively capture its genetic diversity. The distance threshold, which acts to exclude similar sequences, is the key parameter for distillation (discussed below). If a BIN contains only a single record, its sequence is retained as the representative while those with multiple records are distilled in the following way:

1. Duplicate sequences are removed. If just one sequence remains, it is the representative for that BIN.

2. If multiple sequences remain, a single focal sequence is selected from them and added to the BOLDistilled library.
3. Genetic distance is then calculated between that focal sequence and all other sequences in this BIN.
4. All sequences with a distance above the threshold are retained, while those below it are discarded.
5. A new focal sequence is then haphazardly selected from the remaining sequences. The selection process continues (from step 3) in an iterative fashion until no sequences remain.
6. The process is extended to the next BIN until all BINs have been processed.

The Process ID is the unique identifier on BOLD for each record and its associated DNA sequence and taxonomy. BOLDistilled libraries report the Process ID for each public record while those from private records are concealed. BOLDistilled libraries minimize the number of private records by preferentially selecting public records as focal sequences.

In addition to reducing genetic redundancy, an effective reference library must possess a single consensus taxonomic assignment for each BIN. Arriving at this consensus requires the resolution of taxonomic conflicts among specimens. This is carried out by an R script (R Core Team, 2022) that examines the taxonomic hierarchy for every member of each BIN in the complete BOLD library. At each level of the hierarchy (kingdom to species), ≥75% agreement is taken as its taxonomic assignment. For example, if all taxonomic assignments for members of a BIN are congruent down to a genus, but half are assigned to one species and the rest to another species, the BIN taxonomy is only resolved to a genus. Missing taxonomy is not considered discordance, so if even one member of a BIN has, for example, a generic ID, the BIN gains that identification unless discordance is introduced through future curation or new data. We note that Process IDs in BOLDistilled libraries refer to BOLD records containing reference sequences, but the taxonomy associated with individual records in that BIN might differ from the consensus taxonomy in the BOLDistilled library.

We used a 0.75% divergence threshold to generate the BOLDistilled library for COI after trials with varying thresholds. Higher thresholds reduce the overall library size because fewer representatives are included per BIN, but this can compromise representation of intra-BIN genetic variation. Conversely, lower divergence thresholds result in broader intra-BIN genetic diversity at the cost of a larger library. Our tests indicated that a 0.75% divergence threshold led to a nearly ten-fold reduction in the size of the BOLDistilled library while offering enough resolution to accurately infer taxonomy in taxa with high intra-specific genetic variation. This value might change slightly with future study and will be reported in the metadata accompanying each BOLDistilled library.

To validate the BOLDistilled COI library, we compared its performance against the complete 17.5M COI library on BOLD. We analyzed two metabarcoded Malaise trap samples—one from Canada and one from Australia—and the resulting sequences were identified using both the complete and distilled reference libraries. Briefly, we lysed the bulk samples with a guanidine thiocyanate-based lysis buffer and extracted bulk DNA from three replicates of lysate per sample. We amplified the COI barcode using standard methods (Hebert et al., 2018) and sequenced the resulting amplicons on an Oxford Nanopore Technologies (Oxford, UK) PromethION flow cell on a PromethION P2 Solo sequencer following the manufacturer's recommendations for the SPK-LSK114 ligation module. We filtered, demultiplexed, and clustered the reads with a custom pipeline (Prosser, unpublished) and then ran the resulting fasta files through `VSEARCH –usearch_global` using both reference libraries. From these results, we compared the time required to complete analysis, the number of BINs detected, and the identity of BINs.

BOLDistilled libraries are available at the following URL: boldsystems.org/BOLDistilled [note: URL will be activated following acceptance for publication]. The latest reference library for a given gene region or taxonomic group will be available via a download link. Earlier versions will remain available in a linked persistent repository. Each library will include the underlying sequences, their corresponding consensus taxonomy, and a summary of the library's metadata and structure, which allows users to convert them into their desired formats. To that end, we will also make each library available in the formats of popular taxonomic assignment algorithms. All scripts and output data used in this study are available on the Zenodo digital repository (doi: 10.5281/zenodo.15442656).

Results & Discussion

The BOLDistilled library (BOLDistill_COI_Mar2025) contained 82.3% fewer sequences (1.2M records) than the complete BOLD reference library (17.5M records) when a divergence threshold of 0.75% was used. Among these records 24% were resolved to species, 36% to genus, and 93% to family. To compare their performance, we queried both the complete and BOLDistilled reference libraries using two data sets with 15-fold difference in read depths and with less than 1% overlap in BIN composition—the sample from Canada (CDN) had 755,083 reads while the Australian (AUS) sample had 11,244,319 reads—both estimated to contain 400–500 BINs. In our tests, `VSEARCH –usearch_global` took 10–20 minutes to analyze the sequences when the complete BOLD reference library was used (CDN: 592 s; AUS: 1180 s). By comparison, the same analysis with the BOLDistilled library was completed in seconds (CDN: 12 s; AUS: 18 s), a 98% reduction in computational time for both samples. Because multiple samples are often analyzed in a run the incorporation of a BOLDistilled

library into bioinformatic workflows can reduce 24 hours of computation to less than 30 minutes.

To compare taxonomy assignment performance on standard computers, we ran `VSEARCH --usearch_global` on the Australian sample using two laptops—a 2020 MacBook Air M1 equipped with 8 GB RAM and a 2023 MacBook Pro M2Pro equipped with 16 GB RAM. The MacBook Air was unable to run VSEARCH using the complete library due to insufficient memory, though it completed analysis with the BOLDistilled library in 96 s. The MacBook Pro could run both libraries, taking 3848 s (64 min) with the complete library versus 66 s with the BOLDistilled library (98.3% reduction). These results highlight an important feature of BOLDistilled libraries: they support users without access to high-end performance computers, and they abolish a barrier to research in labs with limited funding. Indeed, the ability of BOLDistilled libraries to run locally on low-end computers without an Internet connection makes them ideal for use in remote communities, progressing the democratization of biodiversity research.

Time for computation is also affected by the sequencing platform employed (via read depth) and by the taxonomic diversity of the sample being investigated. While read depth directly affects compute time for processes such as read filtering, demultiplexing, and clustering, taxonomic assignment algorithms depend primarily on the number of BINs requiring identification—read depth of the BINs is largely inconsequential at this stage. However, high read output typically results in more BINs per sample—rare BINs are more likely to be recovered with higher read depth—so both sequencer output and taxonomic diversity of the sample will affect the compute time of BIN identification.

In terms of taxonomic matches, the reference libraries showed similar performance. In the Canadian sample, use of the complete library led to the detection of 447 BINs versus 448 BINs with the BOLDistilled library. In the Australian sample, the complete library detected 483 BINs versus 485 BINs with the BOLDistilled library. In both cases, the BIN array was nearly identical (CDN: 98.1% overlap; AUS: 98.8% overlap). Differences between the complete and distilled libraries typically involved very closely-related BINs—attributable to real biological variation. Consider a query sequence whose best match in the complete library is to BIN ‘A’ and whose second-best match is to closely related BIN ‘B’. If, after BOLDistillation, the best match reference sequence from BIN ‘A’ is removed and the second-best match from BIN ‘B’ is retained, the query sequence will now match to BIN ‘B’. These cases are uncommon, and the inferred taxonomic composition of a sample is near identical whether using the complete or distilled library.

We believe BOLDistilled libraries, or an analogue, should be used as the basis for assigning the taxonomic source of sequences recovered through metabarcoding or eDNA studies. With it, users can incorporate a DNA barcode reference library into their own workflow whether they use VSEARCH (as we have) or other taxonomic assignment algorithms, such as BLAST (Camacho et al., 2009) or SINTAX (Rognes et al., 2016). BOLDistilled libraries converted into formats compatible with popular algorithms are available on boldsystems.org/BOLDistilled.

Each BOLDistilled library represents a snapshot of all DNA barcode data available at the time of its creation, a record that will be maintained in perpetuity to facilitate the reproducibility of analytical results generated using it. By resolving taxonomic inconsistencies within BINs prior to analysis, these libraries also reduce the risk of misidentifications linked to taxonomic uncertainty among individual barcode records. Importantly, they also maintain meaningful intraspecific genetic variation. The above concerns have been highlighted by the metabarcoding community as key shortcomings in past approaches to the construction of reference libraries (Keck et al., 2023). By addressing these deficits, and because they will be updated as BOLD grows, BOLDistilled libraries are positioned to respond rapidly to future demands.

Presently, a BOLDistilled library is only available for COI. The BIN algorithm and our sequence divergence threshold have been fine-tuned based on our collective expertise into this locus. Similar libraries can certainly be produced for other loci (e.g., *rbcLa* or ITS2) and we will generate them based on demand and further exploration of the distillation parameters. While software packages exist to aid the manipulation and curation of publicly available reference sequences (Keck & Altermatt, 2023), BOLDistilled libraries require no further optimization, are fully traceable and versioned, and can be customized by researchers to include only taxa of interest. Looking forward, we believe these libraries should be adopted as standards that maximize the utility of DNA barcode reference libraries for the scientific community.

Figure

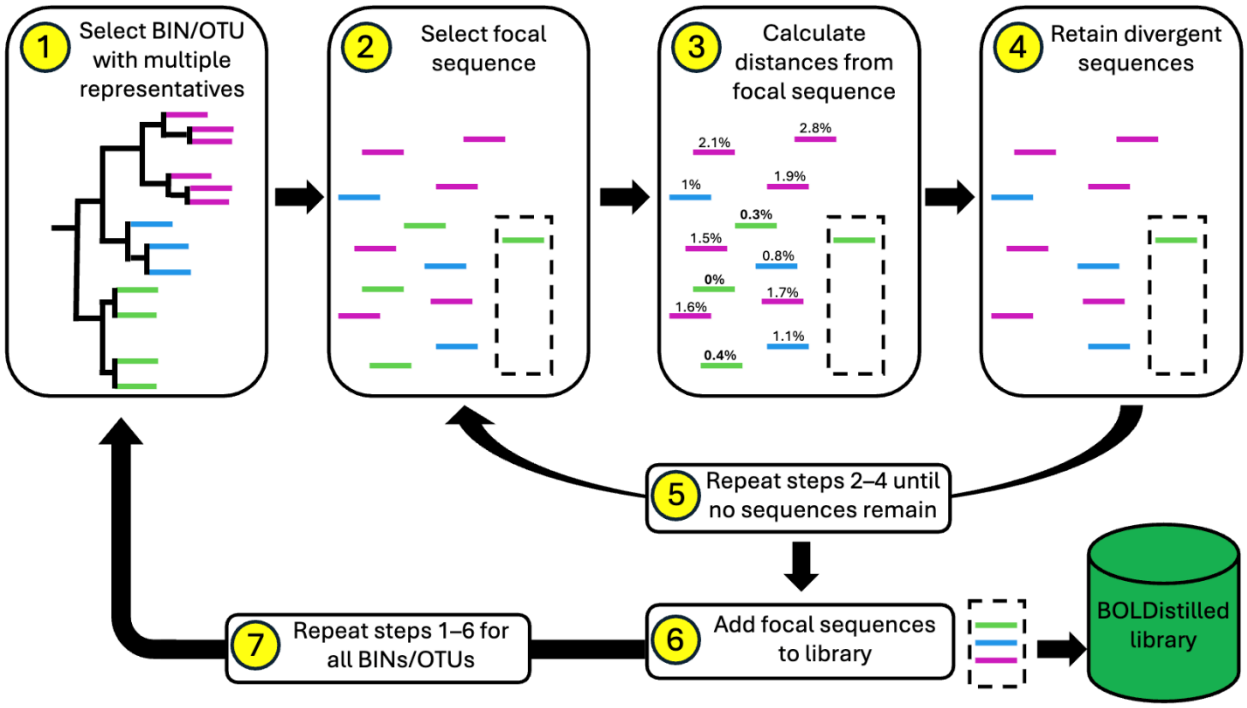

Fig. 1. Illustration of the BOLDistiller algorithm. BINs or OTUs with multiple sequences are selected for distillation (step 1). The tree shows an example of a single BIN with high intra-BIN variation with each colour indicating an intra-BIN cluster from which a single representative should be retained. First, a focal sequence is selected from the pool (step 2) and divergence between the focal sequence and all other sequences is calculated (step 3). Sequences with high divergence (≥0.75% for COI) from the focal sequence are retained (step 4). The process is repeated until no sequences remain (step 5). The set of focal sequences for this BIN are added to the BOLDistilled library (step 6), and the process continues with the next BIN or OTU in the list (step 7).

Acknowledgements

We thank Chris Ho, Sujeevan Ratnasingham, Catherine Wei, Marlee Lyle, and Sameer Padhye for their support and advice in implementing BOLDistilled. This work was supported by the New Frontiers in Research Fund (NFRFT-2020-00073), by the Canada Foundation for Innovation (MSI 42450), and by the Government of Canada through Genome Canada and Ontario Genomics (OGI-233).

Data availability statement

This study contains no original data. The genetic resource resulting from our study is available on Zenodo (doi: 10.5281/zenodo.15442656) under a Creative Commons Attribution 4.0 International license.

Author contributions

- Algorithm design: SWJP, KAT, RMF
- Algorithm implementation: SWJP, RMF
- Writing the paper: SWJP, RMF, KAT, PDNH
- Project supervision: PDNH
- Resource Acquisition: PDNH

Supplemental Files

- BOLDistill_COI_Mar2025_SEQUENCES.fasta
- BOLDistill_COI_Mar2025_TAXONOMY.tsv
- BOLDistill_COI_Mar2025_METADATA.tsv
- BOLDistill_COI_Mar2025_blast (folder containing several files)
- BOLDistill_COI_Mar2025_vsearch (single file)
- BOLDistill_COI_Mar2025s_sintax (folder containing two files)
- BOLDistill.sh
- BOLDistill.R
- BOLDistill.rmd
- BOLDistill_sintax.py

References

Ahmed, M., Back, M. A., Prior, T., Karssen, G., Lawson, R., Adams, I., & Sapp, M. (2019). Metabarcoding of soil nematodes: the importance of taxonomic coverage and availability of reference sequences in choosing suitable marker(s). *Metabarcoding and Metagenomics*, *3*, e36408. https://doi.org/10.3897/mbmg.3.36408

Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA Metabarcoding for Biodiversity-Based Monitoring and Assessment. *Frontiers in Marine Science*, *3*. https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2016.00096

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Carlson, R. (2003). The pace and proliferation of biological technologies. *Biosecurity and Bioterrorism : Biodefense Strategy, Practice, and Science*, *1*(3), 203–214. https://doi.org/10.1089/153871303769201851

Hallam, J., Clare, E. L., Jones, J. I., & Day, J. J. (2021). Biodiversity assessment across a dynamic riverine system: A comparison of eDNA metabarcoding versus traditional fish surveying methods. *Environmental DNA*, *3*(6), 1247–1266. https://doi.org/https://doi.org/10.1002/edn3.241

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V, Janzen, D. H., Hallwachs, W., Naik, S., Sones, J. E., & Zakharov, E. V. (2018). A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics*, *19*(1), 219. https://doi.org/10.1186/s12864-018-4611-3

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. Biol. Sci.*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Keck, F., & Altermatt, F. (2023). Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. *Molecular Ecology Resources*, *23*(2), 511–518. https://doi.org/10.1111/1755-0998.13723

Keck, F., Couton, M., & Altermatt, F. (2023). Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Molecular Ecology Resources*, *23*(4), 742–755. https://doi.org/https://doi.org/10.1111/1755-0998.13746

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2020). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, *45*(3), 373–385. https://doi.org/https://doi.org/10.1111/een.12831

Moore, G. E. (1965). Cramming More Components onto Integrated Circuits. *Electronics*, *38*(8), 114–117.

Muers, M. (2011). Technology: Getting Moore from DNA sequencing. *Nature Reviews Genetics*, *12*(9), 586–587. https://doi.org/10.1038/NRG3059

Pringle, R. M., Kartzinel, T. R., Palmer, T. M., Thurman, T. J., Fox-Dobbs, K., Xu, C. C. Y., Hutchinson, M. C., Coverdale, T. C., Daskin, J. H., Evangelista, D. A., Gotanda, K. M., A. Man in 't Veld, N., Wegener, J. E., Kolbe, J. J., Schoener, T. W., Spiller, D. A., Losos, J.

311 B., & Barrett, R. D. H. (2019). Predator-induced collapse of niche structure and
312 species coexistence. *Nature*, *570*(7759), 58–64. https://doi.org/10.1038/s41586-019-
313 1264-6

314 R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R
315 Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

316 Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System
317 (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364.
318 https://doi.org/https://doi.org/10.1111/j.1471-8286.2007.01678.x

319 Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species:
320 the Barcode Index Number (BIN) system. *PLOS ONE*, *8*(7), e66213-.
321 https://doi.org/10.1371/journal.pone.0066213

322 Ratnasingham, S., Wei, C., Chan, D., Agda, J., Agda, J., Ballesteros-Mejia, L., Boutou, H. A.,
323 El Bastami, Z. M., Ma, E., Manjunath, R., Rea, D., Ho, C., Telfer, A., McKeowan, J.,
324 Rahulan, M., Steinke, C., Dorsheimer, J., Milton, M., & Hebert, P. D. N. (2024). BOLD
325 v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data. *Methods*
326 *in Molecular Biology (Clifton, N.J.)*, *2744*, 403–441. https://doi.org/10.1007/978-1-
327 0716-3581-0_26

328 Rimet, F., Aylagas, E., Borja, A., Bouchez, A., Canino, A., Chauvin, C., Chonova, T.,
329 Ciampor Jr, F., Costa, F. O., Ferrari, B. J. D., Gastineau, R., Goulon, C., Gugger, M.,
330 Holzmann, M., Jahn, R., Kahlert, M., Kusber, W.-H., Laplace-Treyture, C., Leese, F., …
331 Ekrem, T. (2021). Metadata standards and practical guidelines for specimen and DNA
332 curation when building barcode reference libraries for aquatic life. *Metabarcoding and*
333 *Metagenomics*, *5*, e58056. https://doi.org/10.3897/mbmg.5.58056

334 Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile
335 open source tool for metagenomics. *PeerJ*, *4*. https://doi.org/10.7717/peerj.2584

336 Young, M. R., & Hebert, P. D. N. (2022). Unearthing soil arthropod diversity through DNA
337 metabarcoding. *PeerJ*, *10*. https://doi.org/10.7717/PEERJ.12845

338