Int. J. Lim. 2025, 61, 10
© EDP Sciences, 2025
https://doi.org/10.1051/limn/2025009

Available online at:
www.limnology-journal.org

RESEARCH ARTICLE

Assigning taxonomy and traits to DNA sequences of river diatoms in a region with limited taxonomic knowledge using the updated and annotated reference library Diat.barcode

Maria Mercedes Nicolosi Gelis[1,*], Raphaëlle Barry-Martinet[3], Jean-François Briand[3], Teofana Chonova[4], Joaquín Cochero[1], Gilles Gassiole[5], Maria Kahlert[6], Balasubramanian Karthick[17], François Keck[7], Martyn Kelly[8], Hristina Kochoska[9], David Mann[10,12], Pratyasha Nayak[17], Martin Pfannkuchen[11], Tobias Servulo[3], Rosa Trobajo[12], Valentin Vasselon[13], Danijela Vidakovic[14], Laurine Viollaz[2], Carlos Wetzel[15], Jonas Zimmermann[16] and Frederic Rimet[2]

[1] Instituto de Limnología "Dr. Raúl A. Ringuelet" ILPLA CONICET-UNLP, Bv. 120 y 62 n 1437, La Plata, Buenos Aires, Argentina
[2] UMR CARRTEL, INRAE, Université Savoie-Mont Blanc, 75bis av. de Corzent – CS 50511, 74203 Thonon les Bains cedex, France
[3] MAPIEM, Université de Toulon, Toulon, France
[4] Department of Environmental Chemistry, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstr. 133, CH-8600 Dübendorf, Switzerland
[5] MicPhyc, 88 chemin Manouilh, Ilet à Vidot 97433 Salazie, La Réunion
[6] Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, PO Box 7050, SE- 750 07 Uppsala, Sweden
[7] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland. Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department of Aquatic Ecology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland
[8] Bowburn Consultancy, 11 Monteigne Drive, Bowburn, Durham DH6 5QB, UK and School of Geography, University of Nottingham, Nottingham NG7 2RD.
[9] Department of Ecology, Faculty of Science, Charles University, Viničná 7, Prague 2, CZ-12844, Czech Republic
[10] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, Scotland, UK
[11] Center for Marine Research, Ruder Boskovic Institute, Giordano Paliaga 5, 52210 Rovinj, Croatia
[12] IRTA, Marine and Continental Waters, Ctra Poble Nou km 5.5, La Ràpita, Catalonia, Spain
[13] SCIMABIO Interface, 5 Rue des Quatre Ventes, 74200 Thonon-les-Bains
[14] University of Belgrade, Institute of Chemistry, Technology and Metallurgy, National Institute of the Republic of Serbia, Njegoseva 12, 11000 Belgrade, Serbia
[15] Luxembourg Institute of Science and Technology (LIST), Environmental Sensing and Modelling unit, Biodiversity Monitoring and Assessment group, 41 rue du Brill, L-4422 Belvaux, Luxembourg
[16] Botanic Garden and Botanical Museum, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195 Berlin, Germany
[17] Biodiversity and Palaeobiology Group, Agharkar Research Institute, Pune, 411004, India. Department of Botany, Savitribai Phule Pune University, Pune, Maharashtra-411007, India

Received: 11 July 2025; Accepted: 13 October 2025

Abstract – We present a new version of a barcoding reference library dedicated to diatoms, Diat.barcode v12, with newly published sequences, annotated with ecological, and biological traits (size class, life forms, ecological guilds, *etc*) and curated by a college of experts. Diat.barcode incorporates *rbcL* data on all diatoms, though freshwater taxa are better represented. We used this library in two different areas, one where the taxonomic coverage of the library was good (mainland France) and another where it was poor (French Guyana) with about 320 diatom samples collected for river monitoring across both regions. We show that a direct bioinformatic assignment of environmental sequences to traits (ecological guilds, habitats, morphology) has great potential in French Guyana where species knowledge is poor and therefore the proportion of assigned environmental sequences is much lower (12.8%) than trait assignation (30%). We used co-correspondence analyses to show that, unlike mainland France where all species and trait

*Corresponding author: mercedesnicolosi@ilpla.edu.ar

assignation datasets were significantly correlated, in French Guyana only 7 out of 13 trait categories showed a significant correlation. This indicates a significant loss of ecological information through species assignment in French Guyana. Consequently, directly assigning environmental sequences to traits can be useful and provide more ecological information in regions with poor taxonomic knowledge because of the many *rbcL* sequences without taxonomic assignations.

Keywords: Diatom / monitoring / metabarcoding / traits / reference library

1 Introduction

Diatoms are microalgae that are widely used to assess the quality of rivers and lakes (*e.g.* Hering *et al.*, 2006; Rimet, 2012). The method routinely used is based on morphological identifications and frustules counting by microscopy (CEN, 2014). However, biodiversity assessment using molecular methods, and in particular using DNA metabarcoding, is becoming increasingly relevant in biodiversity studies (Keck *et al.*, 2022) and monitoring programs (Leese *et al.*, 2016; Laamanen *et al.*, 2025). This is based on the sequencing of a short DNA fragment easily amplified from DNA extracted from a natural community. Metabarcoding offers an alternative that is of interest to stakeholders for reasons of cost and data throughput. Metabarcoding methods for diatoms are operationally advanced. This progress comes from extensive research, including testing various DNA extraction methods (Vasselon *et al.*, 2017) and selecting a 263 bp *rbcL* gene barcode suitable for species identification (Kermarrec *et al.*, 2013; Vasselon *et al.*, 2017). Dedicated primers have also been developed (Vasselon *et al.*, 2017). Further advancements involve testing and selecting bioinformatics algorithms (Bailet *et al.*, 2019; Rivera *et al.*, 2020), evaluating gene copy number per cell on read number (Vasselon *et al.*, 2018; Pérez Burillo *et al.*, 2020), and conducting proficiency tests (Vasselon *et al.*, 2025). Additionally, DNA preservation has been evaluated through tests ranging from one day to 1 yr (Baricevic *et al.*, 2022). Finally, the methods have undergone intercalibration (Vasselon *et al.*, 2025) and standardization of protocols at a European level (CEN, 2018, 2025; Kelly, 2019).

However, challenges remain, particularly concerning taxonomic assignments to diatom species. Taxonomic identification of species from DNA sequences critically depends on reference libraries that connect genetic sequences to scientific names. In this sense, the effectiveness of metabarcoding depends on the development and upkeep of reliable reference libraries, which must be comprehensive, expertly curated and regularly updated to incorporate new DNA barcodes and taxonomic refinements (Gauvin *et al.*, 2025). The rapid rise of molecular techniques has been accompanied by a large amount of new DNA sequences (*i.e.*, DNA barcodes), which are typically made available through online repositories such as ENA or GenBank (Sayers *et al.*, 2020). Some of these sequences undergo expert taxonomic curation and are incorporated into curated reference libraries. For microalgae in particular, these reference libraries include Phytool (Canino *et al.*, 2021), PhytoRef (del Campo *et al.*, 2018) and µgreen-db (Djemiel *et al.*, 2020). For diatoms, Diat.barcode library has offered an open-access resource since 2012, with a fine-tuned taxonomy at the genus and species levels curated by a network of diatom experts (Rimet *et al.*, 2019). Issues related to reference libraries and their use for metabarcoding purpose, can be summarized into seven challenges (Keck *et al.*, 2023): (i) mislabelling, which refers to errors in the identification of biological material deposited in the reference library; (ii) sequencing errors, when an erroneous DNA sequence (PCR errors, chimeras, poor sequence quality) is linked with biological material; (iii) sequence conflict, where different taxa are assigned to an identical genetic sequence; (iv) taxonomic conflict, which occurs when the same organism is registered several times in the reference database with different taxonomic identities; (v) low taxonomic resolution, where organisms are identified only at higher taxonomic ranks; (vi) missing taxa, an important problem when the proportion of all existing taxa is important (especially in poorly studied regions), and (vii) missing intraspecific variants, as different genetic variants (haplotypes) within a single species are often overlooked. In addition to these challenges, other relevant issues have been identified, such as (viii) the assignment of sequence information from contaminant species to correctly identified main culture species, and (ix) the undersampling of species diversity within specific taxa.

These limitations underscore the need for continuous refinement of methods and reference libraries to improve the accuracy and relevance of taxonomic assignments. Taxonomic assignment, a critical step in bioinformatics workflows, links DNA sequences to a scientific name. This enables the use of ecological preferences specific to taxa (Keck, 2023) in order to calculate biotic indices to measure anthropogenic pressures (Tapolczai *et al.*, 2019). This step is also essential if metabarcoding is to be used for conservation purposes, as many legal and ecological frameworks rely on species-level identification (*e.g.*, the U.S. Endangered Species Act, IUCN lists, invasive species management). The completeness and accuracy of reference libraries and their integration with advanced taxonomic assignment algorithms define the predictive power of metabarcoding analyses and their applicability to biodiversity monitoring (Murali *et al.*, 2018; Keck, 2023), as well as their interoperability with taxonomic backbones such as the GBIF Backbone Taxonomy (GBIF, 2024). However, completeness and quality of reference libraries are often geographically and taxonomically biased. For instance, Marques *et al.* (2021) demonstrated that gaps in reference libraries for fish species increase towards the tropics. Similarly, Weigand *et al.* (2019) found that certain taxonomic groups used in European aquatic monitoring, such as fish, true bugs, and freshwater vascular plants, are better represented in reference libraries compared to groups such as freshwater diatoms and marine molluscs. In China, Li *et al.* (2022) reported that barcode libraries for aquatic taxa frequently reflect geographical biases tied to the origin of the specimens and sequences. These biases are even more pronounced in remote and biodiverse regions where taxonomic knowledge is limited, and many species remain undescribed. Studies on the

diatom flora of Mayotte Island – a tropical island near Madagascar, for example, revealed significant gaps in reference libraries despite efforts to isolate, culture and sequence local strains (Kermarrec *et al.*, 2013; Vasselon *et al.*, 2017). Although advances have been made in creating libraries dedicated to particular tropical areas (*e.g.* Ecuador, Ballesteros *et al.*, 2020), culturing diatoms from field samples to generate reliable DNA barcodes is labor-intensive and often unsuccessful due to the challenges of maintaining viable cultures. Alternative approaches, such as single-cell PCR (Hamilton *et al.*, 2015; Khan-Bureau *et al.*, 2016) and OTU co-abundance networks (Irannia and Chen, 2016), have been proposed to fill these gaps. Another alternative is to focus on integrating environmental sequences from high-throughput sequencing (HTS) runs with morphological observations to improve library completeness (Rimet *et al.*, 2018). Applying strict quality criteria helps ensure the reliability of these additions, but substantial work remains to address the large taxonomic gaps in tropical and remote regions.

In regions with incomplete reference libraries or poorly known taxonomy, trait-based approaches can offer a usefull complementary perspective for ecological and environmental monitoring. By considering functional and morphological traits, such as ecological guilds, researchers can adress some of the limitations associated with species-level taxonomic identification and explore relevant ecological questions (Berthon *et al.*, 2011; Stenger-Kovács *et al.*, 2018). While species composition can vary widely across regions, diatom traits composition often shows considerable overlap, enabling functional comparisons of communities even between remote areas with differing floras (Soininen *et al.*, 2016). This makes trait-based methods particularly valuable for assessing the ecological status of areas where there is limited taxonomic knowledge. Trait-based approaches have already been applied extensively in running waters, where diatom traits have been linked to environmental variables such as nutrient levels, organic pollution, grazing, and shear stress (Berthon *et al.*, 2011; Lange *et al.*, 2016; Tapolczai *et al.*, 2017) and to a lesser degree in lakes (Gottschalk and Kahlert, 2012; Rimet *et al.*, 2015; Zorzal-Almeida *et al.*, 2017). Building on the success of trait-based phytoplankton studies (Salmaso and Padisák, 2007; Kruk *et al.*, 2010), these approaches offer a practical framework to assess functional responses of diatom assemblages to environmental change, particularly in understudied tropical and remote regions. In addition to the interest in using ecological and biological traits of diatoms for biomonitoring and fundamental ecological studies, there is a very pragmatic interest related to identification because identifying diatom traits (mobility, size class, colony shape, *etc*) is very easy to do using microscopy and does not require great expertise. A recent study tested successful trait assignment with metabarcoding data of lake phytoplankton and stressed the potential of such bioinformatic strategies for ecological studies when reference libraries are poorly documented (Tapolczai *et al.*, 2025).

This study introduces a new version (v12) of the annotated reference library Diat.barcode, with the following objectives: 1) to expand and curate the library, integrating new sequences to enhance taxonomic coverage; 2) to annotate the library with ecological and biological traits in order to facilitate trait-based assessment; 3) to evaluate trait-based assignation, comparing the added value of assigning environmental sequences directly to traits compared to scientific names, particularly in regions where taxonomy is poorly known and; 4) leverage phylogenetic conservation of traits: by exploring the assumption that many traits are phylogenetically conserved, we hypothesize that bioinformatic assignment of traits directly to environmental sequences can help to reduce the number of unassigned sequences compared to taxonomic assignment, particularly in regions where taxonomic knowledge is limited, such as the tropics. Indeed, we expect that if most traits are phylogenetically conserved, then the attribution of traits to amplicon sequence variants (ASVs) –without any taxonomic assignment- will be facilitated, since bioinformatic attribution is done on the basis of genetic proximity.

To this end we carried out the assignment tests in two areas where river sampling campaigns were carried out and high-throughput sequencing performed. The first area is mainland France, where Diat.barcode reference library has good coverage at species level, and the second area is French Guyana in South-America which has a high level of diversity that is unknown and has never been described, and for which Diat.barcode library should have significant gaps.

2 Methods

2.1 Integrating newly published sequences into Diat.barcode

To update Diat.barcode from v11 to v12, new *rbcL* barcode sequences published in the NCBI (National Center for Biotechnology Information) database between June 24, 2021, and February 4, 2024, were downloaded. A total of 1,080 new sequences were obtained. The curation process followed the protocol described by Rimet *et al.* (2019), which consists of two main steps. First, the quality and length of the sequences are assessed, and the sequences are incorporated into a multiple sequence alignment. Second, a constrained phylogenetic analysis is performed to ensure that similar sequences (*e.g.*, those belonging to the same clade) are assigned the same taxonomic name. If discrepancies are found, a taxonomic curation procedure (see details in Rimet *et al.*, 2019) is applied to verify and correct the names including: 1) taxonomic homogenization based on peer-reviewed literature results, 2) check of taxonomic synonyms, 3) examination of photos/slides to update the taxonomic identification.

To construct the reference phylogeny, we used the multiple alignment available in the reference library Diat.barcode v12 (Rimet *et al.*, 2019), selecting only long sequences as defined in the Diat.barcode protocol (Rimet *et al.*, 2019). The final alignment consisted of 1,046 positions and included 3,176 sequences from Diat.barcode v11, and 408 newly downloaded sequences from NCBI. These sequences were selected from an initial set of 1,092 sequences, of which 672 were excluded for being shorter than the established threshold. A maximum likelihood phylogenetic analysis was performed using RAxML v8.2.12 (Stamatakis, 2014) on the Migale Bioinformatics Facility (MIGALE and INRAE, 2020). The analysis employed the GTRGamma model and included 100 rapid bootstrap replicates. The computation was run using 8 threads and required 15.9 h to complete.

The shorter sequences were aligned in a FASTA file and included 672 sequences from the 1,092 newly downloaded sequences, and 2,843 sequences from Diat.barcode v11. These

shorter sequences were placed into the reference phylogeny using RaxmlGUI (Silvestro and Michalak, 2012) in "enforce constrain mode" with the "backbone option". The computation was performed on a laptop with an Intel Core i7-8665U CPU @ 1.90 GHz (2.11 GHz boost) and required 50.82 h to complete.

Each new sequence was evaluated within the phylogeny to verify whether its phylogenetic neighbor had a similar taxonomic name. When discrepancies were identified, taxonomic corrections were made. A total of 72 changes in taxonomic names were performed, distributed as follows: 15 updates to scientific names of sequences from previous versions of Diat.barcode (v11 and earlier). For example, "*Cyclotella stylorum"* was updated to "*Stephanocyclus stylorum"*. The other changes (57) were applied in newly downloaded sequences, categorized as: 17 name homogenizations, such as "*Licmophora* sp. A" standardized to "*Licmophora* sp.", 14 corrections of misspellings, such as "*Chaetoceros muellerii"* corrected to "*Chaetoceros muelleri",* 7 taxonomic updates, such as "*Amphora incrassata"* updated to "*Halamphora incrassata",* 7 cases where unidentified species were assigned specific names, such as "*Lindavia* sp." identified as "*Lindavia thienemannii",* and 1 case where an unidentified taxon was assigned a name, "Bacillariophyta sp. CLi-2022a" reclassified as "*Rhoicosphenia* sp.". This process ensured consistency and accuracy in the taxonomic assignments within the updated reference phylogeny.

The number of sequences available in Diat.barcode v12 throughout the different families was represented with a taxonomic tree drawn using the refdb r package (Keck & Altermatt, 2023).

2.2 Annotation of Diat.barcode v12 with biological and ecological traits and test of their phylogenetic signal

Diat.barcode was updated to include biological and ecological traits, assigned to species based on the literature, for both former and newly added sequences. The traits included are chloroplast number, chloroplast shape, biovolume, frustule dimensions, size class, habitat type (marine, freshwater, benthic, planktonic, epipsammic, epipelic, wet soil), mobile, unattached, attached, adnate, pedunculate, stalk, pad attached to substrate, colonial, non-colonial, type of colony (mucous tubule colony, chain colony, zig-zag colony, rosette colony, ribbon colony, stellate colony, arbuscular colony), ecological guilds (high profile guild, motile guild, planktonic guild, low profile guild). Specific bibliography was employed for the classification of diatom traits: Round *et al.* (1990), Krammer and Lange-Bertalot (1986, 1988, 1991a, 1991b) floras, Passy (2007), Rimet and Bouchez (2012), Cox (1996), algaebase.org (Guiry and Guiry, 2025) and diatoms.org (Spaulding, 2021).

The strength of each trait as a predictor of phylogeny was tested using the function phyloSignal of the phylosignal v1.3.1 package (Keck *et al.*, 2016), using the phylogeny with short and long sequences of Diat.barcode with the K.star test function, which computes permutation test for Blomberg's K Star (Blomberg *et al.*, 2003). The K.star statistic is based on an evolutionary model that compares observed values (in our case, the observed value of a trait as given in Diat.barcode) with those expected under a Brownian motion model. The representation of diatom traits in the phylogeny was generated using the barplot.phylo4d function in the phylosignal package.

2.3 Presentation of the environmental datasets and diatom metabarcoding

We used *rbcL* HTS data obtained from two river networks, one in the French overseas department in South America, French Guyana, the other one in the mainland French metropolitan territory. Aquatic biofilms (periphyton) were collected during the river monitoring campaigns carried out within the framework of the European Water Framework Directive, corresponding to the 2013 to 2017 yearly campaigns for the Guyana rivers (85 sites, 164 samples) and the 2017 and 2018 campaigns for the metropolitan rivers (87 sites, 162 samples).

Aquatic biofilms containing the benthic assemblages of diatoms were sampled according to the current standard (NFT 13946, AFNOR, 2014). The recommendations of the technical report, CEN/TR 17245 (CEN, 2018) from the European Committee for Standardization (CEN) have been followed to ensure that these samples are compatible with the subsequent application of molecular biology techniques. Briefly, for each sampling, the biofilm was brushed from at least five submerged stones located in the stream's lotic zone (str*etc*h of a few meters to several dozen meters depending on the river size) and fixed in ethanol at a final concentration of at least 70%, which preserves the DNA. Each sample (50 ml) was homogenized and sub-sampled into two batches for morphological and molecular analysis, respectively.

DNA extractions and PCR were carried out as described in Rivera *et al.* (2020). The suspended diatom biofilm (2 ml) was centrifuged for 30 min at 13,000 rpm. Then, we used the DNA extraction kit Macherey-Nagel NucleoSpin® Soil kit (MN-Soil), following the manufacturer's protocol. DNA quantity and quality were checked using a NanoDrop™ 1000 Spectrophotometer (Thermo Fisher Scientific). We amplified a 263 bp barcode of chloroplast DNA, embedded in the *rbcL* gene, which encodes for Ribulose-1,5-bisphosphate carboxylase/oxygenase using diatom specific primers Diat_rbcL_708F_1 (AGGTGAAGTAAAAGGTTCWTACTTAAA), Diat_rbcL_708F_2 (AGGTGAAGTTAAAGGTTCWTAYTTAAA), and Diat_rbcL_708F_3 (AGGTGAAACTAAAGGTTCWTACTTAAA) as forward primers, and Diat_rbcL_R3_1 (CCTTCTAATTTACCWACWACTG) and Diat_rbcL_R3_2 (CCTTCTAATTTACCWACAACAG) as reverse primers (Vasselon *et al.*, 2017). Each DNA extract was amplified in triplicate using equimolar mixes of the three forward and two reverse primers. Half of the P5 (CTTT CCCTACACGACGC TCTTCCGATCT) and P7 (GGAGTTCAGACGTGTGCT CTTCCGATCT) Illumina adapters were included to the 5' part of the *rbcL* forward and reverse primers, respectively. Additionally, blank samples using water of molecular quality were run in parallel to check for potential contaminants introduced during handling the samples in the lab. Amplifications were performed in a final volume of 25 µl following mix and reaction conditions used by Rivera *et al.* (2020), the number of amplification cycles was set to 33 and the conditions of a cycle were as follow: 95 °C – 1 min, 54 °C – 1 min, 72 °C – 1 min. For each sample, DNA from three replicates was extracted and amplified, after which they were pooled, and

50 µl was sent for sequencing on an Illumina MiSeq platform at GetPlage Toulouse, France (https://get.genotoul.fr). The PCR amplicons were purified and used as templates in a second PCR that used Illumina tailed primers targeting the half of P5 and P7 sequences. Finally, all generated PCR amplicons were dual-indexed and pooled into a single tube. The final pool was sequenced using Illumina MiSeq v3 (2×250 bp) for French Guyana, and MiSeq v2 (2×250 bp) for mainland France.

2.4 Bioinformatics

Demultiplexed MiSeq reads were analyzed with the DADA2 pipeline (Callahan *et al.*, 2016) by adapting the settings to diatom metabarcoding sequence data (available on https://github.com/fkeck/DADA2_diatoms_pipeline). First, forward primers (Diat_rbcL_708F_1, Diat_rbcL_708F_2, Diat_rbcL_708F_3) and reverse primers (R3_1, R3_2) established by Vasselon *et al.* (2017) were removed from R1 and R2 reads using cutadapt 2.9 (Martin, 2011). Then the quality profile of the reads was examined and low-quality ends were trimmed by truncating R1 and R2 reads to 200 and 170 nucleotides, respectively. Filter criteria included the removal of ambiguous bases ("N") and limiting the maximum expected error (maxEE) to 2. The DADA2 error model was applied, and reads were then dereplicated into unique sequence units, and ASVs were inferred based on the error models. Paired-end reads were merged into one sequence, followed by the removal of chimera. At the assignment step, two types of annotations were performed using diat.barcode (v12). First, ASVs were assigned to taxonomy using the *assignTaxonomy* function with a minimum bootstrap confidence of 75, In parallel, the same ASVs were assigned to functional and ecological traits (chloroplast number, size class, marine, freshwater, mobile, unattached, attached, adnate, pedunculate, stalk, pad attached to substrate, colonial, non-colonial, mucous tubule colony, chain colony, zig-zag colony, rosette colony, ribbon colony, stellate colony, arbuscular colony, high profile guild, motile guild, planktonic guild, low profile guild) using the same function and confidence threshold.

2.5 Statistical analysis

The congruence between diatom assemblages derived from both taxonomy and traits assignment was assessed using a symmetric co-correspondence (sCoCa) analysis (Alric *et al.*, 2020). Symmetric co-correspondence analysis (sCoCA) is used when the goal is to identify shared patterns between two ecological communities sampled at the same sites, without designating one as a predictor or response. This method extracts common covariance structures symmetrically, making it ideal for exploratory analysis of co-occurring species assemblages.

For the taxonomy-based matrix, only ASVs with assigned taxonomy were used. For the trait-based approach, a separate matrix was constructed for each trait category (*e.g.*, size class, habitat type, guild, chloroplast number, *etc*.), considering only ASVs assigned. Before performing sCoCA, the gene read counts of each ASV were converted to relative abundances by dividing by the total number of reads in each sample. Each of these trait-specific matrices was then analyzed in relation to the taxonomy-based matrix, resulting in a total of 13 independent sCoCA analyses per region (Mainland France and French Guyana).

To test the significance of the global covariation between the two tables, a Monte-Carlo permutation procedure with 9,999 permutations was used. In each permutation, sCoCA (by considering all axes) is reapplied to obtain a value of the covariance between table Y and rtable Y row-permuted X (so that samples are randomized while preserving the relative abundance of individuals). A null distribution was estimated from covariance calculated for the permuted data. The observed covariance is then compared to the distribution obtained under the null hypothesis. The positions of the samples on the ordination axis of each table are then correlated to show the overall level of covariation between them. All statistical analyses were performed with the R software (R Core Team, 2019) and using the *cocorresp* package (Simpson, 2009) for sCoCA. Analyses were conducted separately for the Mainland France dataset and French Guyana dataset.

3 Results

3.1 Content of Diat.barcode v12

3.1.1 Number of species and sequences

Since the setup of Diat.barcode in 2012, the number of *rbcL* sequences has quintupled, increasing from 1225 to 5896 in the latest version (Fig. 1a). Most of the sequences have lengths between 1200 and 1600 bp. A few of them are much shorter (263 bp) and correspond to sequences deposited as part of metabarcoding studies following the protocols of Rimet *et al.* (2018): this protocol describes how to take advantage of high throughput sequencing of natural samples as a source of taxonomic information for reference barcoding libraries. The number of taxa increased from 673 (version v1) to 1651 (version v12) (Fig. 1b), of which 1421 are identified to species.

The hierarchical relationships of families according to their parent taxonomical links are shown along with the distribution of barcodes between major structural types in the diatoms are shown in Figure 2. In the tree, the classification given by Medlin (2011), which has three classes, Mediophyceae, Coscinodiscophyceae and Bacillariophyceae, and the Fragilariophyceae as described in Round *et al.* (1990) was used.

3.1.2 Library completeness for traits

The new version of Diat.barcode showed comprehensive coverage for most of the traits: 96% of the sequences had trait information for ecological guilds (low-profile, motile, high-profile, planktonic), type of habitat (benthic, planktonic, epipelic, episamic, *etc*), mobile and pioneering species, attached and non-attached species, colonial o no colonial species and type of colony (ribbon, filament, mucous tubule), 80% for chloroplasts number and shape (*e.g.* discoid, plate like, *etc*) and only 69% for size classes (based on biovolumes) (Fig. 3).

3.2 Traits distribution in the phylogeny and test of their phylogenetic signal

According to the test of phylogenetic signal based on K.star statistics, ecological guilds—as well as all other

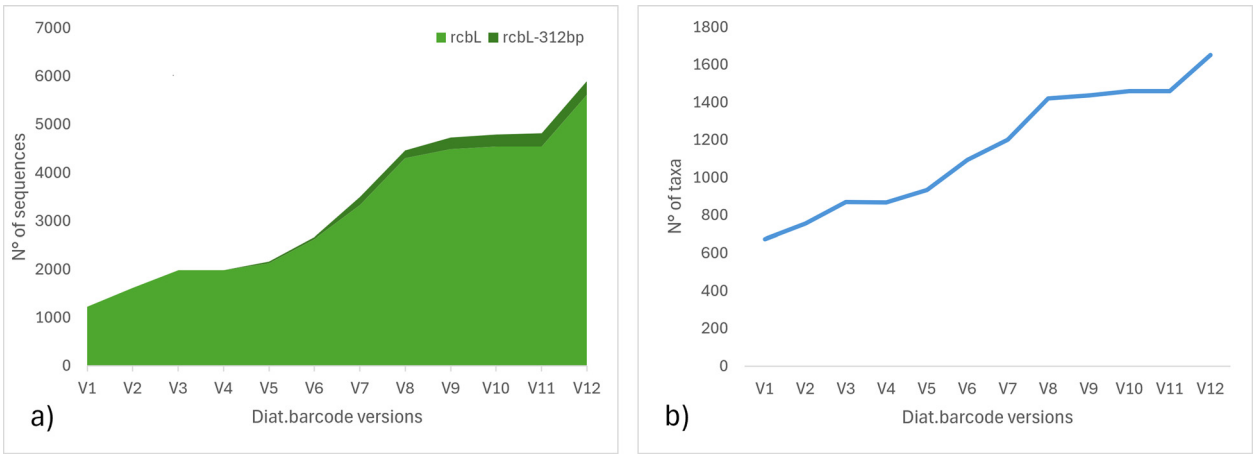

Fig. 1. Content evolution of Diat.barcode reference library across versions. a) Number of rbcL sequences (rcbL) and 312 bp rbcL sequences (rcbL-312bp) corresponding to sequences deposited as part of metabarcoding studies.b) Number of taxa.

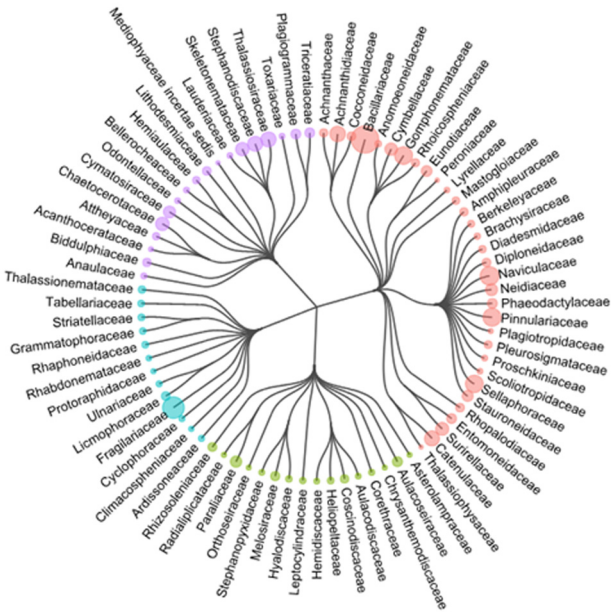

Fig. 2. Taxonomic tree of Diat.barcode v12. Hierarchical relationships of families according to their parent taxonomical links are represented. Size of bubbles gives the number of sequences in the corresponding families. Colors correspond to Classes: purple: Mediophyceae, red: Bacillariophyceae, green: Coscinodiscophyceae, blue: Fragilariophyceae.

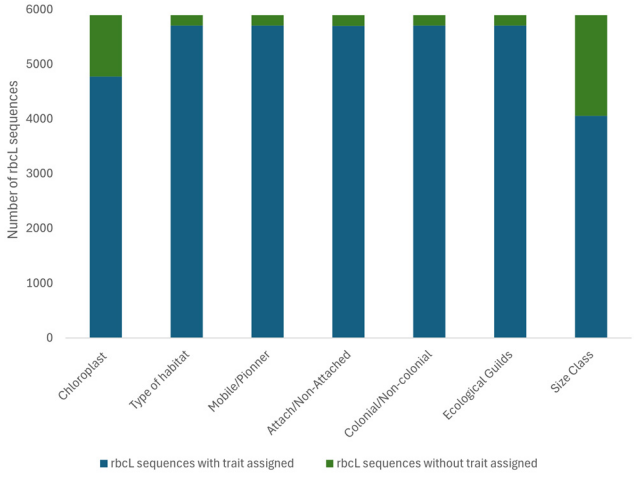

Fig. 3. Number of *rbcL* sequences with available trait data in Diat.barcode (v12) (blue bars) and number of sequences with missing traits in the database (green bars).

traits—show significant phylogenetic signals ($p < 0.001$). This indicates that trait values are conserved across the phylogeny, with phylogenetically neighboring barcodes tending to share similar trait values. As an example of trait distribution in the phylogeny, Figure 4 shows the distribution of ecological guilds in the *rbcL* phylogeny. This distribution is clearly non-random: phylogenetically close sequences tend to belong to the same ecological guilds. Results for the other traits, along with their respective phylogenetic distributions, are presented in Supplementary Figure 1.

3.3 Comparison of a direct assignation to traits vs taxonomical assignation and co-correspondence analysis

In Mainland France, of the 1915 ASVs obtained, 802 were taxonomically assigned, with 767 achieving species level (42%). For trait assignment, an average of 939 ASVs (49%) were successfully assigned (Fig. 5), with the percentage for each individual trait shown in Table 1.

In French Guyana, among the 4437 ASVs obtained, 600 were taxonomically assigned, with 568 achieving species level (12.8%). For trait assignment yielded an average of 1338 ASVs (30%) (Fig. 5), with the percentage for each individual trait shown in Table 1.

The Symmetric Co-Correspondence Analysis (sCoCA) was performed to evaluate the common variance between diatom species and trait assignments in Mainland France and French Guyana. Results based on ecological guilds are

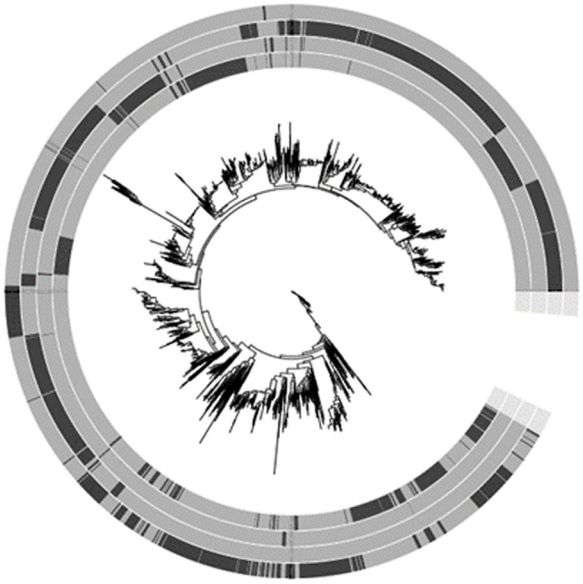

Fig. 4. Ecological guild distribution in the phylogeny (based on *rbcl*). Black color corresponds to a sequence belonging to an ecological guild. From inner to outer circle: high-profile, low profile, motile, planktonic.

presented below, while results for the other traits are included in the supplementary material (Supplementary Tabs. 1 and 2).

In Mainland France, the sCoCA explained 12.86% of the total variation in diatom ecological guilds and 5.46% at the species level, with the first three axes accounting for 30.56% of this common variance (axis 1: 15.29%, axis 2: 8.24%, axis 3: 7.03%; Figs. 6a-b). The correlations between traits and species across these axes were consistently high ($r = 0.93, 0.90, 0.92$). The Monte Carlo permutation test (999 replicates) confirmed the significance of this association ($p = 0.001$). Similar results were obtained for the other trait groups, as shown in Supplementary Fig. 2).

Results for French Guyana were slightly different (Figs.7a-b). The sCoCA explained 17.62% of the variation in ecological guilds and 4.45% at the species level, with the first three axes accounting for 28.76% of the common variance (axis 1: 11.66%, axis 2: 8.72%, axis 3: 8.38%). Correlations between traits and species remained high ($r = 0.89–0.944$). But the Monte Carlo permutation test (999 replicates) yielded non-significant results for aerophilic, epipelic, freshwater, marine, motility and size classes traits. Results for the other traits in French Guyana yielded significant associations ($p < 0.05$) for Benthic, Chloroplast, Colonial, Colony type, Episamic, and Planktonic traits (see Supplementary Fig. 3).

4 Discussion

In this study, we contribute to the implementation of DNA metabarcoding for diatom-based biomonitoring by updating Diat.barcode and increasing the number of curated *rbcL* sequences. Moreover, we incorporated functional and morphological traits into the reference database. A phylogenetic signal was detected for all traits; we can take advantage of this phylogenetic conservation to assign traits directly to environmental sequences in bioinformatic pipelines.

We evaluated the assignment success of both species and traits in two regions with contrasting levels of taxonomic coverage: Mainland France, with good taxonomic representation, and French Guyana, with poor coverage. Our results highlight that direct trait assignment success was higher than taxonomic assignment, especially in the region where the taxonomic coverage of the reference library Diat.barcode was poor, namely French Guyana.

4.1 Bridging taxonomic gaps: the need for expanded molecular libraries and collaborative curation

The continuous expansion of molecular libraries is crucial for improving species-level detection in diatom metabarcoding studies. A previous study found that diatom species widely monitored across many countries tend to have good barcode coverage (*rbcL* and 18S), whereas species thar are restricted to fewer countries often have lower coverage (below 20%) (Weigand *et al.*, 2019).

Despite the increase in the number of sequences and taxa in the new version of Diat.barcode, our results align with these findings, as taxonomic assignment success was markedly lower in French Guyana compared to Mainland France, emphasizing the need for more comprehensive sequencing efforts in tropical and understudied regions. However, the joint effort of experts from several countries to curate this reference library demonstrates the potential for collaborative initiatives to bridge these gaps.

4.2 Biological and ecological trait assignment success and conservation of traits in the phylogeny

Biological and ecological traits provide an alternative way of interpreting diatom assemblages, offering an interesting approach for functional ecology, therefore going beyond taxonomic composition. Previous studies have demonstrated that functional traits can be effectively used to detect environmental change and anthropogenic disturbance (*e.g.* Passy, 2007; Berthon *et al.*, 2011; Soininen *et al.*, 2016). The advantage of using traits as indicators is that they represent a measurable adaptation strategy under particular circumstances; thus, their utility is well justified (Tapolczai *et al.*, 2016). Moreover, trait responses have been reported to be independent of ecoregions (Soininen *et al.*, 2016), making trait-based approaches a way to compare environmental pressures in diverse ecosystems with regionally distinct diatom assemblages. Additionally, trait-based indices have proven to be effective in tropical rivers, where taxonomic knowledge is more limited (Tapolczai *et al.*, 2017), highlighting their potential for biomonitoring in regions with less-explored diatom flora.

In our study, we found that the use of traits, in general, enabled the assignment of more ASVs than use the of taxonomy, especially in French Guyana where many species remain unknown and therefore it is impossible to assign many ASVs. This suggests that trait-based approaches can mitigate the limitations of incomplete taxonomic reference databases, making them particularly useful in biomonitoring programs.

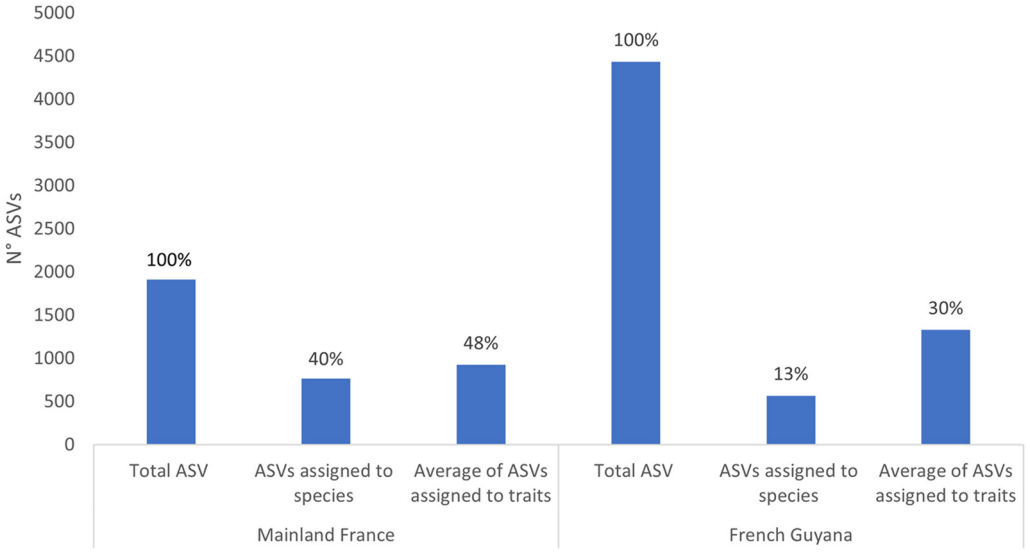

Fig. 5. Number of total ASVs, number of ASV assigned to species, and to traits for diatom communities in Mainland France and French Guyana. Percentages are shown as labels.

Table 1. Summary of the traits species associations and the proportion of ASVs assigned to species and traits. Montecarlo Permutation test, significant differences * $0.001 < p < 0.05$, *** $p < 0.001$, NS not significant. The proportion of ASVs assigned to species was 42% in Mainland France and 13% in French Guyana.

	Mainland france		French guyana	
Trait	Assigned to trait	Monte carlo test	Assigned to trait	Monte carlo test
Aerophilic	67%	***	48%	NS
Benthic	84%	***	78%	*
Chloroplast	31%	***	8%	*
Colonial	31%	***	8%	*
Colony type	31%	***	8%	*
Epipelic	69%	***	51%	NS
Episamic	78%	***	71%	*
Freshwater	31%	***	8%	NS
Guilds	30%	***	8%	*
Marine	31%	***	8%	NS
Motility	31%	***	8%	NS
Planktonic	82%	***	75%	*
Size class	31%	***	10%	NS

Furthermore, our results suggest that when taxonomic assignment is weaker than trait assignment, the correlation between taxonomic and trait-based datasets is lost (non-significant Monte-Carlo test), as shown for French Guyana. A possible explanation is that the species dataset retains only a limited portion of the original information (in French Guyana, for example, 87% of ASVs were not assigned), and therefore only captures a small part of the ecological information. This highlights the value of employing traits. Besides this, Tapolczai *et al.* (2016) suggested that the function of an organism in an ecosystem depends on its morphological and physiological traits, which determine its adaptive strategies within a habitat. Given the high redundancy at the species level (Kelly, 2013), grouping taxa based on shared ecological traits (Salmaso *et al.*, 2015) can provide a complementary framework, as phylogenetically neighboring barcodes often share the same trait values. In our analysis, we also observed a progressive decrease in assignment success from higher to lower taxonomic ranks. Since higher taxonomic ranks are relatively uninformative for ecological studies, the trait-based approach is of real interest (Tapolczai *et al.*, 2025). Nonetheless, it is important to note that trait assignment success varied: while some traits (*e.g.*, type of habitat: benthic, planktonic, episamic, epipelic, aerophilous) were well represented, others (*e.g.*, size class, chloroplast number) showed assignment success comparable to that of species-level taxonomy. Overall, trait-based approaches can broaden the perspectives gained from species-based methods, particularly in contexts of limited taxonomic knowledge.

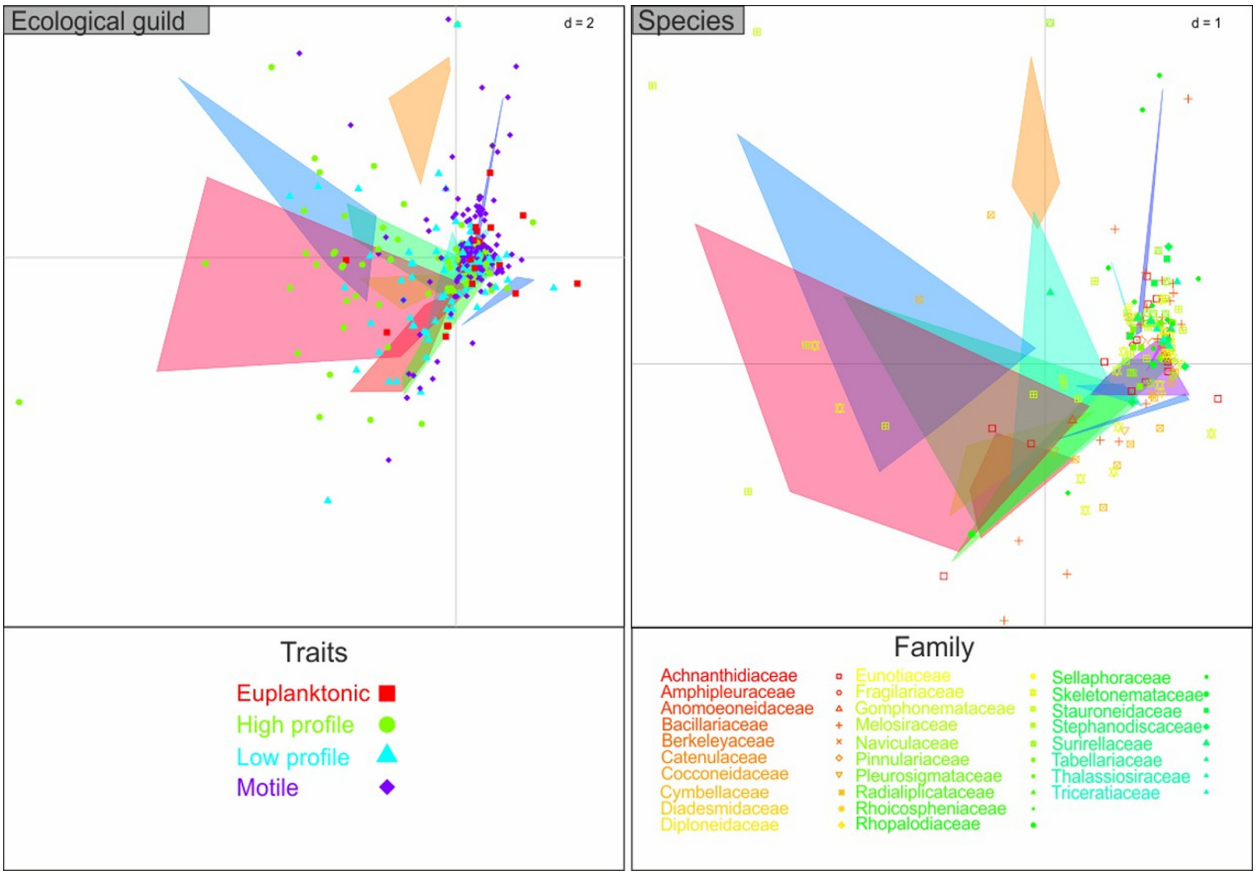

Fig. 6. Ordination biplots for Mainland France obtained from symmetric co-correspondence analysis (sCoCA). Ecological guilds are displayed in the left panel and species in the right panel. To improve the interpretation of the plots, species are colored based on their family and polygons represent the sampling sites from which they were sampled, regardless of their identification.

Phylogenetic niche conservatism refers to the tendency of lineages to retain their niche-related traits through speciation events (Crisp & Cook, 2012), while phylogenetic signal is defined as "the tendency for (phylogenetically) related species to resemble each other more than they resemble species drawn at random from the tree" (Blomberg and Garland, 2002). In our study, the significant phylogenetic signal detected across all traits explains the higher success in trait assignment compared to species assignment. Closely related barcodes tended to share similar trait values, which aligns with previous findings by Keck *et al.* (2016), who reported significant phylogenetic structuring of ecological traits in diatoms—particularly those related to nutrient and organic matter preferences—suggesting that evolutionary history constrains ecological functions. Recently, Tapolczai *et al.* (2025) also found that several phytoplankton traits related to habitat preference are well conserved across the phylogeny. Consistently, we observed that genetically close taxa tend to share similar habitat and ecological pReferences

4.3 Implications for biomonitoring and ecological assessments

The growing application of molecular tools in biomonitoring necessitates robust reference libraries that integrate both taxonomic and functional information and the approach presented here contributes to bridging the inevitable gaps that persist in such resources. The enhanced trait coverage in Diat.barcode facilitates the use of metabarcoding for ecological assessments, particularly in regions where traditional taxonomic expertise is limited. Recent studies have highlighted the advantages of trait-based biomonitoring, demonstrating their applicability in detecting shifts in ecosystem health (Apothéloz-Perret-Gentil *et al.*, 2017; Rivera *et al.*, 2020). Our results indicate that trait-based assignment methods can improve the interpretability of metabarcoding data by providing ecologically meaningful classifications, even when taxonomic knowledge of the area is poor.

These results highlight that trait-based information can be especially valuable in contexts where taxonomy-free and species-based approaches face important limitations. The taxonomy-free approach requires access to large environmental datasets to estimate optimum and tolerance for each ASV, which is not trivial because environmental datasets are often lacking. Moreover, the spatial and geographical origin of samples influences the estimation of ASVs ecological preferenda, limiting the transferability of these estimations between regions with different environmental conditions (different climates, geologies). Similarly, species-based approaches depend heavily on the species considered: while abundant ecological information exists for some easily identifiable species, most species lack sufficient ecological

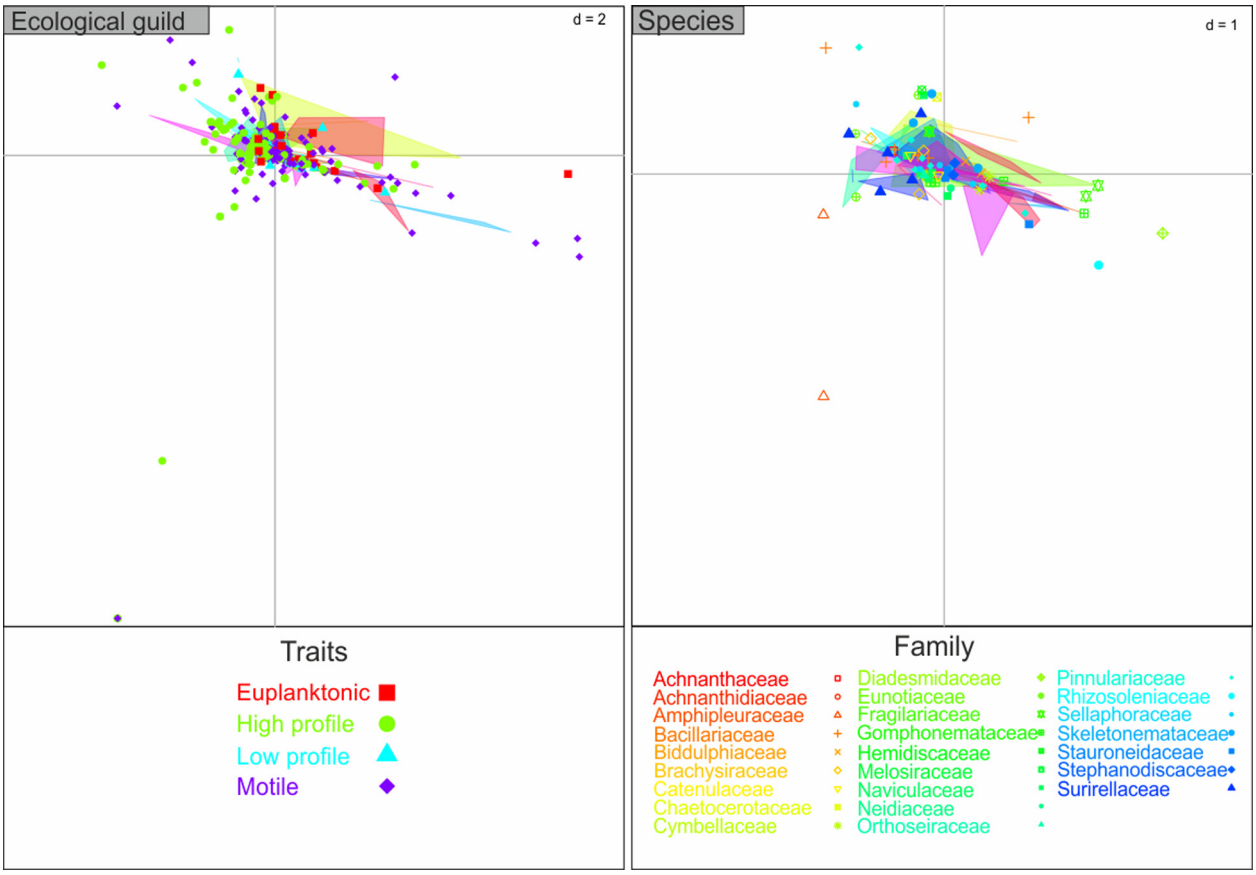

Fig. 7. Ordination biplots for French Guyana obtained from symmetric co-correspondence analysis (sCoCA). Ecological guilds are displayed in the left panel and species in the right panel. To improve the interpretation of the plots, species are colored based on their family and polygons represent the sampling sites from which they were sampled, regardless of their identification.

data, partly due to difficulties in morphological differentiation (Tapolczai *et al.*, 2025). In this context, trait-based approaches offer a complementary and functionally relevant framework that can enhance biomonitoring, particularly in regions with limited taxonomic knowledge or ecological data.

Our objective in applying sCoCA was not to measure diatom–environment relationships directly, but to assess whether trait-based assignments captured ecological patterns comparable to those revealed by taxonomic assignments. In French Guyana, the weaker co-structure likely reflects the limited success of taxonomic assignment, which may have resulted in the loss of some ecological information. Nevertheless, trait-based assignments were able to retain key ecological signals, highlighting their utility even when taxonomic resolution is low.

4.4 The wider use of Diat.barcode

In this paper, we have focused on freshwaters and the application of metabarcoding in regions for which there are as yet few reference sequences, so that many ASVs cannot be assigned to named species. Our examples used the primers developed by Vasselon *et al.* (2017), which have now been used in many river biomonitoring studies worldwide, including a major study of streams and rivers across the whole of the U.S. by Smucker *et al.* (2024). However, Diat.barcode aims to curate all published *rbcL* sequences for diatoms, whether or not the marker used is the 263-bp region designed by Vasselon *et al.* (2017). Several other markers have been developed within the *rbcL* gene. Some of these overlap the Vasselon *et al.* region partially (Wolf & Vis, 2020; Liu *et al.*, 2020) or completely (Kelly *et al.*, 2020), while others do not (An *et al.*, 2018); a short barcode, lying entirely within the Vasselon *et al.* region has been used extensively for studying subfossil sediment assemblages (*e.g.* Stoof-Leichsenring *et al.*, 2012, 2020). The taxonomic resolution of these different markers varies (*e.g.* Pérez-Burillo *et al.*, 2022b), but the basic resource needed for taxonomic assignment in each case is a curated reference database, which can be supplied by Diat.barcode. Furthermore, although diatom metabarcoding has been used so far mostly for surveys and biomonitoring of freshwater rivers and lakes, marine habitats are increasingly being studied, both oceanic (Gibb *et al.*, 2024) and coastal (Pérez-Burillo *et al.*, 2022a; Girard *et al.*, 2025); again, we emphasize that Diat.barcode is intended as a general support for diatom taxonomic assignment, whatever the habitat, though our experience is that the proportion of ASVs assignable to species in marine environments is considerably lower than for freshwaters.

5 Conclusions

In conclusion, the newest Diat.barcode expands molecular diatom research by filling taxonomic gaps and by providing a richer set of functional traits (notably habitat type and ecological guild). Our results show that trait-based assignment can complement species-based approaches, providing ecological meaningful interpretability that help capture key community patterns even in regions with limited taxonomic resolution.

Nonetheless, the weaker co-structure observed in French Guyana reminds us that trait libraries must keep growing, especially for tropical lineages, and that trait definitions need rigorous standardisation. Continued curation of Diat.barcode, together with methodological advances such as hybrid taxon-plus-trait pipelines and better environmental metadata integration, will consolidate the role of molecular tools in freshwater and marine ecosystem assessment.

Acknowledgments

Samples were carried out in the framework of the French river monitoring and the OFB (Office Français de la Biodiversité), the Office de l'Eau de Guyane, the Water agencies, as well as the French ministry in charge of environment protection are thanked for helping in providing us the samples for molecular analyses. The study was supported by Aquaref (Etude B1.15) and the HORIZON project No. 101079234 BIOLAWEB (Boosting Institute of Chemistry Technology and Metallurgy in Water Biomonitoring) coordination and support actions funded by the European Union. CEW thanks the Luxembourg National Research Fund (FNR) grant C19/SR/13680552/DISCO. The study was also supported by ANPCYT-PICT 2021-00372 and PICT 2019-03621.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

Data presented in this study is available in NICOLOSI GELIS, Maria Mercedes; COCHERO, Joaquín; VIOLLAZ, Laurine; BRIAND, Jean-François; BARRY-MARTINET, Raphaëlle; CHONOVA, Teofana; GASSIOLE, Gilles; KAHLERT, Maria; KECK, François; KELLY, Martyn; KOCHOSKA, Hristina; MANN, David; PFANNKUCHEN, Martin; TROBAJO, Rosa; VASSELON, Valentin; VIDAKOVIC, Danijela; WETZEL, Carlos; ZIMMERMANN, Jonas; RIMET, Frédéric, 2025, "Taxonomic or trait assignation of environmental sequences in regions with a poor taxonomic knowledge: case of river diatom metabarcoding with a new version of the annotated reference library Diat.barcode – supplementary data", https://doi.org/10.57745/IRRMXH, Recherche Data Gouv.

Supplementary material

Supplementary Table 1. Results of the Symmetric Co-Correspondence Analysis (sCoCA) assessing the shared variance between diatom species and trait assignments in mainland France.

Supplementary Table 2. Results of the Symmetric Co-Correspondence Analysis (sCoCA) assessing the shared variance between diatom species and trait assignments in French Guyana.

Supplementary Figure 1. Traits distribution in the phylogeny (based on rbcl).

Supplementary Figure 2. Ordination biplots for French Guyana obtained from symmetric co-correspondence analysis (sCoCA). Aerophilous (red solid square) /non aerophilous (light blue solid circles) are displayed in the left panel and species in the right panel. To improve the interpretation of the plots, species are colored based on their family and polygons represent the sampling sites from which they were sampled, regardless of their identification.

Supplementary Figure 3. Ordination biplots for Mainland France obtained from symmetric co-correspondence analysis (sCoCA). **Benthic** (red solid square) /**non benthic** (light blue solid circles) are displayed in the left panel and species in the right panel. To improve the interpretation of the plots, species are colored based on their family and polygons represent the sampling sites from which they were sampled, regardless of their identification.

The Supplementary Material is available at https://www.limnology-journal.org/10.1051/limn/2025009/olm.

References

Alric B, Ter Braak CJ, Desdevises Y, Lebredonchel H, Dray S. 2020. Investigating microbial associations from sequencing survey data with co-correspondence analysis. *Mol Ecol Resour* 20: 468–480.

An SM, Choi DH, Lee H, Lee JH, Noh JH. 2018. Next-generation sequencing reveals the diversity of benthic diatoms in tidal flats. *Algae* 33: 167–180.

Apothéloz-Perret-Gentil L, Cordonier A, Straub F, Iseli J, Esling P, Pawlowski J. 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Mol Ecol Resour* 17: 1231–1242.

Bailet B, Bouchez A, Franc A, Frigerio J-M., Keck F, Karjalainen S-M., Rimet F, Schneider S, Kahlert M. 2019. Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe freshwater and consequences for ecological status. *Metabarcoding Metagenom* 3. https://doi.org/10.3897/mbmg.3.34002.

Ballesteros I, Castillejo P, Haro AP, Montes CC, Heinrich C, Lobo EA. 2020. Genetic barcoding of Ecuadorian epilithic diatom species suitable as water quality bioindicators. *C R Biol* 343: 41–52.

Baricevic A, Chardon C, Kahlert M, Karjalainen SM, Pfannkuchen DM, Pfannkuchen M, Rimet F, Tankovic MS, Trobajo R, Vasselon V, Zimmermann J, Bouchez A. 2022. Recommendations for the preservation of environmental samples in diatom metabarcoding studies. *Metabarcoding Metagenom* 6:e85844.

Berthon V, Bouchez A, Rimet F. 2011. Use of diatom life-forms and ecological guilds to assess pollution in rivers: case study of south-eastern French rivers. *Hydrobiologia* 673: 259–271.

Blomberg SP, Garland T. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *J Evol Biol* 15: 899–910.

Blomberg SP, Garland T, Ives AR, 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57: 717–745.

Canino A, Bouchez A, Laplace-Treyture C, Domaizon I, Rimet F. 2021. Phytool, a ShinyApp to homogenise taxonomy of freshwater microalgae from DNA barcodes and microscopic observations. *Metabarcoding Metagenom* 5: e74096.

CEN. 2014. EN 14407: Water quality – Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. EN 14407:2014. Geneva: Comité European de Normalisation.

CEN. 2018. Water Quality – CEN/TR 17244 – Technical Report for the Management of Diatom Barcodes. CEN Stand.

CEN. 2025. EN 13946:2025: Water quality – Guidance standard for the routine sampling and preparation of benthic diatoms from rivers and lakes. CEN Stand.

Crisp MD, Cook LG. 2012. Phylogenetic niche conservatism: what are the underlying evolutionary and ecological causes? *New Phytol* 196: 681–694.

Cox E. 1996. Identification of Freshwater Diatoms from Live Material. London: Chapman and Hall, 1; 328 p.

del Campo J, Kolisko M, Boscaro V, Santoferrara LF, Nenarokov S, Massana R, Guillou L, Simpson A, Berney C, de Vargas C. 2018. EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS Biol* 16: e2005849.

Djemiel C, Plassard D, Terrat S, Crouzet O, Sauze J, Mondy S, Nowak V, Wingate L, Ogée G, Maron PA. 2020. µgreen-db: a reference database for the 23S rRNA gene of eukaryotic plastids and cyanobacteria. *Sci Rep* 10: 5915.

Gauvin P, Eme D, Domaizon I, Rimet F. 2025. An annotated reference library for supporting DNA metabarcoding analysis of aquatic macroinvertebrates in French freshwater environments. *Metabarcoding Metagenom* 9: e137772.

Gibb RLA, Botha DK, Venkatachalam S, Bizani M, Bornman TG, Dorrington RA. 2024. DNA metabarcoding reveals distinct bacterial and phytoplankton assemblages in the Agulhas Current and the adjacent coastal shelf. *Limnol Oceanogr* 69: 2760–2774.

Girard EB, Pratama AM, del Rio-Hortega L, Volkenandt S, Macher JN, Renema W. 2025. Coastal eutrophication transforms shallow micro-benthic reef communities. *Sci Total Environ* 961: 178252.

Gottschalk S, Kahlert M. 2012. Shifts in taxonomical and guild composition of littoral diatom assemblages along environmental gradients. *Hydrobiologia* 694: 41–56.

Guiry MD, Guiry GM. 2025. AlgaeBase. World-Wide Electronic Publication, University of Galway. https://www.algaebase.org

Hamilton PB, Lefebvre KE, Bull RD. 2015. Single cell PCR amplification of diatoms using fresh and preserved samples. *Front Microbiol* 6: 1084.

Hering D, Johnson RK, Kramm S, Schmutz S, Szoszkiewicz K, Verdonschot PFM. 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. *Freshw Biol* 51: 1757–1785.

Irannia ZB, Chen T. 2016. TACO: Taxonomic prediction of unknown OTUs through OTU co-abundance networks. *Quant Biol* 4: 149–158.

Keck F, Rimet F, Franc A, Bouchez A. 2016. Phylogenetic signal in diatom ecology: perspectives for aquatic ecosystems biomonitoring. *Ecol Appl* 26: 861–872.

Keck F, Rimet F, Bouchez A, Franc A. 2016. phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol Evol* 6: 2774–2780.

Keck F, Blackman RC, Bossart R, Brantschen J, Couton M, Hürlemann S, Altermatt F. 2022. Meta-analysis shows both congruence and complementarity of DNA and eDNA metabarcoding to traditional methods for biological community assessment. *Mol Ecol* 31: 1820–1835.

Keck F, Couton M, Altermatt F. 2023. Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Mol Ecol Resour* 23: 742–755.

Keck F, Altermatt F. 2023. Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. *Mol Ecol Resour* 23: 511–518.

Kelly M. 2013. Data rich, information poor? Phytobenthos assessment and the Water Framework Directive. *Eur J Phycol* 48: 437–450.

Kelly M. 2019. Adapting the (fast-moving) world of molecular ecology to the (slow-moving) world of environmental regulation: lessons from the UK diatom metabarcoding exercise. *Metabarcoding Metagenom* 3:e39041.

Kelly MG, Juggins S, Mann DG, Sato S, Glover R, Boonham N, Sapp M, Lewis E, Hany U, Kille P, Jones T, Walsh K. 2020. Development of a novel metric for evaluating diatom assemblages in rivers using DNA metabarcoding. *Ecol Indic* 118: 106725.

Kermarrec L, Bouchez A, Rimet F, Humbert J-F. 2013. First evidence of the existence of semi-cryptic species and of a phylogeographic structure in the *Gomphonema parvulum* complex (Bacillariophyta). *Protist* 164: 686–705.

Khan-Bureau DA, Morales EA, Ector L, Beauchene MS, Lewis LA. 2016. Characterization of a new species in the genus *Didymosphenia* and of *Cymbella janischii* (Bacillariophyta) from Connecticut, USA. *Eur J Phycol* 51: 203–216.

Krammer K, Lange-Bertalot H. 1986. Bacillariophyceae, Teil 1: Naviculaceae, 876 pp.

Krammer K, Lange-Bertalot H. 1988. Bacillariophyceae, Teil 2: Bacillariaceae, Epithemiaceae, Surirellaceae, 596 pp.

Krammer K, Lange-Bertalot H. 1991a. Bacillariophyceae, Teil 3: Centrales, Fragilariaceae, Eunotiaceae, 576 pp.

Krammer K, Lange-Bertalot H. 1991b. Bacillariophyceae, Teil 4: Achnanthaceae. Kritische Ergänzungen zu Navicula (Lineolatae) und Gomphonema. *Gesamtliteraturverzeichnis Teil* 14, 437 pp.

Kruk C, Huszar VL, Peeters ET, Bonilla S, Costa L, Lürling M, Scheffer M. 2010. A morphological classification capturing functional variation in phytoplankton. *Freshw Biol* 55: 614–627.

Laamanen T, Norros V, Vihervaara P, Jerney J, Kortelainen P, Kujala K, Lambert S, Mäyrä J, Nikula L, Palmroos I, Tolkkinen M, Vuorio K, Meissner K. 2025. Technology readiness level of biodiversity monitoring with molecular methods – where are we on the road to routine implementation? *Metabarcoding Metagenom* 9: e130834.

Leese F, Altermatt F, Bouchez A, Ekrem T, Hering D, Meissner K, Mergen P, Pawlowski J, Piggott J, Rimet F, Steinke D, Taberlet P, Weigand A, Abarenkov K, Beja P, Bervoets L, Björnsdóttir S, Boets P, Boggero A, Bones A, Borja Á, Bruce K, Bursić V, Carlsson J, Čiampor F, Čiamporová-Zatovičová Z, Coissac E, Costa F, Costache M, Creer S, Csabai Z, Deiner K, DelValls Á, Drakare S, Duarte S, Eleršek T, Fazi S, Fišer C, Flot J-F., Fonseca V, Fontaneto D, Grabowski M, Graf W, GuÐbrandsson J, Hellström M, Hershkovitz Y, Hollingsworth P, Japoshvili B, Jones J, Kahlert M, Stroil BK, Kasapidis P, Kelly M, Kelly-Quinn M, Keskin E, Kõljalg U, Ljubešić Z, Maček I, Mächler E, Mahon A, Marečková M, Mejdandzic M, Mircheva G, Montagna M, Moritz C, Mulk V, Naumoski A, Navodaru I, Padisák J, Pálsson S, Panksep K, Penev L, Petrusek A, Pfannkuchen M, Primmer C, Rinkevich B, Rotter A, Schmidt-Kloiber A, Segurado P, Speksnijder A, Stoev P, Strand M, Šulčius S, Sundberg P, Traugott M, Tsigenopoulos C, Turon X, Valentini A, van der Hoorn F B, Várbíró G, Hadjilyra VM, Viguri J,

Vitonytė I, Vogler A, Vrålstad T, Wägele W, Wenne R, Winding A, Woodward G, Zegura B, Zimmermann J. 2016. DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Res Ideas Outcomes* 2:e11321.

Li F, Zhang Y, Altermatt F, Zhang X, Cai Y, Yang Z. 2022. Gap analysis for DNA-based biomonitoring of aquatic ecosystems in China. *Ecol Indic* 137: 108732.

Liu M, Zhao Y, Sun Y, Wu P, Zhou S, Ren L. 2020. Diatom DNA barcodes for forensic discrimination of drowning incidents. *FEMS Microbiol Lett* 367: fnaa 145.

Marques V, Milhau T, Albouy C, Dejean T, Manel S, Mouillot D, Juhel JB. 2021. GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Divers Distrib* 27: 1880–1892.

Medlin LK. 2011. A review of the evolution of the diatoms from the origin of the lineage to their populations. In: Seckbach J. Kociolek J.P. (eds.), The Diatom World, Springer Science + Business Media B. V. pp. 93–118.

MIGALE. 2020. Migale bioinformatics facility. INRAE. https://doi.org/10.15454/1.5572390655343293 E12.

Murali A, Bhargava A, Wright ES. 2018. IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 6: 140.

Passy SI. 2007. Diatom ecological guilds display distinct and predictable behavior along nutrient and disturbance gradients in running waters. *Aquat Bot* 86: 171–178.

Pérez-Burillo J, Trobajo R, Vasselon V, Rimet F, Bouchez A, Mann DG. 2020. Evaluation and sensitivity analysis of diatom DNA metabarcoding for WFD bioassessment of Mediterranean rivers. *Sci Total Environ* 727: 138445.

Pérez-Burillo J, Valoti G, Witkowski A, Prado P, Mann DG, Trobajo R. 2022a. Assessment of marine benthic diatom communities: insights from a combined morphological-metabarcoding approach in Mediterranean shallow coastal waters. *Mar Pollut Bull* 174: 113183.

Pérez-Burillo J, Mann DG, Trobajo R. 2022b. Evaluation of two short and similar *rbcL* markers for diatom metabarcoding of environmental samples: effects on biomonitoring assessment and species resolution. *Chemosphere* 307: 135933.

RCore Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rimet F. 2012. Recent views on river pollution and diatoms. *Hydrobiologia* 683: 1–24.

Rimet F, Bouchez A. 2012. Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. *Knowl Manag Aquat Ecosyst* 406: 1–14.

Rimet F, Bouchez A, Montuelle B. 2015. Benthic diatoms and phytoplankton to assess nutrients in a large lake: Complementarity of their use in Lake Geneva (France-Switzerland). *Ecol Indic* 53: 231–239.

Rimet F, Abarca N, Bouchez A, Kusber WH, Jahn R, Kahlert M, et al. 2018. The potential of high-throughput sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. *Fottea* 18.

Rimet F, Gusev E, Kahlert M, Kelly M, Kulikovskiy M, Maltsev Y, Mann D, Pfannkuchen M, Trobajo R, Vasselon V, Zimmermann J, Bouchez A. 2019. Diat.barcode, an open-access curated barcode library for diatoms. *Sci Rep* 9: 15116.

Rivera SF, Vasselon V, Bouchez A, Rimet F. 2020. Diatom metabarcoding applied to large scale monitoring networks: Optimization of bioinformatics strategies using Mothur software. *Ecological indicators* 109: 105775.

Round FE, Crawford RM, Mann DG. 1990. The Diatoms: Biology and Morphology of the Genera. Cambridge University Press, Vol. 747.

Salmaso N, Padisák J. 2007. Morpho-functional groups and phytoplankton development in two deep lakes (Lake Garda, Italy and Lake Stechlin, Germany). *Hydrobiologia* 578: 97–112.

Salmaso N, Naselli-Flores L, Padisák J. 2015. Functional classifications and their application in phytoplankton ecology. *Freshw Biol* 60: 603–619.

Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I 2020. GenBank. *Nucleic Acids Res* 48: D84– D86.

Simpson GL. 2009. Cocorresp: Co-correspondence analysis ordination methods. (R package version 0.4-0). Retrieved from http://cran.r-project.org/package=cocorresp

Silvestro D, Michalak I . 2012. raxmlGUI: a graphical front-end for RAxML. *Org Divers Evol* 12: 335–337.

Smucker NJ, Pilgrim EM, Nietch CT, Gains-Germain L, Carpenter C, Darling JA, Yuan LL, Mitchell RM, Pollard AI. 2024. Using DNA metabarcoding to characterize national scale diatom-environment relationships and to develop indicators in streams and rivers of the United States. *Sci Total Environ* 939: 173502.

Spaulding SA 2021. Diatoms.org: supporting taxonomists, connecting communities. *Diatom Res* 36: 291–304.

Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stenger-Kovács C, Körmendi K, Lengyel E, Abonyi A, Hajnal É, Szabó B, Padisák J. 2018. Expanding the trait-based concept of benthic diatoms: development of trait- and species-based indices for conductivity as the master variable of ecological status in continental saline lakes. *Ecol Indic* 95: 63–74.

Stoof-Leichsenring KR, Epp LS, Trauth MH, Tiedemann R. 2012. Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. *Mol Ecol* 21: 1918–1930.

Stoof-Leichsenring KR, Dulias K, Biskaborn BK, Pestryakova LA, Herzschuh U. 2020. Lake-depth related pattern of genetic and morphological diatom diversity in boreal Lake Bolshoe Toko, Eastern Siberia. *PLoS One* 15:e0230284.

Soininen J, Jamoneau A, Rosebery J, Passy S I. 2016. Global patterns and drivers of species and trait composition in diatoms. *Global ecology and biogeography* 25: 940–950.

Tapolczai K, Bouchez A, Stenger-Kovács C, Padisák J, Rimet F. 2016. Trait-based ecological classifications for benthic algae: review and perspectives. *Hydrobiologia* 776: 1–17.

Tapolczai K, Bouchez A, Stenger-Kovács C, Padisák J, Rimet F. 2017. Taxonomy-or trait-based ecological assessment for tropical rivers? Case study on benthic diatoms in Mayotte island (France, Indian Ocean). *Sci Total Environ* 607: 1293–1303.

Tapolczai K, Keck F, Bouchez A, Rimet F, Kahlert M, Vasselon V. 2019. Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. *Front Ecol Evol* 7: 1–15.

Tapolczai K, Rimet F, Ćirić M, Ballot A, Laplace-Treyture C, Alric B. 2025. A novel framework for phytoplankton biomonitoring: Trait assignment of 23S rRNA sequences. *Ecol Indic* 173: 113361.

Vasselon V, Rimet F, Tapolczai K, Bouchez A. 2017. Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). *Ecol Indic* 82: 1–12.

Vasselon V, Bouchez A, Rimet F, Jacquet S, Trobajo R, Corniquel M, et al. 2018. Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods Ecol Evol* 9: 1060–1069.

Vasselon V, Rivera SF, Ács É, Almeida SB, Andree KB, Apothéloz-Perret-Gentil L, Bailet B, Baričević A, Beentjes KK, Bettig J, Paix B. 2025. Proficiency testing and cross-laboratory method comparison to support standardisation of diatom DNA metabarcoding for freshwater biomonitoring. *Metabarcoding Metagenom* 9: e133264.

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Ekrem T. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: gap-analysis and recommendations for future work. *Sci Total Environ* 678: 499–524.

Wolf DI, Vis ML. 2020. Stream algal biofilm community diversity along an acid mine drainage recovery gradient using multimarker metabarcoding. *J Phycol* 56: 11–22.

Zorzal-Almeida S, Soininen J, Bini LM, Bicudo DC. 2017. Local environment and connectivity are the main drivers of diatom species composition and trait variation in a set of tropical reservoirs. *Freshw Biol* 62: 1551–1563.

Cite this article as: Nicolosi Gelis MM, Barry-Martinet R, Briand J-F, Chonova T, Cochero J, Gassiole G, Kahlert M, Karthick B, Keck F, Kelly M, Kochoska H, Mann D, Nayak P, Pfannkuchen M, Servulo T, Trobajo R, Vasselon V, Vidakovic D, Viollaz L, Wetzel C, Zimmermann J.Rimet F. 2025. Assigning taxonomy and traits to DNA sequences of river diatoms in a region with limited taxonomic knowledge using the updated and annotated reference library Diat.barcode. *Int. J. Lim.* 61: 10: https://doi.org/10.1051/limn/2025009