Fottea, Olomouc, 18(1): 37–54, 2018
DOI: 10.5507/fot.2017.013

The potential of High–Throughput Sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes

Frédéric Rimet[1,2*], Nelida Abarca[3], Agnès Bouchez[1,2], Wolf–Henning Kusber[3], Regine Jahn[3], Maria Kahlert[4], François Keck[4], Martyn G. Kelly[5], David. G. Mann[6], André Piuz[7], Rosa Trobajo[8], Kalman Tapolczai[1,2], Valentin Vasselon[1,2] & Jonas Zimmermann[3]

[1] *INRA – UMR Carrtel, FR–74200 Thonon–les–Bains, France;* [*] *Corresponding author e–mail: frederic.rimet@inra.fr*
[2] *Université Savoie Mont Blanc, UMR, 73000 Carrtel Chambéry, France*
[3] *Botanischer Garten und Botanisches Museum Berlin–Dahlem, Freie Universität Berlin, Königin–Luise–Str. 6–8, 14195 Berlin, Germany*
[4] *Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, PO Box 7050, SE– 750 07 Uppsala, Sweden*
[5] *Bowburn Consultancy, 11 Monteigne Drive, Bowburn, Durham DH6 5QB, UK*
[6] *Royal Botanic Garden Edinburgh, Edinburgh, EH3 5LR, Scotland, UK*
[7] *Muséum d'Histoire Naturelle, Route de Malagnou 1, Case postale 6434, CH–1211 Genève 6, Switzerland*
[8] *Aquatic Ecosystems, Institute for Food and Agricultural Research and Technology (IRTA), Crta de Poble Nou Km 5.5, Sant Carles de la Ràpita, Catalonia, Spain*

Abstract: Diatoms are used routinely to assess pollution level in rivers and lakes. Current methods are based on identification by light microscopy, which is laborious. An alternative is to identify species based on short DNA fragments and High–Throughput Sequencing (HTS). However a potential limitation is the incomplete coverage of species in reference barcode libraries. Usually these libraries are compiled by isolating cells, before culturing and sequencing them, which is tedious and often unsuccessful. Here we propose the use of *rbc*L sequences from environmental samples analysed by HTS. We set several criteria to ensure good sequence quality and correspondence with the target species observed in microscopy: the sequence needed to be abundant in the sample, and with no insertions nor deletions or stop codon, phylogenetic neighbour taxa had to correspond to neighbour taxonomic taxa expected from morphological observations. Four species from tropical rivers are given as examples, including one that is new to science.

Key words: algae, Bacillariophyta, biomonitoring, data traceability, DNA barcoding, eDNA, ecosystem assessment, metabarcoding, pollution, Water Framework Directive

Introduction

Human activities have had an impact on the environment and in particular on freshwater ecosystems for a long time. Increased fish mortality in the Rhine and Thames in the late 17th and 18th centuries drew attention to the impact of pollution on aquatic ecosystem health (Markert et al. 2003) and, by the mid–19th century, several authors had observed that the composition of microscopic organisms in polluted aquatic ecosystems was different to that in unpolluted habitats (Cohn 1853). In 1908, Kolkwitz and Marsson demonstrated a clear relationship between water quality and organisms, including microalgae, in freshwaters (Kolkwitz & Marsson 1908). Microalgae are often the dominant primary producers of aquatic ecosystems. They display huge taxonomic diversity, and diatoms alone are estimated to have over 100,000 extant species (Mann & Vanormelingen 2013). This diversity, coupled with a high sensitivity to their chemical environment and wide distribution makes them excellent ecological indicators (Stevenson 2014). In the 1950s several authors (e.g. Hustedt 1957; Zelinka & Marvan 1961; Butcher 1947) started to use diatoms for practical

assessment of pollution. Interest grew over subsequent decades and diatoms are now widely used for ecological assessment in both Europe (e.g. RIMET 2012; KELLY et al. 2014), with the Water Framework Directive, and the US (e.g. BARBOUR et al. 1999; POTAPOVA & CHARLES 2007; HAUSSMANN et al. 2016), with the Clean Water Act. Until now standard methods for freshwater ecological assessment based on diatoms (e.g. EUROPEAN COMMITTEE FOR STANDARDISATION 2014 a, b) have required the biofilms on submerged surfaces to be sampled and then the species present in this biofilm to be identified and counted using their siliceous exoskeleton under light microscopy. This requires time and highly trained analysts with good knowledge of the taxonomic literature. This limits output per analyst to no more than 200–300 analyses per year for the most taxonomically intensive methods, which is a bottleneck for ecological monitoring (HAJIBABAEI et al. 2016). Moreover, there can be considerable inter–analyst variation (e.g. BESSE–LOTOSKAYA et al. 2006; KAHLERT et al. 2009, 2012).

So–called "DNA–Barcoding" (HEBERT et al. 2003) has been proposed as an alternative to microscopical identification, using DNA sequencing to recognise diatom species. This concept was expanded to enviromental samples, in "DNA–metabarcoding" (POMPANON et al. 2011), where species are identified in natural samples using their DNA. Several studies have demonstrated that this approach may be applicable to river diatoms (KERMARREC et al. 2013b; ZIMMERMANN et al. 2015; VISCO et al. 2015; RIVERA et al. 2017). However, these studies also showed that a potential limitation of metabarcoding was an insufficient coverage of existing taxa of the reference barcoding library. This needs to be as complete as possible and must be curated to maintain its quality (i.e. taxonomic homogeneity of assignations, sequence quality, and traceability of data and metadata; (RIMET et al. 2016; KUSBER et al. 2012). Barcodes in such libraries are obtained in several ways. Firstly, through single–cell isolations, culturing, and Sanger sequencing (e.g. EVANS et al. 2007; TROBAJO et al. 2009; ZIMMERMANN et al. 2014a; ABARCA et al. 2014). However, cell isolation is a long process which can in many cases be unsuccessful since some species appear not to thrive under culture conditions. Moreover, some species in cultures show deformations of their frustule, which can make them difficult to identify especially if they have been cultivated for a long time. Another drawback is the rapid cell size reduction of some species in culture which also complicates identification (e.g. KI et al. 2008). As a result of a combination of these factors, many species that are important for ecological assessment have not yet been sequenced.

A second means of obtaining barcodes is single–cell PCR which can be used to obtain nucleotide sequence for taxa that cannot be cultured (TAKANO & HORIGUCHI 2006; GOMEZ et al. 2012). Single–cell extraction/PCR has recently been applied to diatoms by HAMILTON et al. (2015) and KHAN–BUREAU et al. (2016), but in these cases, identifications were carried out only on living cells, which, in most cases, may prevent correct species identification (HAMILTON et al. 2015). Moreover, the majority of the sequenced diatoms were relatively large and small–celled taxa (10–20 µm) were excluded.

A final means of obtaining barcodes for reference libraries is direct sequencing of environmental samples in order to avoid the laborious procedures involved in the first two methods (isolation, culturing, single cell sorting). Examples of two different approaches have already been published. One is simple direct Sanger sequencing of samples presenting very low species diversity, as in some Chilean rivers with *Didymosphenia geminata* (LYNGBYE) MART. SCHMIDT blooms (JARAMILLO et al. 2015). However, this method is relatively imprecise and several *rbc*L sequences published do not correspond to the targeted species announced in the paper (e.g. NCBI accession numbers: KR066780, KR066784). Another approach is a PCR of the environmental sample followed by cloning (e.g. KHAN–BUREAU et al. 2016).

In this paper, we present another means of enriching barcode reference libraries. In this case environmental samples were sequenced using High–Throughput Sequencing (HTS) and the outputs compared with the results of light and electron microscope analyses of the same samples. There are several potential benefits compared to previous studies. First, a greater number of sequences per sample than Sanger sequencing should be accessible with HTS. Second, a much bigger number of samples can be sequenced with HTS and thereby costs can be reduced. The challenge lies in selecting sequences from environmental samples that correspond to the species of interest. The questions we want to address are:

1. Is it possible to relate environmental sequences to the target species observed by microscopy and to do so with high reliability?
2. What are the advantages/disadvantages associated with the use of sequences from uncultured diatoms?
3. Which material, data and metadata must be stored with these sequences to ensure good traceability?

Several examples will be given from Mayotte island, a French tropical island of the Comoros archipelago situated in the Mozambique Channel (Africa), where 98 environmental samples from rivers were sequenced using a HTS technology, Ion–Torrent PGM, and also observed with light and scanning electron microscopy. The open–access barcoding library R–Syst::diatom (RIMET et al. 2016) is used as the host database to store the data (www.rsyst.inra.fr/) presented in this paper and the Thonon Culture Collection (TCC) to store the material (www6.inra.fr/carrtel–collection_eng/), as well as the Botanischer Garten und Botanisches Museum Berlin–Dahlem of Berlin (Germany) and the Conservatoire et Jardin Botaniques of Geneva (Switzerland).

Materials and Methods

Study area and sampling methodology. Mayotte (France) is a tropical island with a surface of 374 km², located in the Indian Ocean to the northwest of Madagascar and to the east of Mozambique (12°50'35"S, 45°08'18"E) (Fig. 1). Geologically it is of volcanic origin and is part of the Comoros archipelago. It consists of two main parts, the smaller Petite–Terre (11 km^2) and the Grande–Terre (363 km^2) where the study was performed. The main pressures on its rivers are related to the fast growing population (226,915 inhabitants in 2015; Insee 2016). Samples were collected as part of river pollution assessments; results are reported to the European Commission as part of France's obligations under the Water Framework Directive (European Commission, 2000). For this study, samples were collected during the dry season in July and August 2015.

The sampling procedure followed European standards (European Committee for Standardisation 2014 a,b). Benthic diatoms were collected from at least five stones from the fast–flowing parts of sampling sites. The upper surfaces of the stones were scrubbed with a toothbrush in order to collect the biofilms. The samples were then fixed with ethanol to give a final concentration of at least 70%.

Preparation for microscopy. Diatom valves were cleaned using 40% H_2O_2 and 40% HCl. Cleaned valves were mounted in a resin (Naphrax©, Brunel Microscopes: http://www.brunel-microscopes.co.uk/) with a high refractive index to create permanent slides. For scanning electron microscopy (SEM), dried cleaned diatom valves were coated with gold using a Cressington 108 auto sputter coater© and examined with a Zeiss DSM940A © in the Natural History Museum of Geneva (Switzerland). Samples were analysed by light microscopy as part of a parallel study (Tapolczai et al. 2017). Four samples with the low number of species were selected.

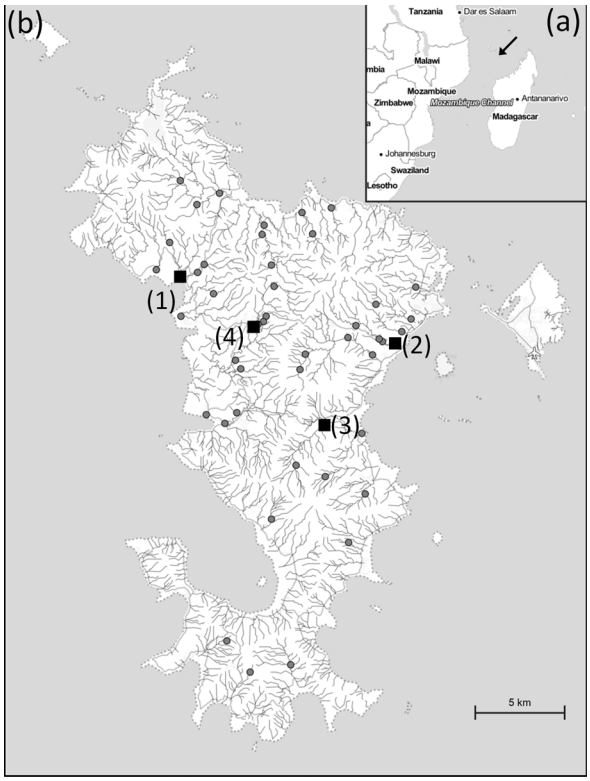

Fig. 1. Study site location: (a) General location of Mayotte in Mozambique channel (arrow); (b) Location of rivers (grey lines), sampling sites in Mayotte (grey dots) and the selected sampling sites discussed in this paper (black squares): (1) Soulou river waterfall, (2) downstream the Gouloué river (Poll 7), (3) downstream the Songaro Mbili river (Poll 29) (4) Mouala river near Mirereni city (Poll5).

DNA extraction. These four environmental samples were centrifuged at 13,000 rpm (equivalent to 18,000 × *g*) for 30 min, the supernatant was then removed. 25 mg of wet pellet was used for each sample. DNA extraction was based on the Sigma–Aldrich GenEluteTM–LPA DNA protocol which was used in previous studies (Kermarrec et al. 2014; Chonova et al. 2016; Vasselon et al. 2017). The final elution volume was 40 µl.

Preparation of the library of amplicons and HTS sequencing: For all four samples, HTS sequencing of a 312–bp fragment of *rbc*L (the exact region is situated between the primers given below) was performed. For each DNA sample, PCR amplification was performed in three replicates on 1 µl of extracted DNA in a mix (25µl final volume) containing: 0.75 Unit of Takara LA Taq® polymerase (for 25 µl of final volume, 0.15 µl of Taq at 5Unit/l), 2.5 µl of 10× LA PCR Buffer II (Mg^{2+}), 1.25 µl of 10 µM of forward and reverse primers, 1.25 µl of 1.51×10^{-4} mol.l^{-1} BSA (Bovine Serum Albumin), 2 µl of 2.5 mM dNTP and completed with 15.6 µl H_2O (molecular biology grade). The primer pair Diat_rbcL_708F (Stoof–Leichsenring et al. 2012) and R3 (Bruder & Medlin 2007) was modified to amplify a broader diversity of diatom as follows: forward primer combine an equimolar mix of Diat_rbcL_708F_1 (AGGTGAAGTAAAAGGTTCWTACTTAAA), Diat_rbcL_708F_2 (AGGTGAAGTTAAAGGTTCWTAYTTAAA) and Diat_rbcL_708F_3 (AGGTGAAACTAAAGGTTCWTACTTAAA); reverse primer combine an equimolar mix of R3_1 (CCTTCTAATTTACCWACWACTG) and R3_2 (CCTTCTAATTTACCWACAACAG). PCR reaction conditions were as follows: initial denaturation of DNA at 95 °C for 15 min followed by 33 cycles with 45 s denaturation at 95 °C, followed by 45 s annealing at 55 °C and 45 s extension at 72 °C. One no–template control (NTC) was used as a negative control.

The 3 replicates of PCR amplicon for each sample were then pooled and cleaned with Agencourt AMPure beads (Beckman Coulter, Brea, USA) following the manufacturer's instructions, except that a 1.5:1 beads:DNA ratio was used specifically to purify the 312–bp fragment. Purified amplicons were assessed for quality and quantified using the 2200 TapeStation (Agilent technologies, Santa Clara, USA) with D1000 screen tape and reagents. The purified amplicons were used to prepare four DNA libraries for HTS with Ion Torrent technology using the NEBNext® Fast DNA Library Prep set for Ion Torrent™ (BioLabs, Ipswich, USA) following the manufacturer protocols for End repair, PCR amplification of adapter ligated DNA (7 cycles) and cleaning steps. Ligation of library adapters to purified amplicons was done using 2µL of P1 adapter (NEB kit) and 2µl of A–X tag adapter provided in Ion Express™ Barcode adapters (Life Technologies, Carlsbad, USA) using 1 tag per amplicon.

The quality, size and concentration of the libraries were checked using the 2200 TapeStation with D1000 High Sensitivity screen tape and reagents. Each library was diluted to 100 pM and all were pooled together with 50 libraries from other environmental samples from Mayotte in a unique mix sequenced using 1 Ion 318™Chip Kit V2 (Life Technologies, Carlsbad, USA) on a PGM Ion Torrent machine by the "Plateforme

Table 1. Sampling sites location, habitat description, and collectors names. Sites acronymes used by the local authorities are given in brackets.

Sampling site	Sampling date	Habitat	Coordinates	Collectors	Dominant target species
Soulou river waterfall	25/07/2015	Unpolluted waterfall	12°46'48.3"S 45°6'6.5"E	Rimet and Tapolczai	*Epithemia hirundiformis* (O. Müller) comb. nov.
Gouloué river, near Passamainty city (Poll7)	20/07/2015	Polluted river, marine influence	12°47'57.9"S 45°12'37.9"E	Rimet and Tapolczai	*Halamphora ghanensis* Levkov
Songaro Mbili river near Dembeni city (Poll29)	24/07/2015	Polluted river	12°50'23.3"S 45°10'28.3"E	Rimet and Tapolczai	*Gomphonema clavatuloides* sp. nov.
Mouala river near Mirereni city (Poll5)	23/07/2015	Polluted river	12°47'25.9"S 45°08'21.6"E	Tapolczai and Vasselon	*G. parvulum* (Kützing) Kützing sensu lato

Génome Transcriptome" (PGTB, Bordeaux, France).

Sequence data processing. Demultiplexing and adapter removal steps were made by the Sequencing Platform which provided a single fastq file for each of the 55 libraries. DNA reads were filtered for length and quality using mothur software (Schloss et al. 2009) in every fastq file with the following settings: a minimum length of 250 bp, a Phred quality score higher than 23 over a moving window of 25 bp, a maximum of 1 mismatch in the forward primer sequence, homopolymers shorter than 8 bp, and absence of ambiguous bases. Reads which were not fully aligned with the *rbc*L barcode were removed. The resulting files were analysed together. Denoising of sequencing error was performed with the pre.cluster command by creating read clusters allowing one nucleotide difference between DNA reads. Chimera removal was done with UCHIME algorithm (Edgar et al. 2011).

The R–Syst::diatom database (Rimet et al. 2016) (database version v5, http://www.rsyst.inra.fr/en), restricted to our 312–bp *rbc*L barcode, was used as the reference database. Taxonomic assignment of DNA reads at species level was made using this reference database and the Naïve Bayesian method (Wang et al. 2007) with a confidence score threshold of 85%. Only DNA reads assigned to the Bacillariophyta phylum (diatoms) were used in further analyses.

After dereplication, uncorrected pairwise distances were calculated between aligned reads to generate a similarity distance matrix. Based on this distance matrix, reads were clustered in Operational Taxonomic Units (OTUs) using the Furthest Neighbour algorithm at 100% similarity level in order to have each OTU represented by a single sequence. Singletons were removed.

Then, for each of the four low–diversity samples, a Blastn was run on the entire NCBI database with each of the 15 to 20 most abundant 312–bp sequences. 312–bp sequences showing a BLAST result congruent with the microscopical identification of targeted species were kept for subsequent phylogenetic analyses.

Phylogenetic analyses. For each of the four samples, the selected 312–bp sequences were aligned with a selection of sequences from R–Syst::diatom database using Muscle (Edgar 2004) in Seaview (Gouy et al. 2010). This selection of sequences was done based on their taxonomic proximity to the 312–bp sequences. The lengths of the Sanger sequences from R–Syst::diatom were at least 1000 bp. The best substitution model was then tested in MEGA7 (Kumar et al. 2016). A first phylogenetic tree was calculated following the best substitution model with raxmlGUI (Silvestro & Michalak 2012) and the sequence selection of R–Syst::diatom, after which we calculated a second phylogenetic tree, adding the 312–bp sequences in the phylogeny and enforcing its topology with the topology of the first tree. This analysis is also available in raxmlGUI under the "enforce constraint menu" and "define topological constraint". Trees were drawn in MEGA7.

Material and data accessibility. All material is accessible through the Thonon Culture Collection (TCC: https://www6.inra.fr/carrtel–collection_eng/) and at the Botanischer Garten und Botanisches Museum Berlin–Dahlem, Freie Universität Berlin (B) and also at the Conservatoire et Jardin Botaniques of Geneva (G) (for the newly described species, all nomenclatural acts have been submitted to PhycoBank for registration of new scientific names; http://phycobank.org). TCC culture collection hosts algal cultures, but also uncultured samples containing species of interest (as permanent slides, and raw and treated material). All metadata are stored in the open–access R–Syst::diatom reference database; a detailed description of this database and its management is given in Rimet et al. (2016). Taxonomy, sequences, photos, sampler names, phenotypic data, etc., can be consulted and downloaded at: http://www.rsyst.inra.fr/

Results

Table 1 and Fig. 1 give the sample locations, their environmental characteristics and the dominant target species.

Morphology and ecology

Halamphora ghanensis Levkov (Fig. 2)

This sample is registered and conserved in the TCC

(TCC956 – uncultured sample) and in the Botanischer Garten und Botanisches Museum Berlin–Dahlem, Freie Universität Berlin (B 40 0041830).

The morphology of the valves from the Gouloué river (Fig. 1, site 2) fitted the description of *H. ghanensis* (Levkov 2009): valve length from 25.2 to 26.4 µm (24 to 27 µm in Levkov), width 5.6 to 6.4 µm (5 to 5.6 µm in Levkov 2009), 13 to 14 dorsal striae/10 µm (14 to 16 in Levkov 2009). We do not think the small differences in measured width and stria density are sufficiently different from the original description to suggest a different species. Morphological features observable in SEM also correspond to the description given in Levkov (2009): dorsal ledge crenulated, dorsal striae biseriate and interrupted by longitudinal bars near the dorsal margin. Striae also appear biseriate internally. There is a poorly developed helictoglossa in the distal raphe endings and fused central helictoglossae at the proximal raphe endings, as in the species description (Levkov 2009). A single row of dorsal areola was observed near the raphe in internal view (a feature not present in *H. acutiuscula* (Kützing) Levkov, which is otherwise morphologically close to *H. ghanensis*).

The sample was collected from Gouloué river, near Passamainty city (Table 1), a polluted river surrounded by a village where many houses discard their wastewater directly into the river (2.5 mg.l^{-1} of dissolved organic carbon, 0.08 mg N.l^{-1} of NH_4^+, 113 mg.l^{-1} O_2 chemical oxygen demand). This section of river is subject to tidal influence, which explains the high conductivity (2260 µS.cm^{-1}) and chloride concentration (706 mg.l^{-1} Cl^-) measured on the day of sampling. We would emphasise, however, that *H. ghanensis* is regularly observed in samples from Mayotte island, in river stretches that are not tidal and where the conductivity is < 300 µS.cm^{-1}. Usually, it forms < 10% of the diatom assemblage (based on 400 valves counted per sample) whereas in the Gouloué river it was abundant (38% of the valves counted). We therefore suspect this species prefers brackish waters rich in organic matter. In Levkov (2009), *H. ghanensis* was reported from a river in Ghana (West Africa), but no chemical measurements were given.

***Gomphonema clavatuloides* Rimet, D.G. Mann, Trobajo et N. Abarca sp. nov. (Fig 3)**

Description: Valve lanceolate–clavate, with the broadest portion of the valve at the central nodule; apex and base rounded. Axial area narrow, linear. Central area small, transversely elongated, made by slight shortening of central striae on both valves sides. Stigma present at the end of a shortened central striae. Length from 25 to 41 µm, breadth from 5.5 to 7.0 µm, striae density from 7 to 9 in 10 µm. Transapical striae moderately radiate, more parallel towards the poles. Areolae lineolate in external view: in internal view, the areolae have the same shape and lie in a furrow. The central punctum has a round opening in external view and is lineolate in internal view. The internal central raphe ends are hooked towards the primary side of the valve (i.e. in the opposite direction to the external distal ends). The external central raphe ends are slightly deflected towards the primary side of the valve and terminate in a drop–shaped expansion. Distally, the raphe ends in terminal fissures that are slightly bent towards the secondary side. Both internal distal raphe

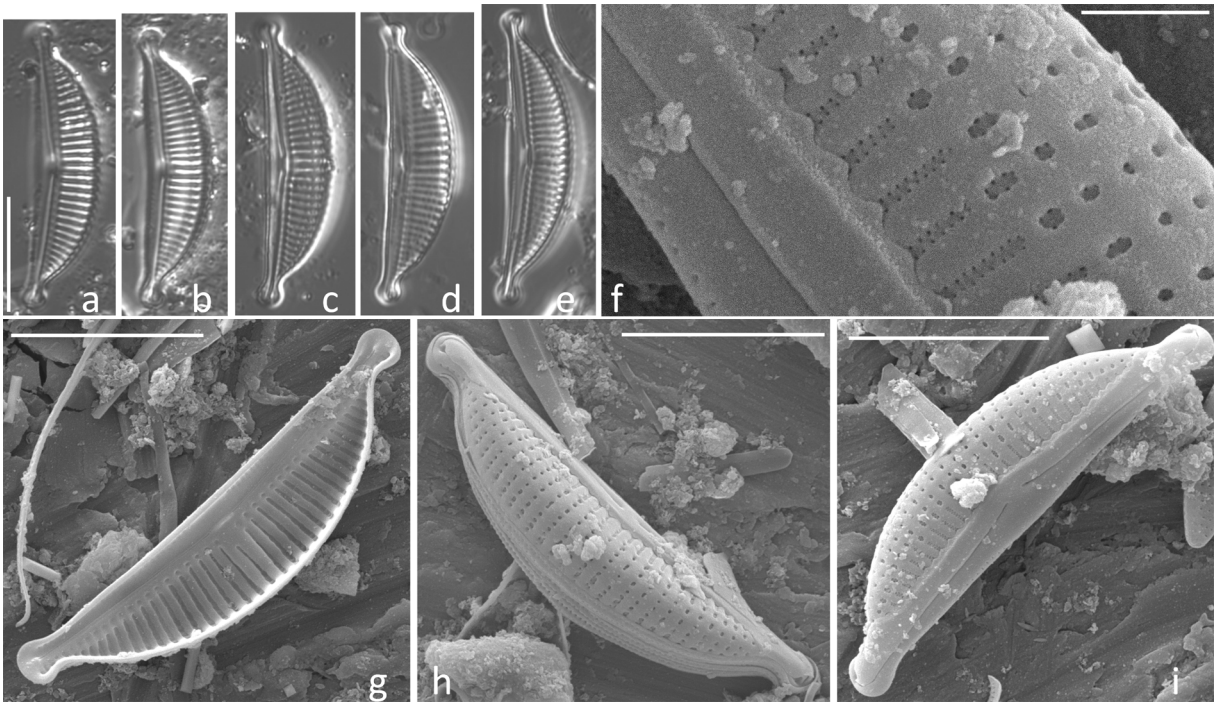

Fig. 2. *Halamphora ghanensis* in downstream Gouloué river: (a–e) Light microscopy, valve views; (f–i) Scanning electron microscopy, (g) internal view; (f, h ,i) external views, (f) detail of areola structure. Scale bar 10 µm (a–e, g–i), 2 µm (f).

ends terminate straight in helictoglossae. Four pores are present on the mantle in extension of each valve stria.

Holotype: B 40 0041829 (Botanischer Garten und Botanisches Museum Berlin–Dahlem, Freie Universität Berlin, Germany) represented by Fig. 3g.
Isotype: Conservatoire et Jardin Botaniques, Geneva, Switzerland, reference number: G00260989.
This sample (slides, raw material) is also registered and conserved in the Thonon Culture Collection of the INRA: TCC955 – uncultured sample deposited in Thonon–les–Bains, France.
Type locality: Songaro Mbili river near Dembeni city (France, Mayotte Island, Northern Mozambique Channel), sampled 24 July 2015 by F. Rimet, coordinates: 12°50'23.3"S, 45°10'28.3"E (Fig. 1 site 3).
Name registration: http://phycobank.org/100011
Etymology: The specific epithet refers to the close resemblance to *G. clavatulum* E. Reichardt.

Similar taxa: This taxon resembles the *G. longiceps* Ehrenberg (*G. clavatum* post auct.)species complex, which also includes similar species and varieties such as *G. clavatulum* E. Reichardt and *G. subclavatum* (Grunow) Grunow (Krammer & Lange–Bertalot 1986; Reichardt 1999). *Gomphonema clavatuloides* has the same general clavate shape as *G. clavatulum* and *G. subclavatum* but differs in its breadth (5.5–7.0 µm: compare 4.7–5.7 µm for *G. clavatulum* and 8–10 for *G. subclavatum*) and striae density (7–9 in 10 µm;

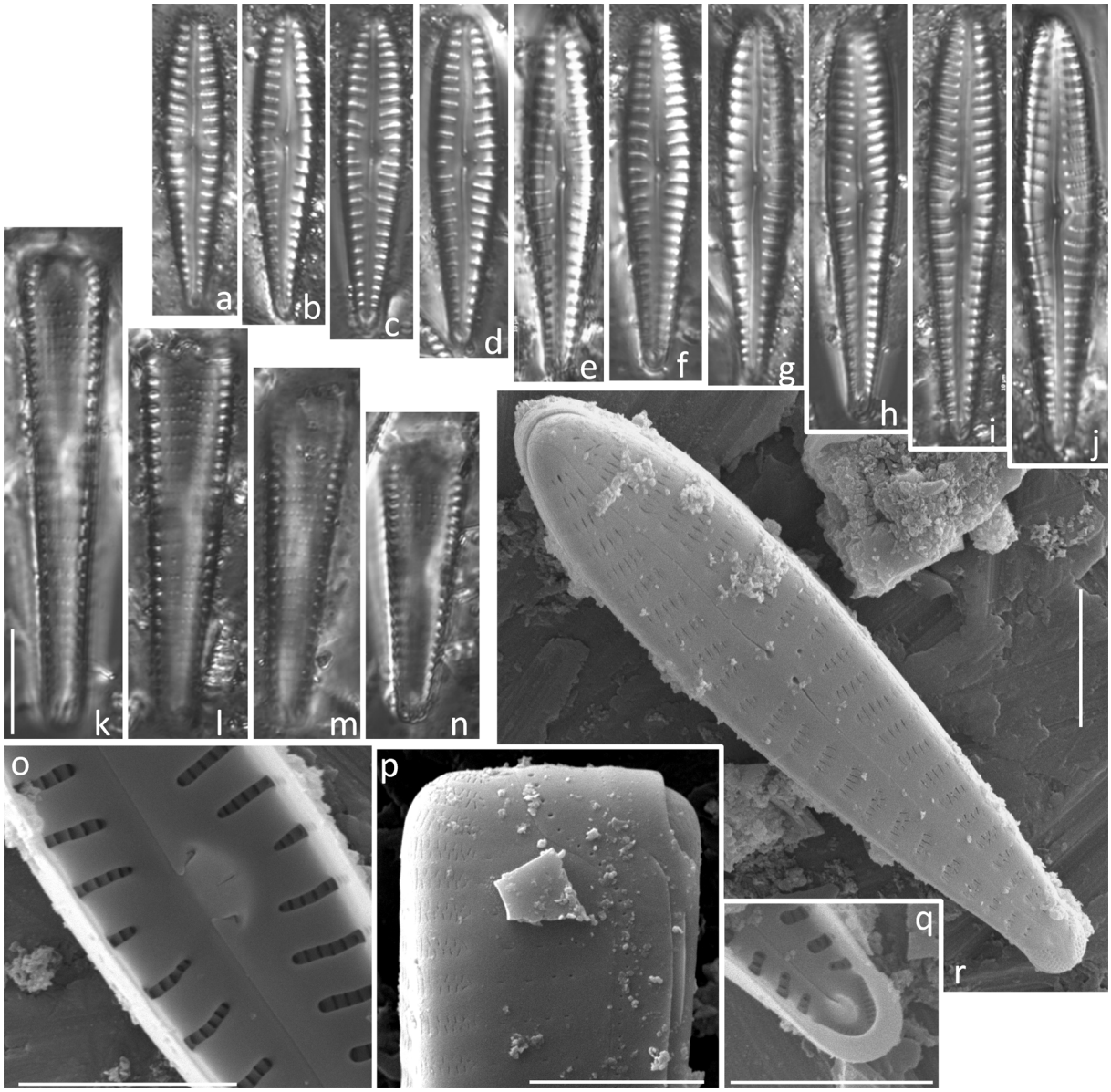

Fig. 3. *Gomphonema clavatuloides,* downstream the Songaro Mbili river: (a–j) light microscopy, valve view; (k–n) light microscopy, girdle view; (o–r) Scanning electron microscopy; (o) internal view, detail of proximal raphe ending and areola structure; (p) external view of girdle view (head pole); (q) internal raphe ending at the foot pole; (r) general external view. Scale bar 10 µm (light microscopy photos), 5 µm (scanning electron microscopy).

compare 10.5–14 µm for *G. clavatulum* and 9–13 for *G. subclavatum*). It might also be confused with small or medium–sized valves of *G. paludosum* E. Reichardt but differs in striation pattern and number of striae in 10 µm (Reichardt 1999).

Ecology: This sample was collected from Songaro Mbili river (site 3, Fig. 1), near Dembeni city, a polluted stream surrounded by subsistence farming and villages; several houses discard their wastewaters directly in the river. The water was almost stagnant with traces of soap (women wash clothes at this place). Conductivity was 280 µS.cm^{-1}, chemical oxygen demand was 66 mg.l^{-1} O_2 (measured in August 2014), but the dissolved phosphorus concentration was low (0.014 mg of P–PO_4^{3-}.l^{-1}). In this sample, *G. clavatuloides* was the dominant species (85.6% of the counted valves).

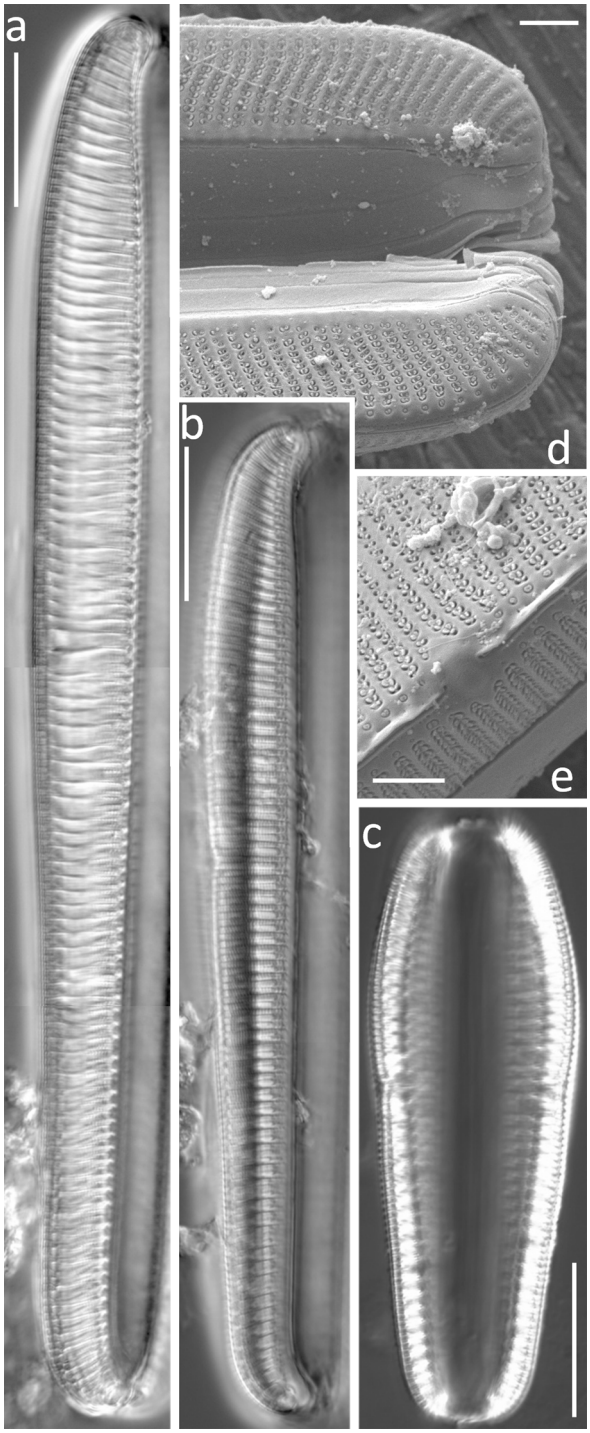

Fig. 4. *Epithemia hirudiniformis,* Waterfall of Soulou river: (a–c) light microscopy; (d, e) Scanning electron microscopy; (d) detail of the foot pole; (e) detail of the proximal ending of the raphe and of the areola structure. Scale bar 20 µm (a–c), 3 µm (d–e).

***Epithemia hirudiniformis* (O. Müller) Rimet, D.G. Mann, R. Trobajo, J. Zimmermann et R. Jahn comb. nov. (Fig. 4)**

Basionym: *Rhopalodia hirudiniformis* O. Müller, Botanische Jahrbücher für Systematik, Pflanzengeschichte und Pflanzengeographie 22, p. 67, pl. I: figs 40–46, 51, 52; pl. II: figs 15–17 (1895).
Name registration: http://phycobank.org/100012

Taxonomy and morphology: This sample is registered and conserved in TCC (TCC954 – uncultured sample) and in the Botanischer Garten und Botanisches Museum Berlin–Dahlem, Freie Universität Berlin (B 40 0041831). We propose a new combination for *Rhopalodia hirudiniformis* in *Epithemia* since, according to Ruck et al. (2016), *Rhopalodia* is paraphyletic with respect to *Epithemia*. When *Rhopalodia* and *Epithemia* are combined in a single monophyletic genus, the correct name is *Epithemia.*

A few species and varieties belonging to former *Rhopalodia* are heteropolar, in particular, *R. hirudiniformis* and its varieties var. *parva* O. Müller, var. *turgida* Fricke, and also the species *R. rhopala* (Ehrenberg) Hustedt. All these taxa were described from east African lakes (Cocquyt 1998; Cocquyt et al. 2018). Regularily, some publications and databases incorrectly spell this species “hirundiniformis”.

The length of the valves found in the Soulou waterfall (site 1, Fig. 1, Fig. 4) ranged from 73.6 to 215 µm and the maximum width from 7 to 14 µm. However, the smallest width measurements may be erroneous as the valves were not always flat during light microscopy. In our sample, the striae and fibulae densities were 11–12 and 5–6 in 10 µm, respectively. These densities correspond to the description of *R. rhopala*, *R. hirudiniformis* and its varieties. However, the length of the frustules in our sample corresponds to *R. rhopala* and *R. hirudiniformis*. Indeed, according to the valve sizes given in Cocquyt (1998) the largest valves (220 µm) correspond to *R. rhopala* while most of the other valves fall into the size range of *R. hirudiniformis* (58–113 µm). The distribution of the length and width measures followed a normal distribution (p > 5%, measurements carried out on 28 valves) so we cannot consider this population as belonging to two different species. Finally, the clear constriction observed on *R. hirudiniformis* var. *turgida* does not correspond to the morphology observed in this sample. For these reasons we decided to identify the valves observed in our sample as *R. hirudiniformis*.

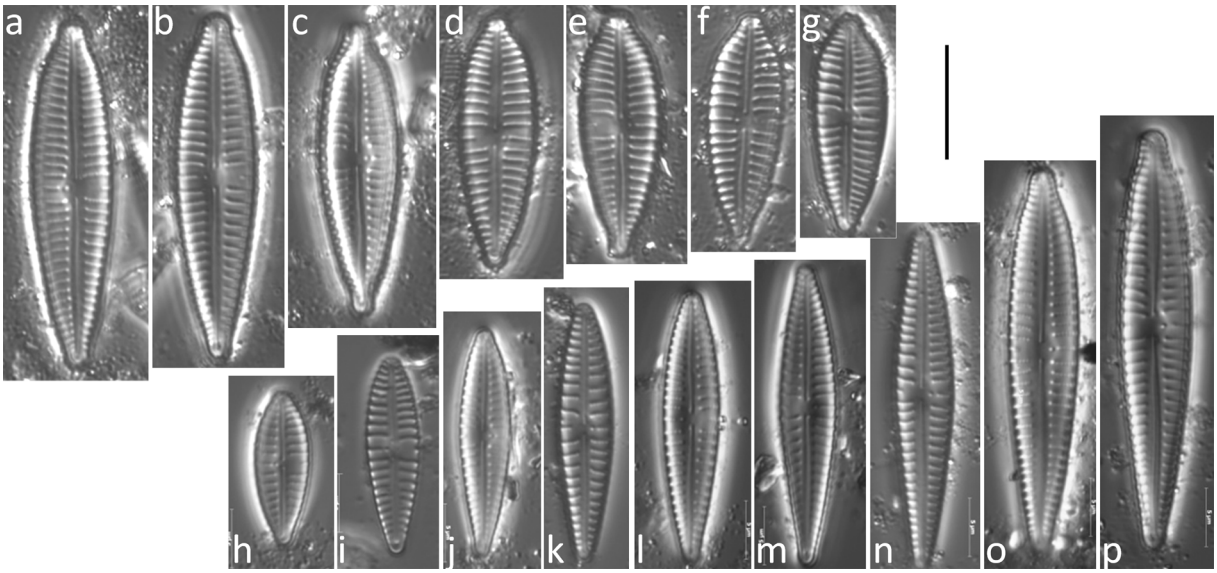

Fig. 5. *Gomphonema parvulum* sensu lato downstream the Mouala river. Light microscopy photos showing the morphological heterogenity of this taxon in this sample: (a–g) morphology corresponding to *G. lagenula*; (h–p) morphology corresponding to *G. parvulum* sensu lato. Scale bar 10 µm.

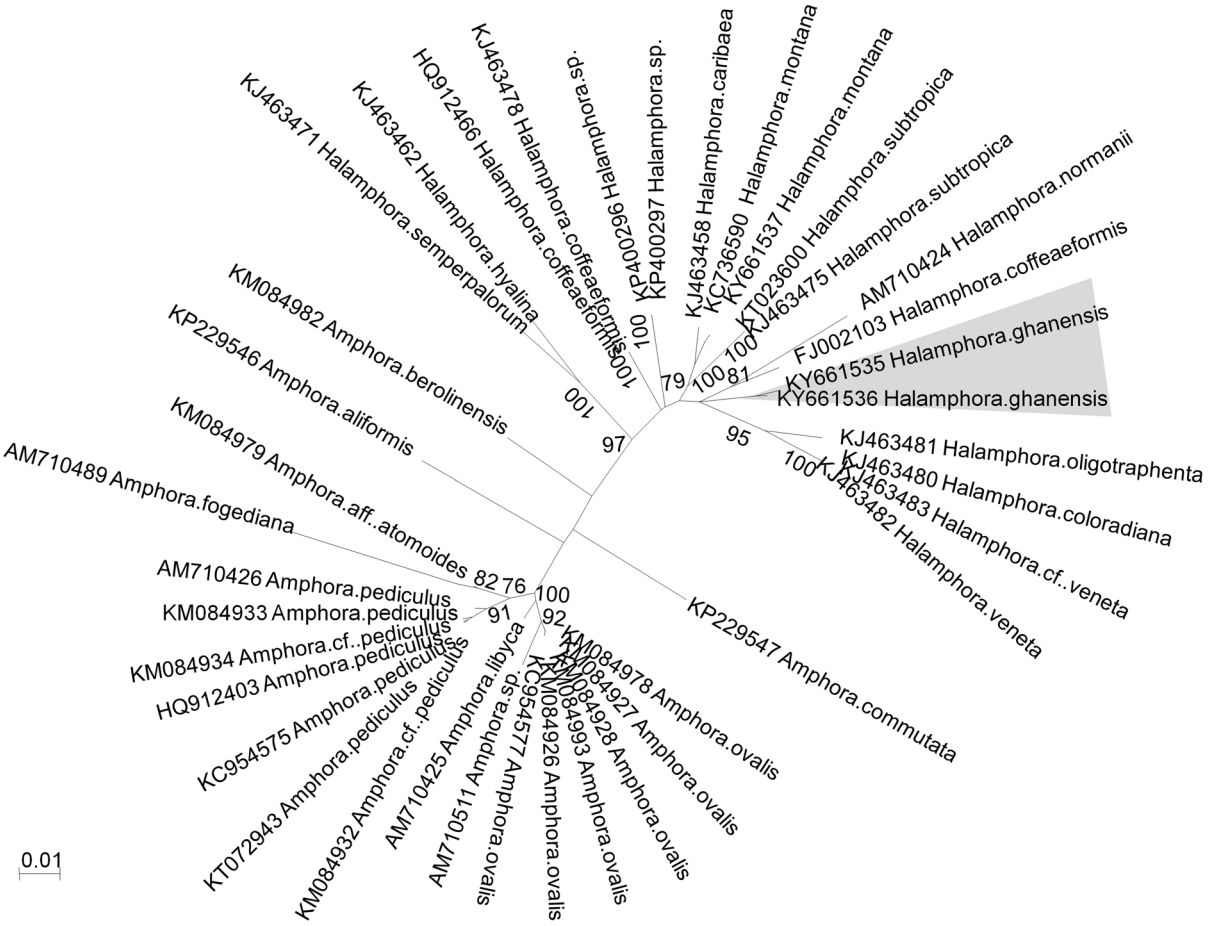

Fig. 6. Constrained unrooted phylogeny of *Halamphora ghanensis.* The two *H. ghanensis* sequences (312 bp) and KT023600, AM710424, AM710425, AM710426, AM710489, AM710511 (lengths of these sequences varies from 385 to 714 bp) were constrained by the phylogeny of the 34 other sequences (926 bp). Maximum likelihood tree with rapid bootstrap and GT Gamma, RaxML. 1000 bootstraps. Bootstrap values above 50% are given for each node. Scale bar: number of substitutions per site.

Ecology: The Soulou waterfall (site 1, Fig. 1) comes from the Chirini river and falls onto a beach in the north west part of Mayotte Island. The sample was taken on the vertical wall of the waterfall where thick aerial biofilms were visible. *E. hirudiniformis* was associated with several species of cyanobacteria (*Oscillatoria* sp., *Pseudanabaena* sp., *Chroococcus* sp., *Plectonema* sp.) and was the only diatom species observed. Moreover, four to eight cells of endosymbiotic cyanobacteria could be observed in each cell of *E. hirudiniformis*. There is a sampling station on the Chirini river 500 m upstream of the waterfall, which shows quite good water quality (230 µS.cm^{-1} conductivity, 30 mg.l^{-1} O_2 chemical oxygen demand, less than 0.1 mg.l^{-1} of NO_3^-–N, 0.01 mg.$^{-1}$ of PO_4^{3-}–P).

***Gomphonema parvulum* (Kützing) Kützing sensu lato (Fig. 5)**

Taxonomy and morphology: This sample is registered and conserved in TCC (TCC958 – uncultured sample) and in the Botanischer Garten und Botanisches Museum Berlin–Dahlem, Freie Universität Berlin (B 40 0041832). Several samples in Mayotte island contained high abundances of *G. parvulum* sensu lato (encompassing *G. narodoense* R. Jahn, N. Abarca, J. Zimmermann et Enke, *G. lagenula* Kützing, *G. parvulum*, and varieties). One of these (Mouala river near Mirereni city) showed significant morphological and genetic diversity. Its morphology is given in Figure 5. Figs 5a–g correspond to the morphology of *G. lagenula*, indeed, the shape of the head pole is consistently more rostrate to capitate than in *G. parvulum* sensu stricto and the general shape of the frustule is lanceolate, as described in Abarca et al.

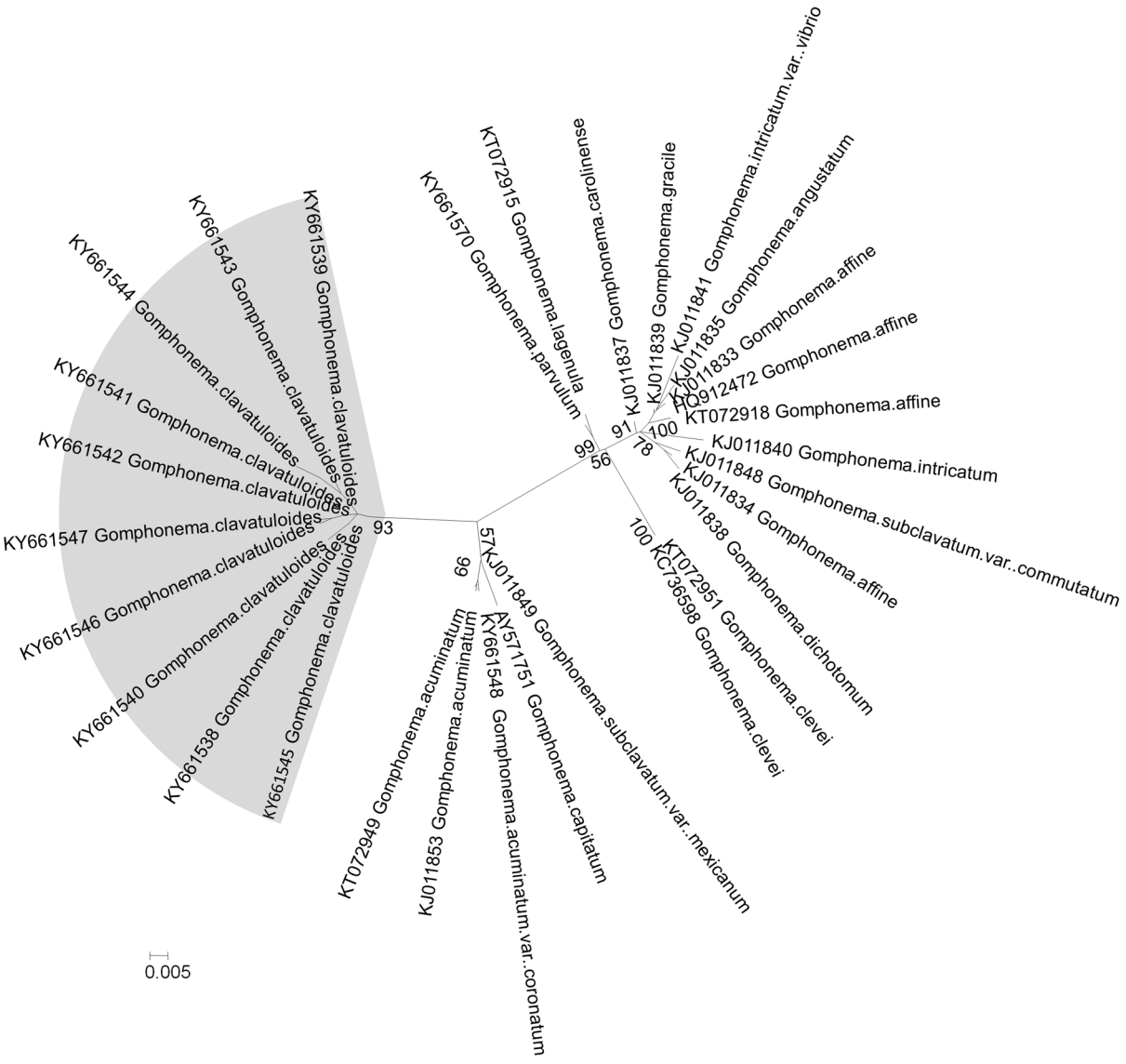

Fig. 7. Constrained unrooted phylogeny of *Gomphonema clavatuloides.* The ten *G. clavatuloides* sequences (312 bp) were constrained by the phylogeny of the 87 other sequences (1093 bp). Maximum likelihood tree with rapid bootstrap and GT Gamma, RaxML. 1000 bootstraps. Bootstrap values above 50% are given for each node. Scale bar: number of substitutions per site.

(2014). Figs 5h–p have a rather different shape; indeed the linear to lanceolate valve is more slender and much less clavate than *G. lagenula*. This second morphodeme could fit *G. parvulum* var. *parvulum* morphodeme exilissimum (strain D12_022) according to the information given by ABARCA et al. (2014). Nevertheless, the stria density of the valves ranged from 9 to 12 in 10 µm and does not correspond to the density in *G. exilissimum* (GRUNOW) LANGE–BERTALOT (12–14 in 10 µm, in HOFMANN et al. 2011).

In this sample, another *Gomphonema* species was observed under microscope, *G. bourbonense* E. REICHARDT.

Ecology: This sample was collected from the Mouala river (site 4, Fig. 1) near Mirereni city, a polluted river receiving wastewaters from several houses. When the diatoms were sampled, the water was deoxygenated (24% O_2 saturation); conductivity was 120 µS.cm^{-1} and chemical oxygen demand was 76 mg.l^{-1} O_2.

Phylogeny

Halamphora ghanensis **(Fig. 6)**

1751 different environmental sequences were obtained from the sample from the Gouloué river, near Passamainty city, of which 432 were singletons. The 15 most abundant sequences were blasted on NCBI. Two sequences had high similarity with *Halamphora montana* (KRASSKE) LEVKOV (96%) and their length was 312–bp. These sequences were selected to build a phylogeny. The list of sequences used as constraints for the 312–bp sequences is given in Supplementary data S1 and the alignment used in the phylogeny is in Supplementary data S5. No indels or stop codons appeared in the 312–bp sequences after alignment. The best model selection was the GTR+G+I model (General Time Reversible model with gamma distribution and Invariable sites) which showed the lowest AICs (Akaike Information Criterion, corrected). A maximum likelihood constrained phylogeny with GTR+G+I model and 1000 rapid bootstraps was calculated and is drawn in Fig 6.

The two 312–bp sequences are included in a group supported by a high bootstrap value (97%). This clade is composed of representatives of *Amphora* and *Halamphora*. Three of the *Amphora* species included were described by WACHNICKA & GAISER (2016), and STEPANEK & KOCIOLEK (2014) have already suggested that they should be transferred into *Halamphora*. These are *Amphora semperpalorum* WACHNICKA et GAISER, *A. subtropica* WACHNICKA et GAISER and *A. caribaea* WACHNICKA et GAISER. *Amphora hyalina* KÜTZING, also included in this group, was described by KÜTZING (1844). The results of our phylogeny confirm the proposition of

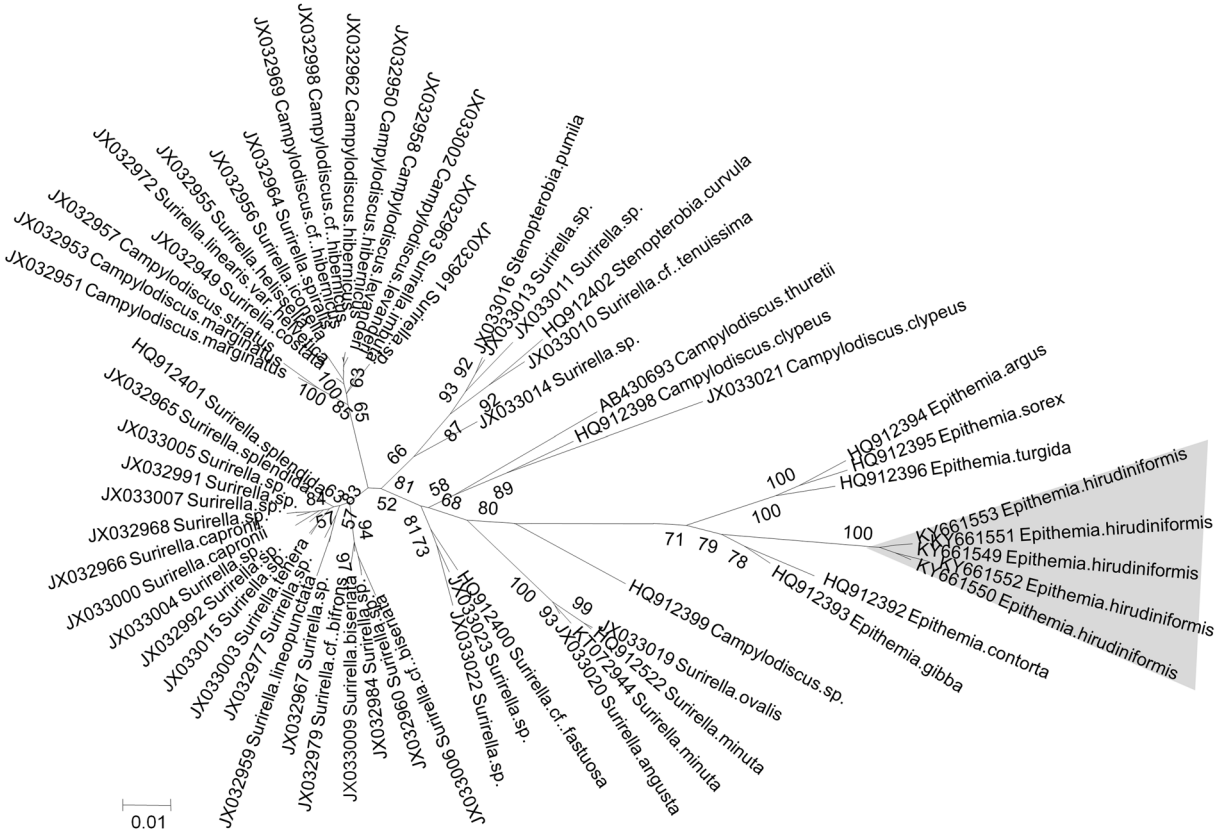

Fig. 8. Constrained unrooted phylogeny of *Epithemia hirudiniformis.* The five *E. hirudiniformis* sequences (290 bp) were constrained by the phylogeny of the 87 other sequences (1362 bp). Maximum likelihood tree with rapid bootstrap and GT Gamma, RaxML. 1000 bootstraps. Bootstrap values above 50% are given for each node. Scale bar: number of substitutions per site.

Stepanek & Kociolek (2014) and the morphological features of these species correspond to the description of *Halamphora* by Levkov (2009). This clade is therefore composed only by *Halamphora* species and we consider that the two 312–bp sequences can be kept in R–Syst::diatom. Moreover we suggest the following new taxonomic combinations:

***Halamphora semperpalorum* (Wachn. et E.E. Gaiser) Rimet et R. Jahn comb. nov.**
Basionym: *Amphora semperpalorum* Wachn. et E.E. Gaiser, Diatom Res. 22: 403, figs 44–48 (2007).
Name registration: http://phycobank.org/100014

***Halamphora hyalina* (Kützing) Rimet et R. Jahn comb. nov.**
Basionym: *Amphora hyalina* Kützing, Die Kieselschaligen Bacillarien oder Diatomeen. W. Köhne, Nordhausen. p. 108, fig. 30/18. (1844).
Name registration: http://phycobank.org/100016

***Halamphora subtropica* (Wachn. et E.E. Gaiser) Rimet et R. Jahn comb. nov.**
Basionym: *Amphora subtropica* Wachn. et E.E. Gaiser, Diatom Res. 22: 407, figs 64–70 (2007).
Name registration: http://phycobank.org/100018

***Halamphora caribaea* (Wachn. et E.E. Gaiser) Rimet et R. Jahn comb. nov.**
Basionym: *Amphora caribaea* Wachn. et E.E. Gaiser, Diatom Res. 22: 399, figs 35–37 (2007).
Name registration: http://phycobank.org/100020

***Gomphonema clavatuloides* (Fig. 7)**
1876 different environmental sequences were obtained from the sample from Songaro Mbili, near Dembeni city, of which 153 were singletons. The 20 most abundant sequences were blasted on NCBI and R–Syst::diatom. 10 sequences had high similarity with *Gomphonema acuminatum* Ehrenberg (95–96%) and their length was 312 bp. Three other sequences also had high similarities with species of *Gomphonema* genus: *G. bourbonense* (one sequence matching with *rbc*L sequences of strains

Fig. 9. Constrained unrooted phylogeny of *Gomphonema parvulum* sequences of the downstream sampling site of the Mouala river. Fourteen *G. parvulum* sequences (275 bp) were constrained by the phylogeny of the 62 other sequences (916 bp). Maximum likelihood tree with rapid bootstrap and GT Gamma, RaxML. 1000 bootstraps. Bootstrap values above 50% are given for each node. Scale bar number of substitutions per site.

TCC441, TCC460, TCC513, TCC450, TCC514, TCC453, TCC452, TCC451) and *G. lagenula* (two sequences matching with *rbc*L sequences of strains TCC470, TCC500, TCC440, TCC432, TCC431, TCC429 and NCBI sequences HG530055, HG530054); these sequences were rejected from further analyses. Sequences showing high similarities with *G. acuminatum* were selected to build a phylogeny. The list of sequences used as constraint for the 312–bp sequences is given in Supplementary data S2 and the alignment used in the phylogeny is in Supplementary data S6. No indels and no stop codons appeared in the 312–bp sequences after alignment. The best model was the GTR+G+I model which showed the lowest AICs. A maximum likelihood constrained phylogeny with GTR+G+I and 1000 rapid bootstraps is given in Fig. 7.

The ten 312–bp sequences form a monophyletic clade supported by a high bootstrap value (93%). The nearest species is *G. subclavatum* var. *mexicanum* (GRUNOW) R.M. PATRICK. Therefore these ten 312–bp sequences can be kept for R–Syst::diatom.

***Epithemia hirudiniformis* (Fig. 8)**

959 different environmental sequences were obtained from the sample from Soulou waterfall, none of which were singletons. The 15 most abundant sequences were blasted on NCBI. 14 sequences had high similarity with *Epithemia gibba* KÜTZING (94–96%) and their length was 290 bp. The 15th had high similarity with *Ulnaria ulna* (NITZSCH) COMPÈRE. This last sequence accounted for 0.5% of the total abundance of the sequences. The list of sequences used as constraits for the 290–bp sequences is given in Supplementary data S3 and the alignment used in the phylogeny is in Supplementary data S7. After alignment, 9 of the 290–bp sequences showed deletions and were removed from the following analyses. 5 sequences were selected and showed no stop codons. The best model selection was the GTR+G+I model which showed the lowest AICs. A maximum likelihood constrained phylogeny with GTR+G+I model and 1000 rapid bootstraps is given in Fig. 8.

The five 290–bp sequences form a monophyletic clade supported by a high bootstrap value (100%). They are inside a larger clade composed of other *Epithemia* species and supported by a high bootstrap value (79%). Therefore only five sequences out of the fourteen 290–bp sequences were kept for R–Syst::diatom.

***Gomphonema parvulum* (Fig. 9)**

1053 different environmental sequences were obtained from the sample from the Mouala river near Mirereni city of which 516 were singletons. The 20 most abundant were blasted on NCBI. 14 sequences had high similarity with *Gomphonema parvulum* sensu lato and their length was 275–bp. These were used to build a phylogeny. The list of sequences used as constraint for the 275–bp sequences is given in Supplementary data S4 and the alignment used in the phylogeny is in Supplementary data S8. No indels and no stop codons appeared in the 275–bp sequences after alignment. The best model selection was the GTR+G+I model which showed the lowest AICs. A maximum likelihood constrained phylogeny with GTR+G+I model and 1000 rapid bootstraps is given in Fig. 9.

Gomphonema parvulum sensu lato (including *G. parvulum* sensu stricto, *G. saprophilum* (LANGE–BERTALOT et E. REICHARDT) ABARCA, R. JAHN, J. ZIMMERMANN et ENKE, *G. narodoense, G. lagenula*) forms a large group supported by 61% bootstrap value. Inside this large group several well supported smaller groups are present: a group with *G. lagenula* strains, a group with *G. saprophilum* strains, a group with *G. parvulum* sensu stricto, a group with *G. narodoense*. However, the position in the tree of several sequences (apart from those of 275–bp) is not well supported.

The 275–bp sequences are present in several groups. Nine sequences are included in the *G. lagenula* group with a bootstrap support of 52%; however, the position of three 275–bp sequences inside the *G. parvulum* sensu lato group is not well supported. Therefore, given the low support of the position of these sequences, we decided not to keep these sequences for R–Syst::diatom.

DISCUSSION

1. Is it possible to relate sequences from HTS analyses of biofilms to a target species observed by light microscopy with high reliability?

Relating sequences to morphological features with high reliability is not an issue in the case of monoclonal cultures since Sanger sequencing delivers a single sequence. Such certainty is, however, harder to obtain when using HTS to sequence biofilm samples collected from the natural environment.

The simplest example presented here is *Epithemia hirudiniformis*, since this was the only species found in microscopy and the 14 most abundant sequences corresponded to this genus. The 15th most abundant sequence matched *Ulnaria ulna.* It was represented by 0.5% of the sequences in the sample, which explains why *U. ulna* was not detected by light microscopy. Such a sample presenting a very low species diversity is similar to that described for *Didymosphenia geminata* in Chilean rivers (JARAMILLO et al. 2015) and it is consequently easy to relate the sequences found with high reliability to the target taxon.

Two other examples, *Halamphora ghanensis* and *Gomphonema clavatuloides*, represent a slightly more complicated situation as each sample included several other species. In both samples six different species were identified using microscopy, while for metabarcoding seven taxa were detected for the *H. ghanensis* sample (3 species and 4 generic assignments where the species could not be determined) and 22 taxa for the *G. clavatuloides* sample

Table 2. Proposed terminology for the naming of sequences from different origins.

#	Newly described	Original material	Morphological data	Molecular data	Terminology examples
1	Yes	Authentic[1] strain of the type (epitype, holotype)	Unialgal culture	Sanger sequencing	*Planothidium caputium* authentic[1] strain D06_014
2	No	No	Unialgal culture	Sanger sequencing	*Planothidium frequentissimum* strain D06_138
3	Yes	Yes	Environmental sample	HTS	*Gomphonema clavatuloides* authentic[1] uncultured sample TCC955 inferred via eDNA
4	No	No	Environmental sample	HTS	*Epithemia hirudiniformis* uncultured sample TCC955 inferred via eDNA *Halamphora ghanensis* uncultured sample TCC956 inferred via eDNA
5	No	No	No	cloning/Sanger	*Planothidium* sp. uncultured sample TF-2014 05DB5_12 inferred via cloning[2]
6	No	No	No	HTS	*Genus species* uncultured sample inferred via eDNA[3]

[1]term for a type-bearing strain (unialgal culture) or isolate (object isolated from environmental sample) because a nomenclatural type is usually a preparation, a slide etc.
[2]sometimes named as clone
[3]only applicable for 100% match with DNA reference library

(16 species and 6 generic assignments where the species could not be determined). Retrieving the sequences of *H. ghanensis* was possible because it was the only member of the Catenulaceae Mereschkowsky in this sample and several reference barcodes of species belonging to this family are present in R–Syst::diatom. Results of BLAST and of the phylogeny confirmed that the sequences are belonging to this family. For *Gomphonema clavatuloide*s, two other species belonging to *Gomphonema* were present in the sample (*G. bourbonensis* E. Reichardt and *G. parvulum*). But here again, identifying the sequences of *G. clavatuloides* was possible, even if this species is new to science, because reference sequences of the two other *Gomphonema* species were available in the reference library and results of BLAST and phylogenies could confirm their membership.

In the case of *G. parvulum* (in the Mouala river) it was much harder to relate sequences to identified target species. This species complex has been studied for a long time (e.g. see Geitler 1972), it is found in rivers worldwide (e.g. Murakami & Kasuya 1993; Silva–Benavides 1996; Ndiritu et al. 2006; Rimet 2009), and it shows significant phenotypic plasticity even in monoclonal culture (Rose & Cox 2014), which may explain why so many varieties and forms have been described. More recent studies integrating morphological and molecular data have clarified this species complex (Kermarrec et al. 2013a; Abarca et al. 2014). In our case, several sequences belonging to several separate clades were present in a single sample and after a careful examination two different morphodemes could be distinguished, which may fit the descriptions of *G. lagenula* and *G. parvulum* var. *parvulum* [morphodeme exilissimum]. It was difficult, however, to relate these two morphodemes to the different clades which were, moreover, not supported with high bootstrap values in the *G. parvulum* phylogeny. The presence of several species belonging to the *G. parvulum* species complex in a single sample has already been observed by Kermarrec et al. (2013). This example shows the limits of the method we are proposing. Therefore these sequences were not included in the reference database R–Syst::diatom.

2. What are the advantages/disadvantages associated with use of uncultured diatom HTS sequences to enrich barcode reference libraries?

Unlike macroorganisms such as aquatic insects, where the DNA barcodes of individuals can be easily Sanger sequenced from a part of their body (e.g. insect legs in Sweeney et al. 2011), diatom cells need to be isolated and cultured to get their barcodes, which is laborious and not always successful. Using HTS for natural samples encompassing several millions of diatom cells and several species makes it potentially possible to obtain thousands of barcode sequences in one HTS run.

One advantage of using HTS is shown by the examples of *Epithemia hirudiniformis, Halamphora ghanensis* and *Gomphonema clavatuloides* reported here. In each case, from 2 to 10 sequences were kept for each species in the reference database R–Syst::diatom. This gives an idea of the intraspecific genetic diversity of species. Obtaining such information with cultures is possible and has already been done (e.g. *Nitzschia palea* (Kützing) W. Smith studied by Rimet et al. (2014);

Pseudo–nitzschia pungens (Grunow ex Cleve) Hasle by Casteleyn et al. 2010) but requires intensive effort to isolate clones representing many clades.

The ease with which large numbers of sequences are generated using HTS is, however, balanced by some drawbacks. In particular, HTS produces more sequencing mistakes than classical Sanger sequencing (Paparini et al. 2015) and comparisons of the performances of different HTS technologies show different results. Loman et al. (2012) showed that MiSeq (Illumina) had the lowest error rates compared to 454 GS Junior (Roche) and Ion Torrent PGM (Life Technologies), but Ion Torrent had the highest throughput (80–100 Mb/h, compared to MiSeq 60 Mb/h and 454 9Mb/h) and 454 GS Junior gives the longest reads (up to 600 bases, compared to MiSeq with 150 bases and Ion Torrent PGM 300 bases). Such differences must be taken into account during data processing. When integrating sequences from HTS in reference databases, sequence quality criteria have to be defined. In this case, we set four criteria:

Criterion 1: only the 15–20 most often sequenced reads were selected because such sequences are more likely to be the best representatives of the population and also are likely to be sequences showing the lowest probabilities of sequencing mistakes (Bragg et al. 2013).
Criterion 2: since *rbc*L is a coding region for a gene, no indels were accepted when aligning the HTS sequences with Sanger sequences from R–Syst::diatom. Sanger sequencing is, until now, considered to be the reference technology in terms of sequencing quality (e.g. Montoya et al. 2016; Khalifa et al. 2016). Such problems were particularly common in the case of *Epithemia hirudiniformis*, where many sequences showed deletions, and were therefore not included in the reference database.
Criterion 3: since *rbc*L is a coding region, after translating the nucleotide sequence into proteins, no stop codons should appear within a coding region. Those containing stop codons must be discarded. However, this problem did not occur for the most abundant sequences in our examples.
Criterion 4: the HTS sequences should show phylogenetic neighbours corresponding to the same neighbour taxa expected from morphological observations.

3. Which material data and metadata must be stored with these sequences in barcoding libraries to ensure good traceability?

The fundamental requirement suggested by Zimmermann et al. (2014a) is that reliable identification of a taxon via DNA barcodes needs unambiguous agreement between genotype and phenotype/morphodeme with a valid binomial. Since this is not possible for HTS generated sequences, in contrast to sequences generated from unialgal cultures, we have to adapt and modify the requirements in order to guarantee a high standard for depositing these sequences.

In the case of HTS–generated sequences from biofilm samples, the material data as well as the linked metadata have to be deposited in curated collections (herbaria such as B, BM, G, P, etc. see index of herbaria in Holmgren et al. 1990) and have to be available through public scientific databases (e.g. R–Syst, AlgaTerra, GGBN, GBIF, BOLD, INSDC). This includes eDNA material deposition (e.g. DNA Bank Network / GGBN) and possible linked voucher material, information concerning the HTS methodology (e.g. DNA extraction, primers, PCR, library preparation, HTS chemistry, HTS platform, paired–end reads or not), details of the bioinformatics pipeline and the algorithms used (e.g. sequence trimming, chimera treatment, thresholds for OTU clustering, modus operandi of species assignment), and availability of the raw reads. As for reference sequences from unialgal diatom cultures, Zimmermann et al. (2016) suggest that metadata should include sampling localities and collectors, basic environmental data, high–resolution LM pictures, morphometrics, taxonomy and nomenclature, maps, literature as well as references to databases where this data is stored.

Even though a 100% unambiguous identification could not be made in the case of *Epithemia hirudiniformis,* the requirements for sound documentation are still applicable for a reference library, but the environmental origin of the sequence also has to be clearly highlighted. This also applies to the other examples, but they have to be treated with great care in order to ensure that a correlation between the presence of frustules and sequences is not just a coincidence. It has been shown in a number of cases that frequency of occurrence of a species within a sample analysed by light microscopy does not coincide with the frequency of occurrences of sequences in an environmental sample (e.g. Jahn et al. 2007). It is therefore essential that data are critically checked, the details of their creation are transparent and the pitfalls of the method discussed here are pointed out. We also recommend that sequences identified using these approaches are distinguished from those obtained from unialgal cultures in barcode libraries.

Table 2 suggests terminology to prevent any confusion when such sequences are used in metabarcoding studies.

Conclusions: limits of the proposed methodology and perspectives

The methods and examples given here are clearly related to a particular application, which is routine assessment of aquatic ecosystems where species identification from natural samples is needed. We have shown that it is possible to enrich a reference barcoding library at low cost, taking advantage of sequencing data from routine samples collected for ecological assessment. Given the HTS technology used (PGM Ion Torrent), the sequence length was 312 bp. Such sequence length has been shown to be long enough for applied topics such as the one addressed here: diatom species identification for ecological assessment (e.g. Kermarrec et al. 2014;

Zimmermann et al. 2015; Visco et al. 2015). It will be straightforward to expand such methodology to wider monitoring networks, if taxonomic experts are available who can follow the recommendations we propose here. Given that current reference barcoding libraries (and R–Syst::diatom in particular) cover only a small part of the diversity of diatoms in freshwater ecosystems, such HTS approaches may be essential if the libraries are to be completed quickly in order to be used for routine assessments.

On the other hand, one must make no mistake about the objective of this method. DNA barcodes are often inappropriate for phylogenetic studies, especially for defining deep nodes of classification trees (Hajibabaei et al. 2007). For such studies, much longer sequences are required and indeed, multigene approaches are now commonly recommended and are providing a much better understanding of diatom evolution (e.g. Ruck et al. 2016; Theriot et al. 2015; Nakov et al. 2014).

Finally, a possibility emerging from studies accumulating numerous DNA barcodes from particular species is that it can be a starting point of population genetics studies or can give indications of genetic diversity at an infraspecific level (Hajibabaei et al. 2007). Coupled with ecological, physiological and geographical information, diatom species boundaries could be re–evaluated, as was done recently for some green microalgae (Darienko et al. 2015). Diatom species descriptions have until recently been based only on morphological features. Only a few studies (e.g. Rovira 2013; Trobajo et al. 2013; Kelly et al. 2015, for the *Nitzschia inconspicua* Grunow species complex , Jahn et al. 2017 for the *Planothidium lanceolatum* group) follow the precepts of integrative taxonomy which aims to delimit species on sets of different criteria such as morphology, DNA, physiology, ecology and biogeography (Dayrat 2005). This should prompt reconsideration of the boundaries of many poorly delimited diatom species and, potentially, enhance their value for ecological assessment.

In conclusion, the approach described here offers a pragmatic and universally applicable way to enrich existing diatom reference barcode libraries with HTS generated barcodes, especially when the sequences are as well documented as classical voucher specimens, no matter which region/gene is used. Nonetheless, we believe that unialgal diatom cultures should still be the backbone of reference libraries, because this is still the method with lowest amount of error.

Acknowledgements

Sampling, microscopy and sequencing were funded by the Agence française pour la biodiversité. Sequencing was carried out in the Genome Transcriptome Facility of Bordeaux (INRA Pierroton) and we thank Alain Franc, Philippe Chaumeil, Jean–Marc Frigerio and Franck Salin for helpful discussions. We also thank DNAqua–Net (European Cost Action CA15219) which helped discussion between some of the authors.

References

Abarca, N.; Jahn, R.; Zimmermann, J. & Enke, N. (2014): Does the cosmopolitan diatom *Gomphonema parvulum* (Kützing) Kützing have a biogeography? – PLoS ONE 9: 1–18.

Barbour, M.T.; Gerritsen, J.; Snyder, B.D. & Stribling, J.B. (1999): Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish. Second edition. – EPA 841–B–99–002. US Environmental Protection Agency, Office of Water, Washington, DC.

Besse–Lotoskaya, A.; Verdonschot, P. & Sinkeldam, J. (2006): Uncertainty in diatom assessment: sampling, identification and counting variation. – Hydrobiologia 566: 247–260.

Bragg, L.M.; Stone, G.; Butler, M.K.; Hugenholtz, P. & Tyson, G.W. (2013): Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. – PLoS Comput. Biol. 9: 1–18.

Bruder, K. & Medlin, L.K. (2007): Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms (Bacillariophyta). I. The genus *Placoneis*. – Nova Hedwigia 85: 331–352.

Butcher, R.W. (1947): Studies in the ecology of rivers. IV. The algae of organically enriched water. – J. Ecol. 35: 186–191.

Casteleyn, G.; Leliaert, F.; Backeljau, T.; Debeer, A.E.; Kotaki, Y.; Rhodes, L.; Lundholm, N.; Sabbe, K. & Vyverman, W. (2010): Limits to gene flow in a cosmopolitan marine planktonic diatom. – PNAS 107: 12952–12957.

Chonova, T.; Keck, F.; Labanowski, J.; Montuelle, B.; Rimet, F. & Bouchez, A. (2016): Separate treatment of hospital and urban wastewaters: a real scale comparison of effluents and their effect on microbial communities. – Sci. Total Environ. 542: 965–975.

Cocquyt, C. (1998): Diatoms from Northern Basin of Lake Tanganyika. – Bibliotheca Diatomologica, vol. 39, J. Cramer, Berlin and Stuttgart.

Cocquyt, C.; Kusber, W.–H. & Jahn, R. (2018): *Epithemia hirudiniformis* and related taxa within the subgenus *Rhopalodiella* subg. nov. in comparison to *Epithemia* subg. *Rhopalodia* stat nov. (Bacillariophyceae) from East Africa. – Cryptogamie, Algologie 39: 1–28.

Cohn, F. (1853): Über lebendige Organismen im Trinkwasser. – Z. klin. Med. 4: 229–237.

Darienko, T.; Gustavs, L.; Eggert, A.; Wolf, W. & Pröschold, T. (2015): Evaluating the species boundaries of green microalgae (*Coccomyxa*, Trebouxiophyceae, Chlorophyta) using integrative taxonomy and DNA Barcoding with further implications for the species identification in environmental samples. – PLoS ONE 10: 1–31.

Dayrat, B. (2005): Towards integrative taxonomy. – Biol. J. Linn. Soc. 85: 407–415.

Edgar, R.C.; Haas, B.J.; Clemente, J.C.; Quince, C. & Knight, R. (2011): UCHIME improves sensitivity and speed of chimera detection. – Bioinformatics 27: 2194–2200.

Edgar, R.S. (2004): MUSCLE: multiple sequence alignment with high accuracy and high throughput. – Nucleic Acids Res. 32: 1792–1797.

European commission (2000): Directive 2000/60/EC of the European Parliament and of the Council of 23rd October 2000 establishing a framework for Community action in the field of water policy. – Official Journal of the

European Communities 327: 1–72.

European Committee for Standardisation (2014a): EN 13946 – Water quality – Guidance for the routine sampling and preparation of benthic diatoms from rivers and lakes. – 18 pp., Afnor, La Plaine St Denis, France.

European Committee for Standardisation (2014b): EN 14407 – Water quality – Guidance for the identification and enumeration of benthic diatom samples from rivers and lakes. – 13 pp., Afnor, La Plaine St Denis, France.

Evans, K.M.; Wortley, A.H. & Mann, D. G. (2007): An assessment of potential diatom "barcode" genes (*cox*1, *rbc*L, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). – Protist 158: 349–364.

Geitler, L. (1972): Sippen von *Gomphonema parvulum*, Paarungsverhalten und Variabilität pennater Diatomeen. – Österr. Bot. Z. 120: 257–268.

Gomez, F.; Lopez–Garcia, P.; Dolan, J.R. & Moreira, D. (2012): Molecular phylogeny of the marine dinoflagellate genus *Heterodinium* (Dinophyceae). – Eur. J. Phycol. 47: 95–104.

Gouy, M.; Guindon, S. & Gascuel, O. (2010): SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. – Mol. Biol. Evol. 27: 221–224.

Hajibabaei, M.; Singer, G.A.C.; Hebert, P. & Hickey, D.A. (2007): DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. – Trends Genet. 23: 167–172.

Hajibabaei, M.; Baird, D.J.; Fahner, N.A.; Beiko, R. & Golding, G.B. (2016): A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. – Phil. Trans. R. Soc., B 371: 20150330.

Hamilton, P B.; Lefebvre, K. & Bull, R. (2015): Single cell PCR amplification of diatoms using fresh and preserved samples. – Front. Microbiol. 6: 1084.

Haussmann, S.; Charles, D.F.; Gerritsen, J. & Belton, T.J. (2016): A diatom–based biological condition gradient (BCG) approach for assessing impairment and developing nutrient criteria for streams. – Sci. Total Environ. 562: 914–927.

Hebert, P.; Cywinska, A.; Ball, S.L. & deWaard, J.R. (2003): Biological identifications through DNA barcodes. – Proc. R. Soc. Lond., B 270: 313–321.

Hoffman, G.; Werum, M. & Lange–Bertalot, H. (2011): Diatomeen im Süsswasser–Benthos von Mitteleuropa. – A.R.G. Gantner, Ruggell, Liechtenstein.

Holmgren, P.K.; Holmgren, N.H. & Barnett, L.C. (eds) (1990): Index herbariorum, ed. 8. Part 1. The herbaria of the world. – 704 pp., New York Botanical Garden.

Hustedt, F. (1957): Die Diatomeenflora des Flusssystems der Weser im Gebiet der Hansestadt Bremen. – Abh. naturwiss. Ver. Bremen 34: 181–440.

Insee (2016): Estimation de la population au 1er janvier par région, département, sexe et âge de 1975 à 2015. – Internet Communication, consulted at www.insee.fr, 9 September 2016.

Jahn, R.; Abarca, N.; Gemeinholzer, B.; Mora D., Skibbe, O.; Kulikovskiy. M.; Gusev, E.; Kusber, W.-H. & Zimmermann, J. (2017): *Planothidium lanceolatum* and *Planothidium frequentissimum* reinvestigated with molecular methods and morphology: four new species and the taxonomic importance of the sinus and cavum. – Diatom Res. 32: 75–107. http://dx.doi.org/10.1080/0269249X.2017.131254

Jahn, R.; Zetzsche, H.; Reinhardt, R. & Gemeinholzer, B. (2007): Diatoms and DNA barcoding: A pilot study on an environmental sample. In: Kusber, W.H. & Jahn, R. (eds): Proceedings of the 1st Central European Diatom Meeting. – pp. 63–68. Botanic Garden and Botanical Museum Berlin–Dahlem, Freie Universität Berlin.

Jaramillo, A.; Osman, D.; Caputo, L. & Cardenas, L. (2015): Molecular evidence of a *Didymosphenia geminata* (Bacillariophyceae) invasion in Chilean freshwater systems. – Harmful Algae 49: 117–123.

Kahlert, M.; Albert, R.L.; Anttila, E.L.; Bengtsson, R.; Bigler, C.; Eskola, T.; Galman, V.; Gottschalk, S.; Herlitz, E.; Jarlman, A.; Kasperoviciene, J.; Kokocinski, M.; Luup, H.; Miettinen, J.; Paunksnyte, I.; Piirsoo, K.; Quintana, I.; Raunio, J.; Sandell, B.; Simola, H.; Sundberg, I.; Vilbaste, S. & Weckstrom, J. (2009): Harmonization is more important than experience – results of the first Nordic–Baltic diatom intercalibration exercise 2007 (stream monitoring). – J. Appl. Phycol. 21: 471–482.

Kelly, M.; Urbanic, G.; Acs, E.; Bennion, H.; Bertrin, V.; Burgess, A.; Denys, L.; Gottschalk, S.; Kahlert, M.; Karjalainen, S.M.; Kennedy, B.; Kosi, G.; Marchetto, A.; Morin, S.; Picinska–Faltynowicz, J.; Poikane, S.; Rosebery, J.; Schoenfelder, I.; Schoenfelder, J. & Varbiro, G. (2014): Comparing aspirations: intercalibration of ecological status concepts across European lakes for littoral diatoms. – Hydrobiologia 734: 125–141.

Kelly, M.G.; Trobajo, R.; Rovira, L. & Mann, D.G. (2015): Characterizing the niches of two very similar *Nitzschia* species and implications for ecological assessment. – Diatom Res. 30: 27–33.

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

Kermarrec, L.; Franc, A.; Rimet, F.; Chaumeil, P.; Humbert, J.F. & Bouchez, A. (2013b): Next–generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. – Mol. Ecol. Res. 13: 607–619.

Kermarrec, L.; Franc, A.; Rimet, F.; Chaumeil, P.; Frigerio, J.M.; Humbert, J.F. & Bouchez, A. (2014): A next–generation sequencing approach to river biomonitoring using benthic diatoms. – Freshw. Sci. 33: 349–363.

Khalifa, M.E.; Varsani, A.; Ganley, A.R.D. & Pearson, M.N. (2016): Comparison of Illumina *de novo* assembled and Sanger sequenced viral genomes: A case study for RNA viruses recovered from the plant pathogenic fungus *Sclerotinia sclerotiorum*. – Virus Res. 219: 51–57.

Khan–Bureau, D.A.; Morales, E.A.; Ector, L.; Beauchene, M.S. & Lewis, L.A. (2016): Characterization of a new species in the genus *Didymosphenia* and of *Cymbella janischii* (Bacillariophyta) from Connecticut, USA. – Eur. J. Phycol. 51: 203–216.

Ki, J.S.; Cho, S.Y.; Katano, T.; Jung, S.W.; Lee, J.; Park, B.S.; Kang, S.H. & Han., M.S. (2009) Comprehensive comparisons of three pennate diatoms, *Diatoma tenuae*, *Fragilaria vaucheriae*, and *Navicula pelliculosa*, isolated from summer Arctic reservoirs (Svalbard 79°N), by fine scale morphology and nuclear 18S ribosomal DNA. – Polar Biol. 32: 147–159.

Kolkwitz, R. & Marsson, M. (1908): Ökologie der pflanzliche Saprobien. – Ber. Deutsch. Bot. Ges. 26: 505–519.

Krammer, K. & Lange–Bertalot, H. (1986): Bacillariophyceae 1. Teil: Naviculaceae. – In: Ettl, H.; Gerloff, J.; Heynig, H. & Mollenhauer, D. (eds): Süsswasserflora von Mitteleuropa, Vol. 2/1. – 876 pp., G. Fischer, Stuttgart & New York.

Kumar, S.; Stecher, G., & Tamura, K. (2016): MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. – Mol. Biol. Evol. 33: 1870–1874.

Kusber, W.H.; Abarca, N.; Skibbe, O.; Zimmermann, J. & Jahn, R. (2012): Reference library of DNA–barcoded diatoms – A use case for publishing data via the GBIF database AlgaTerra. – In: Sabbe, K.; Van De Vijver, B. & Vyverman, W. (eds): Abstracts. 22nd International Diatom Symposium, Aula Academica, Ghent – p. 65, VLIZ Special Publication 58 (available at http://www.vliz.be/events/ids2012/ABSTRACTBOOK%20IDS%202012.pdf).

Kützing, F.T. (1844): Die kieselschaligen Bacillarien oder Diatomeen. – 152 pp., W. Köhne, Nordhausen.

Levkov, Z. (2009): *Amphora* sensu lato. – In: Lange–Bertalot, H. (ed.) Diatoms of the European Inland Waters and Comparable Habitats. – Vol. 5, 916 pp., A.R.G. Gantner, Ruggell, Liechtenstein.

Loman, N.J.; Misra, R.V.; Dallman, T.J.; Constantinidou, C.; Gharbia, S.E.; Wain, J. & Pallen, M.J. (2012): Performance comparison of benchtop high–throughput sequencing platforms. – Nat. Biotechnol. 30: 434–439.

Mann, D.G. & Vanormelingen, P. (2013): An inordinate fondness? The number, distributions and origins of diatom species. – J. Euk. Microbiol. 60: 414–420.

Markert, B.A.; Breure, A.M. & Zechmeister, H.G. (2003): Definitions, strategies and principles for bioindication/biomonitoring of the environment. – In: Markert, B.A.; Breure, A.M. & Zechmeister, H.G. (eds): Bioindicators & Biomonitors Principles, Concepts and Applications. – pp. 3–39, Elsevier, Amsterdam.

Montoya, V.; Olmstead, A.; Tang, P.; Cook, D.; Janjua, N.; Grebely, J.; Jacka, B.; Poon, A.F.Y. & Krajen, M. (2016): Deep sequencing increases hepatitis C virus phylogenetic cluster detection compared to Sanger sequencing. – Infect. Gen. Evol. 43: 329–337.

Murakami, T. & Kasuya, M. (1993): Teratological variations of *Gomphonema parvulum* Kützing in heavily polluted drainage channel. – Diatom 8: 7–10.

Nakov, T.; Ruck, E.; Galachyants, Y.; Spaulding, S.A. & Theriot, E.C. (2014): Molecular phylogeny of the Cymbellales (Bacillariophyceae, Heterokontophyta) with a comparison of models for accommodating rate variation across sites. – Phycologia 53: 359–373.

Ndiritu, G.G.; Gichuki, N.N. & Triest, L. (2006): Distribution of epilithic diatoms in response to environmental conditions in an urban tropical stream, Central Kenya. – Biodivers. Conserv. 15: 3267–3293.

Paparini, A.; Gofton, A.; Yang, R.; White, N.; Bunce, M. & Ryan, U.M. (2015): Comparison of Sanger and next generation sequencing performance for genotyping *Cryptosporidium* isolates at the 18S rRNA and actin loci. – Exp. Parasitol. 151: 21–27.

Pompanon, F.; Coissac, E. & Taberlet, P. (2011): Metabarcoding, une nouvelle façon d'analyser la biodiversité. – Biofutur 319: 30–32.

Potapova, M. & Charles, D.F. (2007): Diatom metrics for monitoring eutrophication in rivers of the United States. – Ecol. Indic. 7: 48–70.

Reichardt, E. (1999): Zur revision der Gattung *Gomphonema*. Die Arten um *G. affine/insigne, G. angustatum/micropus, G. acuminatum* sowie gomphonemoide Diatomeen aus dem Oberoligozän in Böhmen. – In: Lange–Bertalot, H. (ed.) Iconographia Diatomologica, Vol. 8. – 203 pp., A.R.G. Gantner, Ruggell, Liechtenstein.

Rimet, F. (2009): Benthic diatom assemblages and their correspondence with ecoregional classifications: case study of rivers in north–eastern France. – Hydrobiologia 636: 137–151.

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

Rimet, F.; Trobajo, R.; Mann, D.G.; Kermarrec, L.; Franc, A.; Domaizon, I. & Bouchez, A. (2014): When is sampling complete? The effects of geographical range and marker choice on perceived diversity in *Nitzschia palea* (Bacillariophyta). – Protist 165: 245–259.

Rimet, F.; Chaumeil, P.; Keck, F.; Kermarrec, L.; Vasselon, V.; Kahlert, M.; Franc, A. & Bouchez, A. (2016): R–Syst::diatom: An open–access and curated barcode database for diatoms and freshwater monitoring. – Database (Oxford) 2016: baw016: 1–21.

Rivera, S.F.; Vasselon, V.; Jacquet, S.; Bouchez, A.; Ariztegui, D. & Rimet F. (2017): Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. – Hydrobiologia 807: 37–51. https://doi.org/10.1007/s10750-017-3381-2

Rose, D. & Cox, E.J. (2014): What constitutes *Gomphonema parvulum*? Long–term culture studies show that some varieties of *G. parvulum* belong with other *Gomphonema* species. – Plant Ecol. Evol. 147: 366–373.

Rovira, L. (2013): The ecology and taxonomy of estuarine benthic diatoms and their use as bioindicators in a highly stratified estuary (Ebro Estuary, NE Iberian Peninsula): a multidisciplinary apporach. – 295 pp., PhD dissertation, University of Barcelona.

Ruck, E.; Nakov, T.; Alverson, A.J. & Theriot, E.C. (2016): Phylogeny, ecology, morphological evolution, and reclassification of the diatom orders Surirellales and Rhopalodiales. – Mol. Phylogenet. Evol. 103: 155–171.

Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; Sahl, J.W.; Stres, B.; Thallinger, G.G.; Van Horn, D.J. & Weber, C.F. (2009): Introducing mothur: open–source, platform–independent, community–supported software for describing and comparing microbial communities. – Appl. Environ. Microbiol. 75: 7537–7541.

Silva–Benavides, A.M. (1996): The epilithic diatom flora of a pristine and a polluted river in Costa Rica, Central America. – Diatom Res. 11: 105–142.

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

Stepanek, J.G. & Kociolek, J.P. (2014): Molecular phylogeny of *Amphora* sensu lato (Bacillariophyta): an investigation into the monophyly and classification of the amphoroid diatoms. – Protist 165: 177–195.

Stevenson, R.J. (2014): Ecological assessments with algae: a review and synthesis. – J. Phycol. 50: 437–461.

Stoof–Leichsenring, K.R.; Epp, L.S.; Trauth, M.H. & Tiedemann, R. (2012): Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. – Mol. Ecol. 21: 1918–1930.

Sweeney, B.W.; Battle, J.M.; Jackson, J.K. & Dapkey, T.

(2011): Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? – J. North Am. Benthol. Soc. 30: 195–216.

Takano, Y. & Horiguchi, T. (2006): Acquiring scanning electron microscopical, light microscopical and multiple gene sequence data from a single dinoflagellate cell. – J. Phycol. 42: 251–256.

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 607/608: 1293–1303.

Theriot, E.C.; Ashworth, M.P.; Nakov, T.; Ruck, E. & Jansen, R.K. (2015): Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. – Mol. Phylogenet. Evol. 89: 28–36.

Trobajo, R.; Clavero, E.; Chepurnov, V.; Sabbe, K.; Mann, D.G.; Ishihara, S., & Cox, E.J. (2009): Morphological, genetic and mating diversity within the widespread bioindicator *Nitzschia palea* (Bacillariophyceae). – Phycologia 48: 443–459.

Trobajo, R.; Rovira, L.; Ector, L.; Wetzel, C.E.; Kelly, M., & Mann, D.G. (2013): Morphology and identity of some ecologically important small *Nitzschia* species. – Diatom Res. 28: 37–59.

Vasselon V.; Domaizon I.; Rimet F.; Kahlert M. & Bouchez A. (2017): Application of high–throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? – Freshw. Sci. 36:162–177.

Visco, J.; Apothe–loz–Perret–Gentil, L.; Cordonier, A.; Esling, P.; Pillet, L. & Pawlowski J. (2015): Environmental monitoring: inferring the diatom index from next–generation sequencing data. – Environ. Sci.Technol. 49: 7597–7605.

Wachnicka, A.H. & Gaiser, E.E. (2016): Characterization of *Amphora* and *Seminavis* from South Florida, U.S.A. – Diatom Res. 22: 387–455.

Wang, Q. G.; Garritz, G.M.; Teidje, J.M. & Cole, J.R. (2007): Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. – Appl. Environ. Microbiol. 73: 5261–5267.

Zelinka, M. & Marvan, P. (1961): Zur Präzisierung der biologischen Klassifikation der Reinheit fliessender Gewässer. – Arch. Hydrobiol. 57: 389–407.

Zimmermann, J.; Abarca, N.; Enke, N.; Skibbe, O.; Kusber, W.H. & Jahn, R. (2014): Taxonomic refernce libraries for environmental barcoding: a best practice example from diatom research. – PLoS ONE 9: 1–24.

Zimmermann, J.; Glöckner, G.; Jahn, R.; Enke, N. & Gemeinholzer, B. (2015): Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. – Mol. Ecol. Res. 15: 526–542.

Zimmermann, J.; Kusber, W.H.; Droege, G. & Jahn R. (2016): GBOL2 – Increasing the accessibility of eDNA barcoding data. – GGBN Newsletter 5: 7–8.

Supplementary material

the following supplementary material is available for this article:

Table S1. Sequences used for the constrained phylogeny of *Halamphora ghanensis* in the Gouloué river, near Passamainty city (Poll7). Underlined accession numbers were deposited in the framework of this study. TCC (Thonon Culture Collection) numbers are given.

Table S2. Sequences used for the phylogeny of *Gomphonema clavatuloides* in the Songaro Mbili river near Dembeni city (Poll29). TCC (Thonon Culture Collection) numbers are given.

Table S3. Sequences used for the phylogeny of *Epithemia hirudiniformis* in the Soulou waterfall. TCC (Thonon Culture Collection) numbers are given.

Table S4. Sequences used for the phylogeny of *Gomphonema parvulum* sensu lato in the Mouala river near Mirereni city (Poll5). TCC (Thonon Culture Collection) numbers are given.

Supplement S5–S8. Fasta files.

This material is available as part of the online article (http://fottea.czechphycology.cz/contents)

Received January 26, 2017
Accepted April 11, 2017