Morphology and metabarcoding: a test with stream diatoms from Mexico highlights the complementarity of identification methods

Demetrio Mora[1,4], Nélida Abarca[1,5], Sebastian Proft[1,6], José H. Grau[2,7], Neela Enke[1,8], Javier Carmona[3,9], Oliver Skibbe[1,10], Regine Jahn[1,11], and Jonas Zimmermann[1,12]

[1]Botanischer Garten und Botanisches Museum Berlin, Freie Universität Berlin, Königin-Luise-Straße 6-8, 14195 Berlin, Germany
[2]Museum für Naturkunde Berlin, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Invalidenstraße 43, 10115, Berlin, Germany and BEGENDIV: Berlin Center for Genomics in Biodiversity Research, Königin-Luise-Straße 6-8, 14195 Berlin, Germany
[3]Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior s/n, Ciudad Universitaria, Coyoacán, 04510 Ciudad de México, México

Abstract: Diatoms are among the most commonly used bioindicators. Correct taxonomic identifications are critical to their use as bioindicators because closely related diatom species can respond differently to water physiochemical characteristics and pollutants. However, diatom identification based on morphology can be time consuming, and requires highly specialized taxonomic skills. To optimize diatom identification, DNA metabarcoding is increasingly used because it is generally less time consuming and may be more accurate than morphological identification. To date, however, neither DNA metabarcoding nor DNA barcoding diatom studies have been conducted in Mexico. Thus, we studied epilithic diatoms from streams in Central Mexico with a combination of morphological and metabarcoding techniques, and compared the diatoms identified and quantified by each method. We also assembled a barcode reference library based on clonal culturing. This library is composed of 190 strains that belong to 72 species in 24 genera. The morphological analysis of environmental samples resulted in the identification of 204 infrageneric taxa in 42 genera, but clonal culturing from the same samples retrieved 12 additional infrageneric taxa and 1 additional genus, thereby revealing concealed diversity. The metabarcoding approach resulted in the identification of 266 infrageneric taxa that belonged to 35 genera. Together, these methods detected 49 genera. Of these genera, 14 were identified only by morphology, 29 were identified by both methods, and 6 were only identified by metabarcoding. Of the 266 taxa we retrieved with metabarcoding, we confidently assigned 94 infrageneric taxa because a direct morphological or barcode sequence correlation was possible. Thirty-four of these 94 taxa were only detected with the metabarcoding method. One-fourth (23) of the assignments were only possible because of the barcode reference library we developed during this study, because there were no existing barcode sequences that matched these barcodes in the International Nucleotide Sequence Database Collaboration databases. Large disparities existed between relative abundances based on valve counts and sequence reads of the most abundant taxa, even after we corrected for cell biovolume. Overall, we conclude that the combination of morphological and molecular methods increases the detection and identification of diatoms.

Key words: DNA metabarcoding, morphological identifications, High-Throughput Sequencing (HTS), DNA barcoding, barcode reference libraries, epilithon, 18S V4 rRNA gene

Diatoms, the most diverse group of algae, are unicellular photoautotrophic eukaryotes characterized by their silica cell walls (Round et al. 1990). Diatoms are among the most commonly used biological indicators because of their species-specific response to water quality variables and pollutants (Hering et al. 2006, Kelly et al. 2009). Indeed, several local and regional monitoring programs and networks use diatom-based indices to monitor biotic integrity and water

E-mail addresses: [4]demetriomora@gmail.com; [5]n.abarca@bgbm.org; [6]sebastianproft@yahoo.de; [7]jose.grau@mfn-berlin.de; [8]nenke@scienza-berlin.de; [9]jcj@ciencias.unam.mx; [10]o.skibbe@bgbm.org; [11]r.jahn@bgbm.org; [12]j.zimmermann@bgbm.org

DOI: 10.1086/704827. Received 21 November 2017; Accepted 30 April 2019; Published online 24 July 2019.
Freshwater Science. 2019. 38(3):000–000.

quality (Kelly 1998, Prygiel 2002, Potapova and Charles 2007).

Taxonomic identification is a crucial step in recording accurate and meaningful biomonitoring results because closely related and cryptic species can have different environmental optima and tolerances (Morales et al. 2001, Poulíčková et al. 2008). However, taxonomic identification of diatoms based on morphology can be time consuming, needs in-depth knowledge of the diatom diversity in the study area, and requires awareness of the morphological plasticity species may display under varying environmental conditions and throughout their life-cycle (Zimmermann et al. 2015). DNA barcoding (Hebert et al. 2003) and DNA metabarcoding (Taberlet et al. 2012a) based on DNA barcoding have been proposed as alternatives to overcome the impediments to morphology-based identification of organisms.

DNA barcoding is a method used to identify species based on DNA sequences that are linked to morphologically identified specimens. Ideally, barcoding sequences are short enough to be sequenced in 1 Sanger sequencing run with universal primers and can unambiguously identify a species independent of its life-cycle stage (Hebert et al. 2003, Moritz and Cicero 2004, Zimmermann et al. 2011). DNA metabarcoding, or metabarcoding, is a method that can identify multiple species from bulk samples by relying on reference barcode sequences for species identification. DNA for metabarcoding can be extracted from bulk samples of soil, water, or air, containing a 'soup of biodiversity' (Taberlet et al. 2012b, Yu et al. 2012).

The standard metabarcoding approach consists of several steps that involve processing environmental samples (biofilm, water, soil, sediment) to obtain DNA sequences of the organisms present in those samples via High-Throughput Sequencing (HTS). This is followed by bioinformatics treatments that result in lists of species or Molecular Operational Taxonomic Units (MOTUs). These lists can then be used for applications such as biomonitoring. In most cases, however, the diversity measured by metabarcoding exceeds the diversity measured by morphological analyses (Zimmermann et al. 2015, Groendahl et al. 2017). In some cases, this reveals diversity that is indiscernible with morphological methods.

The choice of a barcoding marker that gives the desired taxonomic resolution is of critical importance in diatom metabarcoding, because most diatom studies require species-level resolution. In diatoms, the preferred DNA metabarcoding markers are *rbc*L and the 18S V4 region (Kermarrec et al. 2014, Zimmermann et al. 2015). The V4 region of the 18S rRNA gene is considered the universal barcode for protists including diatoms (Pawlowski et al. 2012), although the quest for the ideal marker for diatoms continues. The main constraint that results from DNA-based identification approaches based on barcodes is the natural intraspecific and intragenomic variability and interspecific divergence of the barcoding marker. This issue is particularly problematic when a single, traditionally recognized species or bioindicator taxon has a variety of genotypes at the barcoding region. This can lead to sequences that correspond to different genotypes within the same taxon clustering into different MOTUs, thereby artificially inflating the taxonomic richness relative to morphological techniques (Brown et al. 2015, Bálint et al. 2016).

In diatoms, the main sources of the barcodes used to identify sequences in metabarcoding studies are clonal cultures obtained from single cell isolations. Clonal culturing has the advantage of allowing barcode sequences to be correlated with morphologically analyzed valves (Zimmermann et al. 2014, Stachura-Suchoples et al. 2016). Unfortunately, however, diatom culturing is time-consuming. Single cell PCR amplifications may be an alternate way to obtain barcodes for taxa that are difficult to culture (Lang and Kaczmarska 2011, Chen et al. 2013, Hamilton et al. 2015). However, corroborating taxon identity with single-cell amplification is difficult because the valves of the isolated cells are normally destroyed in the DNA extraction process. According to Skibbe et al. (2018), however, it is sometimes possible for several genetically identical cells to be isolated, for example by isolating various cells originating from the same stalks (e.g., *Gomphoneis tegelensis* R. Jahn et N. Abarca), which can provide enough material for both DNA extractions and morphological examinations. A recent study proposed the use of HTS data as an additional way to obtain barcodes, and suggested criteria to ensure the barcodes truly correspond to the species observed by microscopy (Rimet et al. 2018).

To date, neither eDNA metabarcoding nor DNA barcoding studies have been conducted for diatoms in Mexico, despite the potential use of these methods in freshwater diversity studies and in biomonitoring. To compare the performance of morphology and metabarcoding in the identification and quantification of diatom abundances, our objectives were to: 1) compare the number of taxa retrieved by morphological analysis and metabarcoding of environmental samples, 2) create a regional vouchered barcode reference library to aid taxonomic assignments from sequences derived from the metabarcoding approach, 3) compare taxon abundances derived from morphology and metabarcoding approaches, and 4) test the suitability of HTS data as a source of barcode sequences.

METHODS

General approach

The goal of this study was to compare the performance of morphology and metabarcoding methods in the identification and quantification of diatom species. We did this by analyzing epilithic diatoms from streams collected from Central Mexico. For the morphological approach, we used the data from light microscopy (LM) and scanning electron microscopy (SEM) observations obtained by Mora

et al. (2017) with further LM and SEM examinations that improved taxon detection and identification. For metabarcoding, we sequenced DNA by HTS and assigned taxa with the bioinformatic pipeline MetBaN (Proft et al. 2017) in combination with a phylogenetic-based coalescent model. We also generated the first barcode reference library of diatoms from Mexico based on clonal culturing to increase the number of reference sequences for identifying taxa to species level with the metabarcoding approach. We compared the abundances of the most common taxa obtained by morphology (LM counts) and metabarcoding (sequence read numbers). Finally, we examined the potential of retrieving barcode sequences from our HTS dataset. For the comparison of morphology and metabarcoding, we analyzed a subset of 18 samples from the 42 samples collected by Mora et al. (2017). For clonal culturing, we used all 42 samples.

Study area

The Lerma-Chapala River Basin is located in Central Mexico and covers an area of 53,590 km^2. It lies within 2 biodiversity hotspots—Mesoamerica and the Madrean Pine-Oak Woodlands (Myers et al. 2000, Cotler et al. 2006). It is geologically and climatically heterogeneous and has well defined rainy (June–October) and dry seasons (November–May). This basin is one of the most important areas in the country for agriculture and industry and has a population of >15 million inhabitants, but it is also one of the most environmentally degraded basins in the country (Aparicio 2001, Wester et al. 2005).

The studied streams (Fig. 1, Table 1) can be divided into 3 groups according to the physical and chemical composition of their waters (Mora et al. 2017): streams in the 1^{st} group have acidic waters and low specific conductivity; the 2^{nd} group has circumneutral waters, low specific conductivity, and the highest P concentration on average among the 3 groups; the 3^{rd} group is characterized by well mineralized waters with circumneutral waters and the lowest average N concentrations of the 3 groups.

Sampling

Composite epilithon samples were collected by Mora et al. (2017) by brushing cobbles across a transversal section of the streams with disposable toothbrushes, suspending the brushed material in a total volume of 60 ml. Samples were then homogenized and divided into 3 subsamples of 20 ml each, which were used for 3 different purposes: I) frozen (−24°C) for HTS, II) no prior treatment for the establishment of clone cultures to build the barcode library, and III) fixed in 70% alcohol for morphological analyses.

Morphological analysis of environmental samples

Subsamples (subsample group III) were cleaned by heating with 35% hydrogen peroxide at 80°C, followed by a series of rinses with distilled water. Permanent slides were mounted with the high refraction index mounting medium Naphrax®. Observations were performed by LM and SEM. For SEM observations, cleaned material was mounted on stubs and sputter-coated with gold-palladium. Further taxa were identified to the lowest taxonomic level possible with the same identification references cited in Mora et al. (2017). Quantification of taxon abundance was done by counting a minimum of 500 valves in LM. For a detailed description of these methods see Mora et al. (2017).

Barcode reference library of the Lerma-Chapala River Basin

Isolation, culturing, and harvesting of clonal cultures We isolated single-cells from aliquots of each environmental subsample (subsample group II) with micro-capillary glass pipettes under either a LM or a stereo LM. We used both high (LM) and low (stereo LM) magnification to increase the size diversity of the taxa isolated. Prior to isolating cells, we diluted the samples to decrease organism density and make it easier to isolate individual cells. We also did a series of isolation and re-isolation on microscope slides to ensure single-cell isolations. The isolated cells were then transferred to 5-cm diameter Petri dishes that contained culture medium (Alga-Gro; Carolina Biological Supply Company, Burlington, North Carolina). Most cells were grown on culture medium at the concentration recommended by the manufacturer, but we also used culture media that was either ½ or ¼ of the manufacturer-recommended concentration to increase the number of taxa that grew in our cultures (Ferris et al. 1996, Connon and Giovannoni 2002). The cultures were grown in Memmert® Growth Chambers (Memmert, Schwabach, Germany) at 17 to 20°C with a 12h day: night photoperiod. After a successful clonal culture had been established, i.e., growing at exponential phase and noncontaminated by other diatom or protist cells, the culture was divided into 3 subsamples to be used for A) DNA extraction, B) a reserve, and C) morphological analysis. Each of the subsamples was then transferred to a new Petri dish, maintained in the growth chamber for 1 to 3 wk, and subsequently harvested.

Molecular analysis of clonal cultures We transferred the cultured material from subsample group A to 15-mL plastic centrifuge tubes for molecular analysis. We then centrifuged the tubes at 2000 rpm for 10 min, removed the supernatant, and transferred the pellets to 1.5-mL tubes. We extracted the DNA with a NucleoSpin® Plant II Mini Kit (Macherey and Nagel, Düren, Germany) following manufacturer instructions. We checked the DNA fragment size and concentrations with gel electrophoresis (1.5% agarose gel) and NanoDrop® (Peqlab Biotechnology LLC; Erlangen, Germany) and stored the DNA samples at −20°C for future

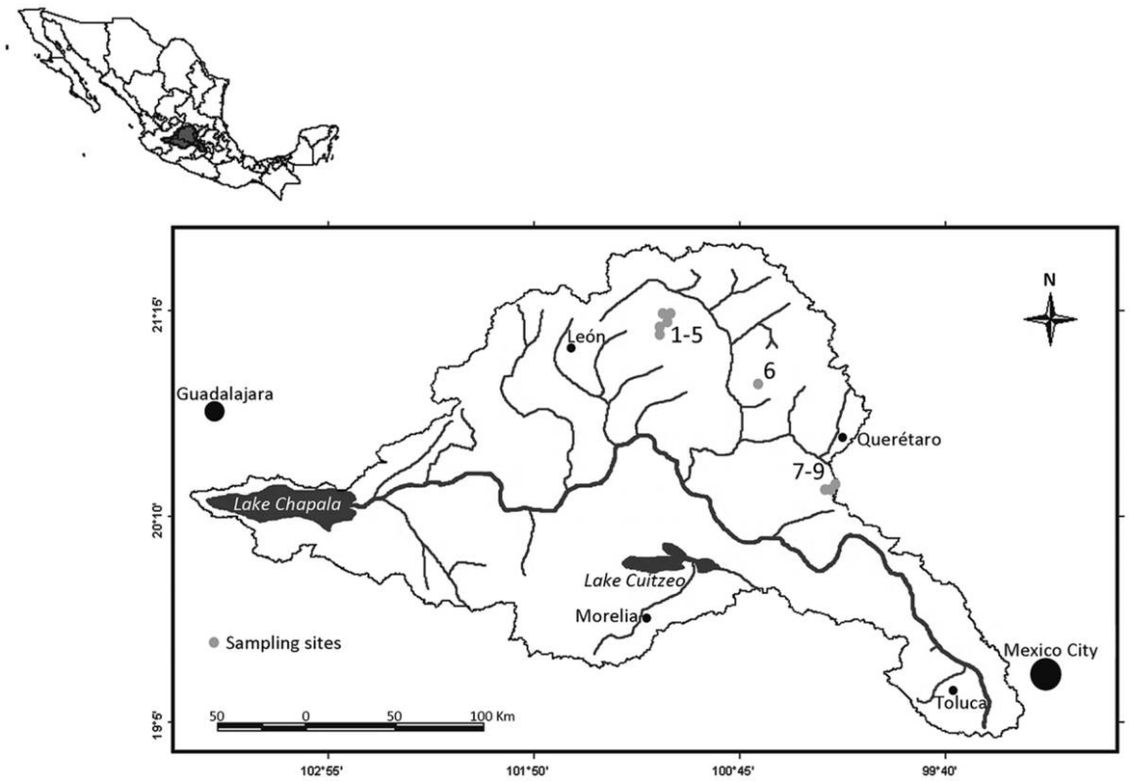

Figure 1. The location of the Lerma-Chapala Basin (dark gray area in the map of Mexico) and the locations of the 9 sites sampled within the basin (light gray dots). The numbers next to the light gray dots refer to the name of the sampling site in Table 1.

use. We amplified the V4 region of the 18S rRNA gene, which has ~390 to 410 base pairs (bp), with the primers and PCR regime from Zimmermann et al. (2011). We visualized PCR products in 1.5% agarose gel and cleaned them with MSB Spin PCRapace® (Invitek LLC, Berlin, Germany) following manufacturer instructions. We measured DNA concentrations with NanoDrop® (Peqlab Biotechnology) and normalized samples to a total DNA content >100 ng/μL for sequencing. M13 tails were used as sequencing primers following Zimmermann et al. (2014) and Ivanova et al. (2007). Sanger sequencing was conducted by Starseq® (GENterprise LLC; Mainz, Germany). We edited the sequences in PhyDE version 0.9971 (Müller et al. 2005). The final DNA sequences were submitted to the ENA (European Nucleotide

Table 1. Streams sampled in the Lerma-Chapala Basin, Central Mexico, including name, geographical coordinates, elevation (m asl) and the numbers assigned to samples collected in each stream. Samples 1 to 9 were taken during the dry season (February 2014), and samples 10 to 18 during the rainy season (September 2014).

	Latitude (N)	Longitude (W)	Elevation (m asl)	Sample
1. La Mesa	21°05′28.69″	101°08′18.98″	2215	1, 10
2. Calvillo	21°06′50.40″	101°08′04.10″	2138	2, 11
3. Peña Colorada	21°09′03.84″	101°05′58.96″	2110	3, 12
4. Paredones	21°11′20.60″	101°06′53.40″	2089	4, 13
5. La Laborcilla 1	21°11′04.70″	101°06′14.60″	2076	5, 14
6. El Membrillo	20°50′21.22″	100°38′43.46″	2114	6, 15
7. Los Ailes 1	20°19′58.72″	100°15′17.09″	2358	7, 16
8. Laguna de Servín 1	20°18′18.10″	100°17′38.10″	2409	8, 17
9. Laguna de Servín 2	20°18′45.20″	100°17′25.60″	2409	9, 18

Archive) (http://www.ebi.ac.uk/ena/) with the phyton package *EMBL2checklists* (Gruenstaeudl and Hartmaring 2019). ENA accession numbers are LS975140-LS975324.

We centrifuged subsample group B and transferred them to 1.5-mL tubes. We stored the tubes in the freezer (−24°C) as reserve material in case further DNA extractions were necessary.

Morphological analysis of clonal cultures We separately transferred the cultivated material from the C subsamples to 15-mL plastic centrifuge tubes and filled them with 35% hydrogen peroxide to oxidize the organic material. After 2 d, we removed the peroxide residue by rinsing the samples with distilled water 4 ×—an initial rinse followed by 3 additional rinses every other day. We used the cleaned samples to make 1 permanent slide per subsample with Naphrax®. We examined and photographed the diatoms on the permanent slides under a Zeiss Axio Imager M2 LM that was connected to a camera (AxioCam HRc; Zeiss, Oberkochen, Germany).

Aliquots of cleaned, cultivated material were air dried, mounted on stubs, and examined under a Hitachi FE 8010 (Hitachi, Tokyo, Japan) scanning electron microscope (SEM) operated at 1.0 kV. Taxon identification was conducted with the same identification references cited in Mora et al. (2017).

DNA metabarcoding by HTS

Subsample group I from the field were defrosted, transferred to 15-mL plastic centrifuge tubes, and centrifuged at 5000 rpm for 5 min. We then removed the supernatant and transferred the pellets to 1.5-mL tubes. We extracted the DNA with a NucleoSpin® Plant II Mini Kit (Macherey and Nagel) following manufacturer instructions. We quantified DNA concentrations with a Qubit 2.0 fluorometer (Invitrogen, Carlsbad, California) and adjusted the volume to a concentration of 20 ng/μL. We used PCR to amplify the V4 region (18S) with the Nextera primers DIV4for: 5′-GCGGTAATTCCAGCTCCAATAG-3′ and DIV4rev3: 5′-CTCTGACAATGGAATACGAATA-3′ following Zimmermann et al. (2011) with a modification for 300-bp paired-end sequencing for Illumina MiSeq (Visco et al. 2015). PCR amplifications were done in duplicate for each sample, and each amplification had a total volume of 25 μL that was comprised of 0.5 μL dNTP mix (25 mM each dNTP), 0.25 μL BSA (10 mg/mL), 0.25 μL DMSO, 1 μL of each forward and reverse primers (10 pm/μL), 0.4 μL of Herculase II Fusion DNA Polymerase (Agilent Technologies Inc., Santa Clara, California), 5 μL Herculase II reaction buffer, 1 μL of template DNA (20 ng/μL), and 15.6 μL of HPLC grade water. The PCR regime included an initial denaturation at 94°C (2 min), 35 cycles of denaturation at 94°C (45 seconds), annealing at 52°C (45 s), elongation at 72°C (1 min), and a final elongation at 72°C (10 min). We visualized the PCR products with electrophoresis on 1% agarose gels and pooled duplicate PCR products to make a final volume of 50 μL. Aliquots of 25 μL of the amplicons were purified with HighPrep PCR paramagnetic beads (Magbio Genomics, Gaithersburg, Maryland).

We then did a 2nd PCR (indexing PCR) on the purified samples to ligate a unique combination of tags to the 5′ end of the primer. The indexing PCR reactions of 25 μL were comprised of 0.25 μL dNTP mix, 1 μL DMSO, 0.625 μL of each primer, 0.25 μL of Herculase, 5 μL Herculase II reaction buffer, 10 μL of template DNA, and 7.25 μL of HPLC grade water. We started the indexing PCR regime with denaturation at 94°C (2 min), 8 cycles of denaturation at 95°C (20 s), annealing at 52°C (30 s), elongation at 72°C (30 s), and a final elongation at 72°C (3 min). We purified the PCR products with HighPrep PCR paramagnetic beads and quantified them with Quant-iT PicoGreen dsDNA Assay Kits (Invitrogen). We prepared the library with MiSeq Reagent Kit V3 (Illumina, San Diego, California) following manufacturer instructions, such that samples were normalized to equal nM DNA concentrations. We then pooled the samples, denatured them to 4 nM, diluted them to 20 pM, mixed them with 5% denatured and diluted PhiX (30 μL of PhiX and 570 μL of library), and loaded them onto the MiSeq cartridge.

We used MetBaN 1.01 (Proft et al. 2017) for bioinformatic analyses. MetBaN is a bioinformatics pipeline that implements a modular and flexible phylogenetic based species delimitation approach by streamlining metabarcoding and phylogenetic software packages. Within MetBaN, we only used the 1st modules, mostly from the *OBITOOLS* package (Boyer et al. 2016). We first merged the samples that consisted of paired-end reads with the *illuminapairedend* module, and retained only the merged reads that were >150 bp and had complete primer sequences. We then removed the primers with the *ngsfilter* module function and merged identical sequences to prevent redundant classifications. Singleton, chimeric, and low-quality sequences were then filtered out. Finally, we pooled all filtered sequences and clustered them into MOTUs at a 6 bp difference identity threshold. Subsequent taxonomic assignments were done by matching the sequences to the EMBL nucleotide sequence database (Kanz et al. 2005). MOTUs without hits were retained as unclassified. Demultiplexed fastq files were added to the Zenodo repository (https://doi.org/10.5281/zenodo.1318593).

We then divided our dataset of EMBL-assigned sequences into smaller datasets. These datasets were created to refine those taxonomic assignments with a phylogenetic-based coalescent model approach (PCMA) after Zimmermann et al. (2015). This method identifies taxonomic boundaries from the variation in branching rates of a phylogenetic tree (Monaghan et al. 2009). We constructed 12 datasets:

Achnanthidiaceae, Bacillariaceae, centrics, Cocconeidaceae, Cymbellales, Eunotiaceae, Fragilariophycidae, Mastogloiales, Sellaphoraceae-Pinnulariaceae-*Caloneis-Mayamaea*-Stauroneidaceae, Naviculales (excluding Sellaphoraceae-Pinnulariaceae-*Caloneis-Mayamaea*-Stauroneidaceae), Surirellales-Rhopalodiales, and Bacillariophyta for genus-unassigned sequences. Each dataset contained the environmental sequences generated by HTS, reference sequences from the barcode reference library we made via clonal culturing, reference sequences from the BGBM Diatom Sequence Reference Database (unpublished), and annotated diatom sequences from the NCBI nucleotide database. We aligned the datasets in the software MEGA version 6.06 (Tamura et al. 2013), which uses the MUSCLE (Edgar 2004) alignment algorithm following Zimmermann et al. (2015). We then visualized and manually improved the alignments in PhyDE version 0.9971 (Müller et al. 2005).

We conducted phylogenetic analyses on those datasets with Maximum Likelihood (ML), as implemented by RAxML version 8 (Stamatakis 2006, 2014, Stamatakis et al. 2008) in the CIPRES platform (Miller et al. 2010). We used a model of sequence evolution that was general time reversible, had a gamma distribution (Γ), and included an estimate of the proportion of invariable sites I (Tavaré 1986). We did 1000 replicates of this model for bootstrap analysis. We then used FigTree version 1.4.2 to draw the phylogenetic trees (Rambaut 2014). To taxonomically assign MOTUs unambiguously at species level, we only considered well-supported clades (≥60% bootstrap support) that were also correlated to morphological data (from environmental samples or clonal cultures). To assign less supported clades (<60% bootstrap support) we followed Zimmermann et al. (2015): sequences in clades clustered together with reference sequences from the International Nucleotide Sequence Database Collaboration (INSDC) databases (i.e., DDBJ, EMBL-EBI and NCBI) but with no morphological match were given the abbreviation cf. (*confer*) before the epithet, e.g., *Anomoeoneis* cf. *sphaerophora*; and sequences in distinct clades were considered as unspecified members of a genus (e.g., *Tryblionella* sp.). We used a 95% identity threshold to assign distinct genera after our experience setting this threshold at different values because the 95% threshold balanced the actual intrageneric variation of the amplified region and the artefactual variation created during sequencing (Brown et al. 2015).

We visualized how well morphology (LM and SEM of environmental samples and cultures) and DNA metabarcoding were able to identify taxa with Venn diagrams.

Comparison of abundance data

We evaluated whether the morphological and metabarcoding methods gave similar estimates of relative diatom abundance by comparing the abundances of the 5 most common taxa found with each method. To make the results from the 2 methods comparable, we first transformed the valve counts and sequence reads into relative abundances. After observing large disparities between the abundances found by each method, we assessed whether applying Correction Factors (CFs) for biovolume to the metabarcoding data improved abundance estimates compared to abundances obtained by morphology. We used the CFs calculated by V. Vasselon et al. (unpublished), even though there were taxonomic differences, because the diatoms they studied were similar in size and therefore in biovolume. We used the CF for *Achnanthidium minutissimum* for *Achnanthidium* sp. 1+5 (we merged *A.* sp. 1 and sp. 5 in our analyses because the barcoding marker did not differentiate these 2 distinct morphodemes) and for *Achnanthidium* cf. *tropicocatenatum* because these taxa are similar in size. We used the CF for *Cocconeis placentula* for *Cocconeis* sp. 2 and the CF for *Navicula cryptotenella* for *Navicula notha*, also because of size similarities.

Retrieval of barcode sequences from HTS data

We explored the potential to obtain barcode sequences for individual taxa from our HTS data following some of the guidelines from Rimet et al. (2018) for the *rbc*L gene. Sequences that are good candidates for barcodes include those that: 1) are among the most abundant in a sample, 2) are phylogenetic neighbors of the same neighbor taxa expected from morphological observations, and 3) have neither indels nor stop codons. The 18S rRNA gene is a non-coding region, unlike *rbc*L, so we relaxed the criteria requiring no indels to a maximum of 1 indel after aligning the putative barcode sequences with sequences from closely related taxa. We followed the same protocols for sequence alignment, phylogenetic tree construction, and tree drawing that we followed for the PCMA.

RESULTS

Barcode reference library of the Lerma-Chapala River Basin

We isolated 111 cultures (strains) from the 18 samples we used to compare morphology and metabarcoding in this study. These 111 cultures corresponded to 46 infrageneric taxa in 21 genera. In combination with the strains cultured from all 42 samples taken by Mora et al. (2017), we cultivated a total of 190 strains that belong to 72 infrageneric taxa in 24 genera (Table S1). Of the 190 strains we produced, 100 strains have sequences that are novel for the 18S V4 rRNA gene, because no identical sequences were available in the International Nucleotide Sequence Database Collaboration (INSDC) databases (DDBJ, EMBL-EBI and NCBI). Of the 24 total genera we identified, 5 genera were underrepresented in the INSDC databases because they had <10 entries for the 18S locus. Of those 5 underrepresented genera, the 9 sequences generated from the 9 strains of

Table 2. Infrageneric taxa detected by morphology and metabarcoding in streams of the Lerma-Chapala Basin, Central Mexico. M1 = infrageneric taxa identified per genus during light microscopy counts of 500 valves. M2 = taxa identified after the valves counts by either light microscopy (LM) or scanning electron microscopy (SEM). R= taxa identified by LM and SEM from cultures of the established barcode reference library in this study. M3 = taxa detected across combined morphological categories (M1, M2 and R). HTS = infrageneric taxa identified per genus by DNA metabarcoding from High-Throughput Sequencing data.

Genus	M1	M2	R	M3	HTS
Achnanthes	–	1	1	1	–
Achnanthidium	8	10	3	10	14
Amphora	1	1	–	1	2
Anomoeoneis	–	–	–	–	1
Brachysira	3	5	1	5	1
Caloneis	4	5	1	5	1
Chamaepinnularia	2	2	–	2	–
Cocconeis	2	3	–	3	10
Craticula	4	5	–	5	8
Cyclostephanos	1	1	–	1	–
Cyclotella	2	2	–	2	3
Cymbella	–	1	–	1	4
Cymbopleura	1	1	–	1	2
Diadesmis	1	1	1	1	1
Diatoma	–	–	–	–	3
Encyonema	8	9	1	9	9
Encyonopsis	2	3	–	3	–
Eolimna	3	3	–	3	–
Epithemia	3	3	–	3	3
Eunotia	4	12	1	12	7
Fistulifera	1	1	1	1	8
Fragilaria	3	5	2	5	17
Frustulia	2	4	–	4	–
Geissleria	1	1	–	1	1
Gomphonema	15	20	6	22	22
Halamphora	3	3	–	3	–
Humidophila	1	1	–	1	–
Iconella	–	1	–	1	1
Luticola	3	5	–	5	–
Mayamaea	3	3	2	4	7
Melosira	–	–	–	–	1
Navicula	14	15	2	16	42
Navigiolum	1	1	–	1	–
Neidium	–	3	–	3	–
Nitzschia	23	29	5	30	48
Nupela	1	2	1	2	–
Pinnularia	6	13	3	14	7
Planothidium	4	4	2	4	7
Pseudofallacia	1	1	–	1	–
Reimeria	1	1	–	1	4
Rhopalodia	1	1	–	1	2
Sellaphora	11	14	8	19	7
Simonsenia	–	–	1	1	–
Stauroneis	1	3	1	3	1
Stephanodiscus	–	–	–	–	1
Surirella	1	3	1	3	4
Thalassiosira	–	–	–	–	1
Tryblionella	–	–	–	–	2
Ulnaria	2	2	2	2	14
Total	148	204	46	216	266

Simonsenia cf. *delognei* (Grunow) Lange-Bertalot were the 1st records of this genus for the 18S locus in the INSDC databases. The 2 sequences generated for *Brachysira altepetlensis* D. Mora, R. Jahn et N. Abarca were added to the single sequence for this genus in INSDC databases. Regarding *Diademis* Kützing, 3 sequences of the genus were already in INSDC, so our study contributed 4 new sequences generated from the 4 strains cultured of *Diadesmis confervacea* Kützing. The sequence generated for *Nupela wellneri* (Lange-Bertalot) Lange-Bertalot added to the 2 sequences for this genus in INSDC databases. Finally, the 2 sequences generated here for *Tryblionella* W. Smith, one for *Tryblionella calida* (Grunow) D.G. Mann and the other for *Tryblionella hungarica* (Grunow) D.G. Mann, added to 6 existing sequences for the genus in INSDC.

Morphological analysis of environmental samples

We found a total of 204 infrageneric taxa (species and varieties) in 42 genera from the data analyzed by Mora et al. (2017) with further observations made in the present study that improved taxon detection and identification (Table 2, Fig. 2A). From that total, we found 148 infrageneric taxa in 38 genera in the LM count data to determine relative abundances. The additional 56 infrageneric taxa were found by scanning the whole slides under LM to look for rare taxa, and also by SEM examinations (Table 2, Fig. 2A). Among the additional 56 infrageneric taxa, we detected 4 genera not recorded in the 38 genera from the LM counts. Those 4 genera were *Achnanthes* Bory and *Neidium* Pfitzer, which were only observed under LM, and *Cymbella* Agardh and *Iconella* Jurilj, which were found with both LM and SEM (Table 2). Of the counted taxa, the 5 most abundant across all samples were, in decreasing order, *Achnanthidium* sp. 5, *Achnanthidium* sp. 1, *Gomphonema parvulum* (Kützing) Kützing, *Cocconeis* sp. 2, and *Achnanthidium* cf. *tropicocatenatum* Marquardt, C. E. Wetzel et Ector (Table S2).

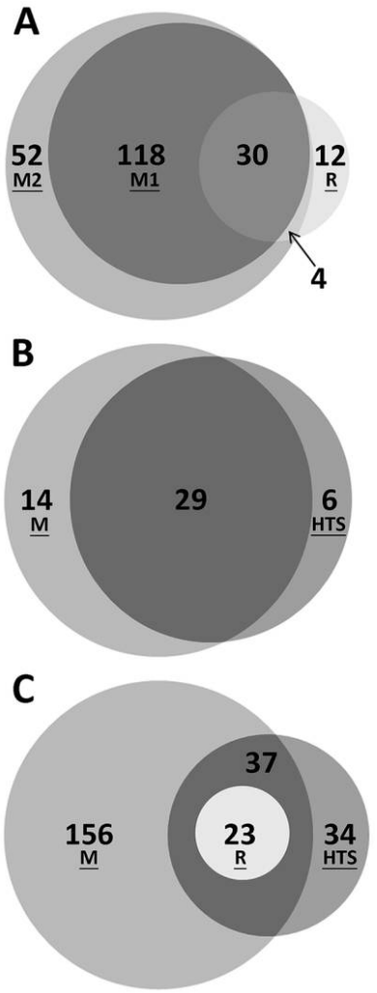

Figure 2. Venn diagrams comparing the performance of morphology and DNA metabarcoding in diatom identifications. A.—Morphological richness across all environmental samples and clonal cultures. M1 = infrageneric taxa identified by counting 500 valves per sample under light microscopy (LM). M2 = infrageneric taxa identified after additional scanning of the slides after the 500-valve counts were completed under LM, as well as taxa observed during SEM examinations. R = infrageneric taxa identified from clonal cultures isolated from the 18 samples that were the focus of this study. B.—Genera identified by morphology (M) and metabarcoding (HTS). C.—Infrageneric taxa identified by morphology and metabarcoding (only assigned taxa to species level from metabarcoding shown). M = taxa identified by morphology, HTS = taxa identified by metabarcoding, and R = taxa retrieved by metabarcoding whose taxonomic assignment was only possible with the reference library of barcode sequences established in this study. For a color version of this figure see Fig. S1.

Morphological diversity detected by clonal culturing

We identified 46 infrageneric taxa in the clonal cultures, but only 34 of them were also found in the microscopy examinations of environmental samples, which means that 12 taxa were found only after culturing. This result increased the total richness to 216 infrageneric taxa in 43 genera (Fig. 2A, Table 2) from the 204 infrageneric taxa in 42 genera identified only from the environmental samples. *Simonsenia* strains were neither observed by LM nor by SEM from environmental samples, adding 1 genus to the 42 found in environmental samples.

DNA metabarcoding by HTS

The Illumina MiSeq sequencing run generated 2,738,628 reads from our full data set. After we deleted singleton and chimeric reads, we retained 1,156,360 quality reads. Diatoms comprised 3.8% (43,703) of the quality reads, according to the BLASTn (Altschul et al. 1990) of the EMBL nucleotide database. We obtained a total of 2181 MOTUs from 43,703 diatom reads.

The phylogenetic-based coalescence model approach (PCMA) resulted in 350 taxonomic units (hereafter referred to as infrageneric taxa). To further remove potential sequencing noise, we removed infrageneric taxa that were comprised of only 1 doubleton or 1 tripleton, and had no association with morphology (no valves observed) or a reference sequence (no reference sequences in the corresponding tree branch). These removals reduced the number of taxa to 331 in 35 genera. We additionally removed 65 of these infrageneric taxa because we could not assign them to a described genus based on the 95% identity threshold, which resulted in a richness of 266 infrageneric taxa in 35 genera. From the remaining 266 infrageneric taxa we were able to confidently assign 94 of them unambiguously because a morphological correlation was possible, or a correlation to a sequence from our own barcode reference database or to sequences from NCBI in well-supported clades ($\geq$ 60 bootstrap support).

The 5 most abundant taxa across all samples were, in decreasing order, *Gomphonema parvulum*, *Navicula notha* J.H. Wallace, *Cocconeis* sp. 2, *Nitzschia palea* (Kützing) W. Smith, and *Ulnaria* cf. *ulna* (Nitzsch) Compère (Table S2).

Comparison of diatom composition from morphology and metabarcoding

The morphological analysis recovered 216 taxa, in contrast to the 266 we recovered with metabarcoding. We found 43 genera based on morphological identification and 35 based on metabarcoding. In total, we found 14 genera only by morphological identification, 29 with both morphological identification and metabarcoding, and 6 only by metabarcoding (Fig. 2B, Table 2). The combination of the total morphological richness of 216 taxa with the 94 infrageneric taxa unambiguously assigned to species level by metabarcoding resulted in a total of 250 infrageneric taxa (Fig. 2C). We detected 60 of these taxa with both methods and 34 only by metabarcoding. The barcode reference library we present here allowed the assignment of 23 infrageneric taxa by

metabarcoding. Without our reference library, those 23 infrageneric taxa would have been left unassigned because no matching reference sequences were available in the INSDC databases before our study (Fig. 2C).

Comparison of relative abundances

We compared the relative abundances of the 5 taxa that were most common in either the morphology and metabarcoding analyses, resulting in 7 taxa including *Achnanthidium* cf. *tropicocatenatum*, *Achnanthidium* sp. 1+5, *Cocconeis* sp. 2, *Gomphonema parvulum* sensu lato (s.l.), *Navicula notha*, *Nitzschia palea* s.l., and *Ulnaria* cf. *ulna*. We pooled the abundances of *Achnanthidium* sp. 1 and *A.* sp. 5 because they could not be differentiated in the metabarcoding approach (identical barcode sequences). We treated *Gomphonema exilissimum*, *G. lagenula*, and *G. parvulum* as a single taxon (*Gomphonema parvulum* s.l.) because the barcoding marker did not differentiate between them. We treated *Nitzschia palea* (*N. palea*, *N. palea* var. *debilis*, and *N. palea* var. *tenuirostris*) in a broad sense for the same reason. Based on graphs of the relative abundance of each species at each site, disparities in abundances were apparent between both methods (Fig. 3). Applying CFs to the metabarcoding abundance data affected the disparities between these methods differently (Fig. 4). On one hand, the differences in abundance between morphology and metabarcoding decreased in *A.* cf. *tropicocatenatum* (from 5.6% to 5.2%), *Achnanthidium* sp. 1+5 (from 27% to 26.1%), *Navicula notha* (from 7.5% to 0.2%), *Nitzschia palea* s.l. (from 5.8% to 0.3%), and *Ulnaria* cf. *ulna* (from 5.5% to 1%) (Fig. 4). On the other hand, the differences in abundance from the 2 methods increased for *Cocconeis* sp. 2 (1.6% to 6.9%), and *G. parvulum* s.l. (from 4.9% to 9.3%) (Fig. 4).

HTS data as a source of barcodes

After in-depth examination of HTS data obtained in our study, we proposed barcode sequences of the V4 region (18S) for 2 taxa, *Iconella delicatissima* (F. W. Lewis) Ruck et Nakov (Fig. 5A, B) and *Navicula notha* Wallace (Fig. 5C–G).

Iconella delicatissima (F. W. Lewis) Ruck et Nakov (Fig. 5 A, B) ≡ *Stenopterobia delicatissima* (Lewis) Van Heurck

The genus *Iconella* has been recently resurrected to accommodate *Stenopterobia* and the 'robustoid' members of *Surirella* and *Campylodiscus* (Ruck et al. 2016a, b, Jahn et al. 2017a). We found *Iconella* sequences only in sample 18. The morphological examination of this sample confirmed the presence of *Iconella delicatissima*. *Surirella angusta* was the only other member of the Surirellales found in this sample, but this species belongs to *Surirella* sensu stricto (Ruck et al. 2016b). A total of 52 sequence reads were obtained for *Iconella delicatissima* from the 6873 total reads obtained from sample 18 (Laguna de Servín 2), making it the 13th most abundant taxon in this sample. After BLASTing this sequence on NCBI, the 3 most similar sequences belonged to the genus *Stenopterobia*, with 98% similarity to the 2 most similar sequences (*Stenopterobia pumila* and *S. curvula*) and 95% similarity to the 3rd most similar sequence (*Stenopterobia* sp. 50). Other sequences with 95% similarity include 1 sequence of *Surirella* sp. and 2 sequences of *Campylodiscus levanderi*. After we aligned our sequence with 17 sequences from the resurrected genus *Iconella* (Ruck et al. 2016b), including those of the aforementioned taxa, our sequence had only 1 indel in comparison with the 3 available sequences of *Stenopterobia* and no indels compared with other included surirelloid species. In the unrooted ML phylogenetic tree, *Iconella delicatissima* clusters with *Stenopterobia pumila*, *S. curvula*, and *S.* sp. 50 in a well-supported clade (bootstrap value = 96%) (Fig. S2). This new barcode sequence has been submitted to the ENA under accession number LS990839.

Navicula notha Wallace (Fig. 5 C–G)

We found sequences of *N. notha* in 17 out of the 18 samples we analyzed with HTS. The read abundance of this taxon across all samples was the 2nd highest after *G. parvulum* s.l. We also found *N. notha* in our morphological examinations of these samples. We used samples 4, 5, and 14 as the barcode sources, because *N. notha* was the only *Navicula* representative found from valve counts in these samples. Further, *N. notha* reached high relative valve abundances in these samples (2.9% in sample 4; 4.5% in sample 5; 5.4% in sample 14). The relative abundance of sequence reads in those samples was as high as 16% (sample 4), 34% (sample 5), and 22% (sample 14). The sequences from these 3 sites were identical, so we only submitted 1 sequence to the ENA (from site 5, accession number LS990785). The closest sequence that corresponded to *N. notha* on the NCBI database was *Navicula cryptotenelloides*, which had a 4 bp difference (99% similarity). Other sequences with high similarity (98%) to our sequence included *Navicula cryptotenella* and *N. reinhardtii*. However, we did not observe *N. cryptotenelloides* or *N. reinhardtii* in our samples, so we are confident that our sequence does not correspond to these taxa. Further, we are confident that our sequence does not correspond to *N. cryptotenella* even though we observed this taxon by microscopy because it was only present at 1 site, it was not recorded during the 500-valve counts to calculate relative abundances, and we only detected it after additional scans of the slides were performed. We aligned our *N. notha* sequence to 29 *Navicula* sequences present in the NCBI nucleotide database. This alignment had no indels. *Navicula notha* clustered with *N. cryptotenelloides*, *N. cryptocephala*, *N. cryptotenella*, and *N. reinhardtii*

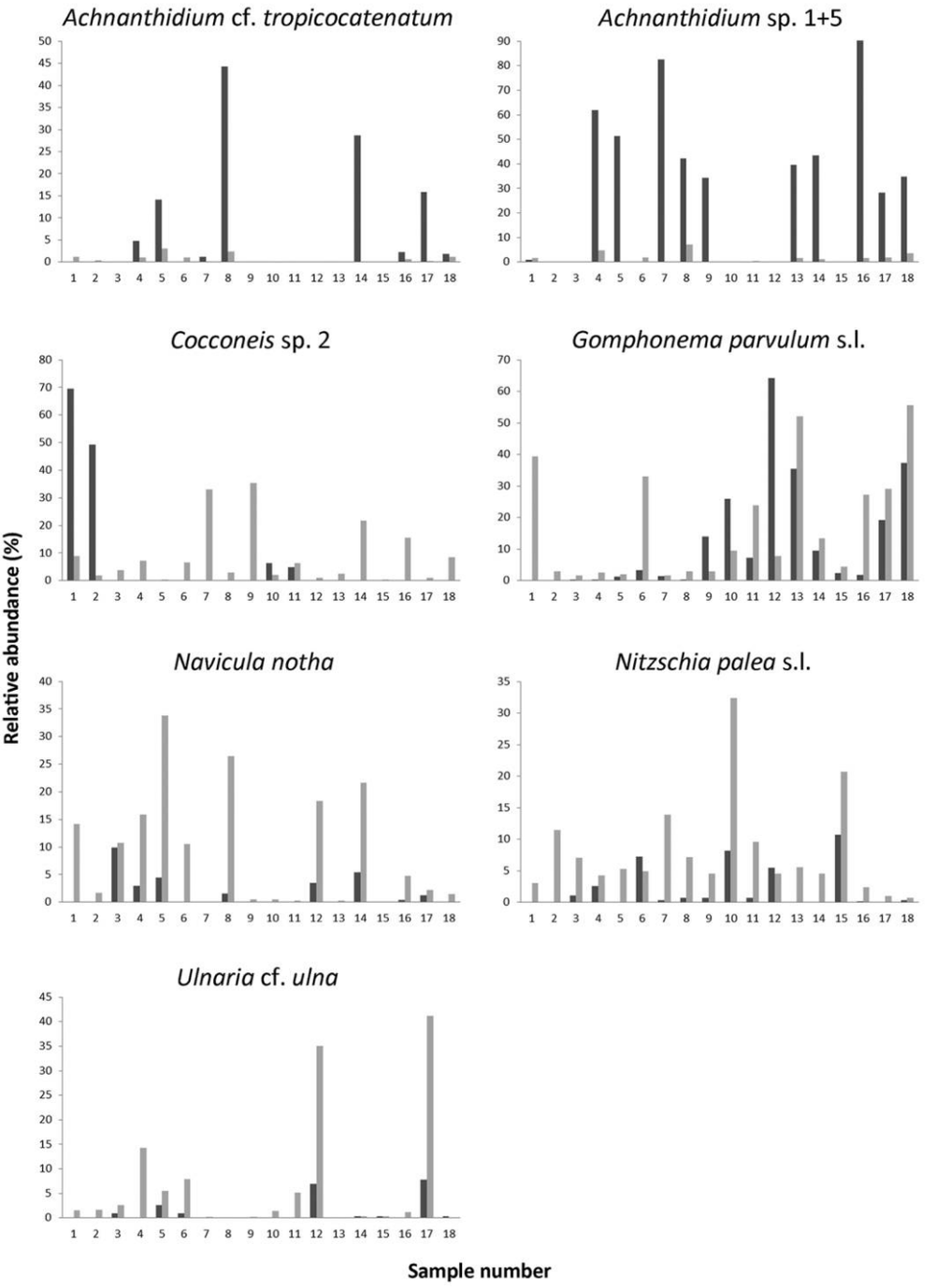

Figure 3. Relative abundances of the 7 most abundant taxa obtained from morphology (dark gray) and metabarcoding (light gray) in each of the 18 samples.

(Fig. S3) in a well-supported clade (bootstrap value = 84%) in the unrooted ML phylogenetic tree.

DISCUSSION

Barcode reference library of the Lerma-Chapala River Basin

The regional vouchered barcode reference library presented here is the 1st of its kind for stream diatoms from Mexico. Only 13 strains of epicontinental diatoms from Mexico had previously been cultured, sequenced, and published (Zimmermann et al. 2011, Abarca et al. 2014, Jahn et al. 2017b). There are other entries in the INSDC databases that refer to diatoms from epicontinental locations in Mexico, but all of them correspond to uncultured organisms with no vouchered material. These facts highlight the importance of having vouchered material to allow traceabil-

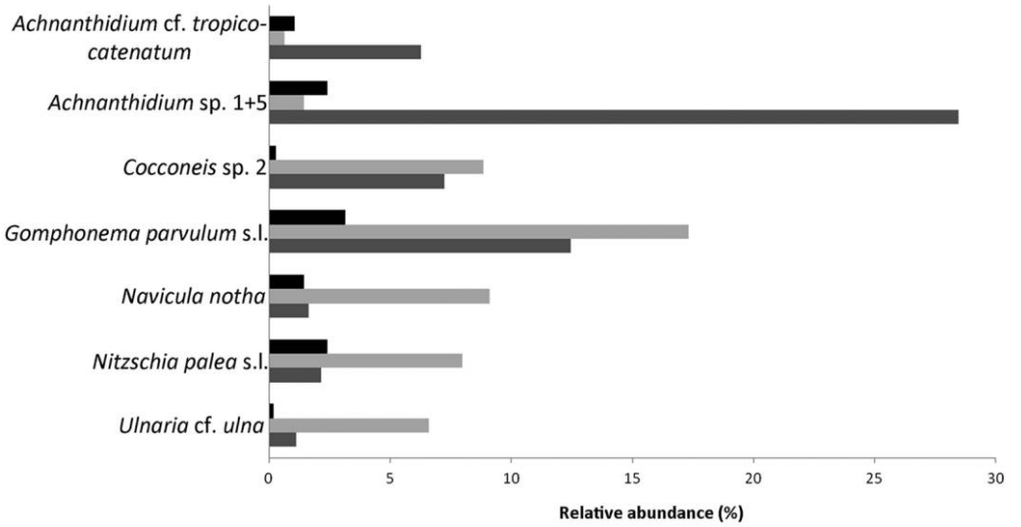

Figure 4. Cumulative relative abundances of the 7 most abundant taxa obtained from morphology (dark gray), metabarcoding (light gray), and metabarcoding after application of correction factors (black).

ity of the data and ensure the availability of reserve material for both morphological and molecular studies for other diatom studies (e.g., phylogenetic and biogeographic).

From the 190 strains established here, 87 strains corresponded to already described species, 49 strains were closely related (cf. *confer*) to already described species, 7 were similar (aff. *affinis*) to other species, and 47 strains were only named at the genus level even after a thorough morphological and bibliographical examination. Some of these unnamed, similar, or closely related strains should probably be described as new species. Our finding of such a large fraction (~50%) of unidentified taxa is not surprising because our samples were collected in Central Mexico, a region not extensively studied for diatoms and for which no monographs have been published (Mora Hernández 2018). A similar proportion of unidentified taxa was found for a small sample (n = 26 strains) of polar diatoms (58%) (Stachura-Suchoples et al. 2016). Even in thoroughly investigated regions like Berlin, 10% of the diatom species identified by Zimmermann et al. (2014) were newly described.

Diatom composition detected by microscopy and metabarcoding

The 500-valve counts based on LM from Mora et al. (2017) led to the identification of 148 taxa. Fifty-six further taxa were found by these authors and us after additional scanning slides under LM and during SEM examinations, raising the richness to 204 infrageneric taxa in 42 genera. This increase in species with increasing sampling effort (in this case, increasing the number of slides observed under microscopy) is a well-known property (Von Falkenhayn 2008, Gotelli and Colwell 2011) of samples such as ours that are characterized by high diversity and a high proportion of rare taxa.

We found a higher taxonomic richness after the PCMA of our metabarcoding data than we did from morphological analyses: 266 from metabarcoding versus 216 taxa identified morphologically (including 12 taxa only identified from cultures). When we took only the unambiguously assigned infrageneric taxa into account (94), ⅔ corresponded to taxa found by microscopy in this study. The assignment of ¼ of those taxa (23) was only possible with our barcode reference sequences. This number represents a fraction of the diatom diversity of the region, but the

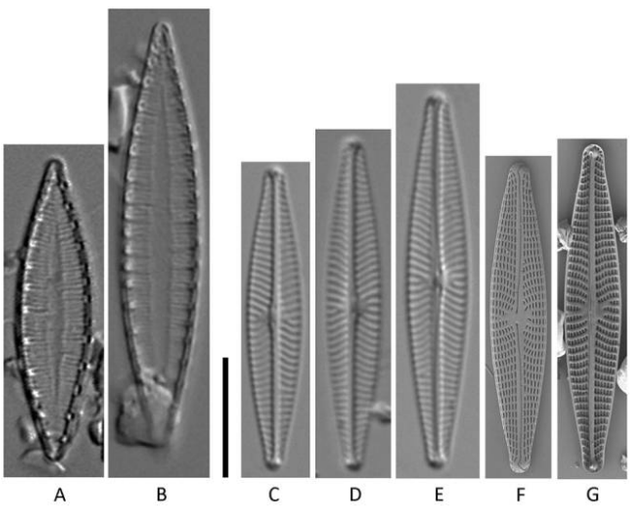

Figure 5. Taxa for which barcodes were retrieved from HTS data. A and B = *Iconella delicatissima* from Laguna de Servín 2 (sample 18) observed under light microscopy. C–G = *Navicula notha*. C–E were observed under light microscopy and F and G were observed with scanning electron microscopy. All specimens were collected from La Laborcilla 1 (sample 5). The black scale bar = 10 μm.

barcode reference library we present here represents a milestone in documenting diatom taxonomic composition in streams of Mexico. In our study alone, the reference library we created allowed us to confidently assign 23 distinct taxa that would have otherwise been left unassigned at the infrageneric level.

Metabarcoding could lead to overestimation of diversity because it normally retrieves more species or MOTUs than morphologically identified taxa (see below; Zimmermann et al. 2015, Vasselon et al. 2017b). However, metabarcoding is a good way to screen for biodiversity because it can show gaps in the described groups of organisms and in barcode reference libraries. Thus, metabarcoding can lead to the refinement of diversity assessments. The main reason cited in the literature for incongruence of taxa lists obtained from morphology and metabarcoding analyses is the incompleteness and lack of accuracy of reference databases (e.g., Zimmermann et al. 2014, 2015, Apothéloz-Perret-Gentil et al. 2017, Vasselon et al. 2017b), which hinders correct taxonomic assignment of environmental sequences. Missing taxa in reference databases would not normally be identified in environmental sequences, whereas sequences with the wrong taxonomy in databases will generate inaccurate identifications (Kermarrec et al. 2014, Zimmermann et al. 2014, Lejzerowicz et al. 2015).

The proportion of sequence reads of diatoms in this study (3.8%) was low compared with the total number of high-quality reads we obtained (43,703 out of 1,156,360). This could be the result of an actual low abundance of diatoms in the sampled sites or a low proportion of live diatom cells. Alternatively, the relatively low proportion of diatom sequences could be because the primers we used in this study had a higher affinity for other protists during PCR, such as Chrysophytes, which comprised 70% of the total reads. This result is unsurprising because Chrysophytes are an abundant and diverse group in microbial freshwater food webs (del Campo and Massana 2011, Grossmann et al. 2016).

Concealed diversity revealed by clonal culturing

We found 12 infrageneric taxa in 7 genera in the cultures that we did not find by microscopy in our environmental samples. For example, we successfully cultivated clones from the genus *Simonsenia*, but we did not detect this genus in the environmental samples analyzed by LM and SEM. These results give rise to the question: how much of the diversity in a sample remains concealed even after exhaustive microscopy examination? In our study, 6% of the diversity (12 out of 216 taxa) was only detected through laboratory culturing. This result could have occurred if the culture media and culturing conditions (i.e., light, day/night cycle and temperature) allowed taxa to grow that were otherwise too rare to be detected through even thorough microscopy examinations. In this case, the suitable culturing conditions would have enabled these taxa to reach abundances high enough to be observed and picked up during the cell isolations. Previous studies of cyanobacteria (Ferris et al. 1996) and marine bacteria (Connon and Giovannoni 2002) have reported that culture media at very low concentrations (e.g., 3 orders of magnitude lower than commonly used) can lead to the identification of taxa undetected by morphological examinations and by standard culturing techniques. Here, some strains were cultured in media at concentrations ½ and ¼ of the manufacturer recommendations, in addition to the recommended concentration. These lower concentrations could partially explain why this concealed diversity was uncovered.

Richness overestimation

The taxa list retrieved by metabarcoding (266 genus-assigned plus 65 unassigned to genus) was larger than the 216 morphology-based list even after a thorough morphological examination of samples. Generally, the number of species or MOTUs that are generated by DNA metabarcoding deviates from the number of taxa observed morphologically (Cowart et al. 2015, Groendahl et al. 2017), normally recovering more taxa than morphology-based approaches. Several biological, environmental, and technical factors contribute to this.

The most important biological factor that could cause DNA-based approaches to overestimate richness is the natural intraspecific and intragenomic variability of the barcoding marker. This variability is particularly problematic when a single traditionally recognized species or bioindicator taxon has multiple genotypes at the barcoding region (Brown et al. 2015, Bálint et al. 2016, Pawlowski et al. 2018). When a single taxon has multiple genotypes at the barcoding region, members of that taxon may cluster into different MOTUs, artificially inflating taxonomic richness. High intraspecific genetic variation is common in nearly all bioindicator groups, such as aquatic insects (Alp et al. 2012, Elbrecht et al. 2014) and diatoms (Rynearson and Armbrust 2000, Trobajo et al. 2009). Moreover, some taxa show high intragenomic polymorphism, such as nematodes (Bik et al. 2013), foraminifera (Weber and Pawlowski 2014), and prokaryotes (Sun et al. 2013). Intragenomic polymorphism has not been widely assessed in diatoms (Alverson and Kolnick 2005) but it could also contribute to MOTU inflation.

MOTU richness can also be artificially inflated through technical errors at different steps of sample processing, especially during amplification (Fonseca et al. 2012, Kermarrec et al. 2013b, Bálint et al. 2016, Elbrecht et al. 2017b) and sequencing (Meacham et al. 2011, Schirmer et al. 2015). In contrast to amplification and sequencing, DNA extraction methods do not affect MOTUs richness significantly in diatoms (Vasselon et al. 2017a).

The MOTU delimitation approach is another factor that influences richness estimation and interpretation. How-

ever, MOTUs do not necessarily correspond to species, and can fail to identify meaningful ecological or phylogenetic units in a straightforward manner (Ryberg 2015, Bálint et al. 2016). To ameliorate this issue, MOTUs have been further analyzed by phylogenetic-based approaches that assign MOTUs to specific taxa (Monaghan et al. 2009, Zimmermann et al. 2015).

Morphology-based assessments can also lead to richness overestimation because diatoms are oxidized before microscopy slides are prepared. Thus, taxonomic identifications, counts to calculate abundance, and diversity estimations rely on the valves of dead cells that could be transported from locations other than the target assemblage (Sawai 2001, Potapova and Charles 2005). The proportion of live diatoms found in lotic environments varies greatly, ranging from 2 to 98% (Gillett et al. 2011). Despite this large variation, the oxidation method for slide preparation gives taxonomic confidence because it allows the visualization of the fine structure of the cell walls, which is needed for identification (Gillet et al. 2009).

Differences in abundance data

Discrepancies in abundance estimates based on valve counts and sequence reads are widely debated by the metabarcoding community, particularly in terms of the use of the number of sequence reads to infer taxa abundances and its application to biomonitoring (Elbrecht and Leese 2015, Vasselon et al. 2018). The barcoding marker and its ability to discriminate among closely related species, as well as primer specificity, are very important and can hinder the otherwise straightforward use of sequence reads to determine species abundances (Elbrecht et al. 2017a). These issues were evident in our results, especially for *Gomphonema* and *Nitzschia*, because the 18S V4 barcode region does not have the discriminatory power to differentiate among some closely related species within these genera. The existence of semi-cryptic diversity within *G. parvulum* has been documented (Kermarrec et al. 2013a), and only a multi-marker phylogeny coupled with detailed micromorphology examinations are able to disentangle some of the species within this species complex that was once thought to have a cosmopolitan distribution (Abarca et al. 2014). For *N. palea*, both morphology and metabarcoding found 3 taxa, but we treated this species as a complex because it was not possible to dissociate the nominate variety from the varieties *N. palea* var. *debilis* and *N. palea* var. *tenuirostris* we obtained from the morphological analysis from the 3 taxa we found through metabarcoding. Morphological, genetic, and mating studies of *N. palea* concluded that this taxon is comprised of 3 or more species but molecular and mating experiments do not separate these taxa into the varieties traditionally recognized morphologically (Trobajo et al. 2009, 2010).

The disparities observed in abundances with each method (Fig. 3) for *N. palea* and *G. parvulum* might have consequences for their use as bioindicators because it was impossible to match taxa assigned by metabarcoding with species recognized morphologically. This limitation might be overcome with taxonomy-free approaches to bioindication that avoid any taxonomic assignments to morphologically recognized taxa (Apothéloz-Perret-Gentil et al. 2017, Tapolczai et al. 2019); this is done by a direct calibration and calculation of ecological values of the MOTUs, without prior species assignment.

Cell size may be another factor in the abundance disparity we observed, as shown in our results for *Achnanthidium* and *Ulnaria* – taxa that differ markedly in cell size. On one hand, the *Achnanthidium* species compared here have a size of 7 to 23 μm and represented 28.5% of the total diatom abundance in the morphological analyses, but comprised only 1.4% of the total sequence read counts. On the other hand, the cells of *Ulnaria* cf. *ulna* analyzed here ranged in size from 70 to 200 μm and represented only 1.1% of the total valve abundance, but consisted of 6.6% of the total read counts. There may be an association between cell size, biovolume, and gene copies of the SSU rDNA (Zhu et al. 2005, Godhe et al. 2008) that could partially explain the disparities in abundances between these methods. Thus, the contrast in cell size, and therefore in biovolume, could explain why *Achnanthidium* spp. were underrepresented in the read counts relative to valve abundance, compared with the relatively large *Ulnaria* cf. *ulna*, which was overrepresented in read abundances relative to valve counts.

Applying CFs based on the cell biovolume of diatoms can reduce the disparity between morphological and molecular abundance data by as much as 45% in both mock communities and environmental samples (Vasselon et al. 2018). However, when we applied CFs for the 18S gene (V. Vasselon et al. unpublished) we saw no overall improvement in the disparity among the abundances recorded by morphology and metabarcoding. One of the reasons for this lack of overall improvement may be the 7 CFs we applied, because only 2 corresponded to the actual species for which they were proposed. The other 5 CFs we used are from taxa that belong to the same genus and are similar in size. Our results point out that species-specific CFs may be needed. The technical errors generated during amplification and sequencing previously described may also affect the disparities observed from both methods.

HTS data as a source of barcodes

Several challenges and limitations make it difficult to establish complete barcode libraries for diatoms, which currently rely mostly on clonal cultures. A primary challenge is the time-consuming process of single cell isolation and culture maintenance. Furthermore, culturing can often be unsuccessful because of recalcitrance of species to culturing

conditions (Mann and Chepurnov 2004, Rimet et al. 2018). HTS data is an alternative source of barcodes (Rimet et al. 2018). As we demonstrate here with *I. delicatissima* and *N. notha*, HTS data can be used as a source of barcodes if data are analyzed carefully.

Retrieving *I. delicatissima* was relatively straightforward because it was the only representative of the genus *Iconella* we found. In this sample, another representative of the Surirellales, *S. angusta*, was observed which might have hindered our findings, but we had reference sequences for *S. angusta* obtained via culturing and Sanger sequencing, so we were able to easily distinguish these species.

Retrieving *N. notha* was more difficult because of the high abundance of *Navicula* taxa in our samples, as evidenced not only by sequences derived from HTS but also from microscopy observations. Thus, we obtained sequences of this taxon from 3 samples in which *N. notha* was the only representative of the genus *Navicula* as detected by microscopy observations.

Conclusion

Our study demonstrates that the combination of morphological (LM and SEM) and molecular (metabarcoding via HTS) methods applied to environmental samples, in combination with a regional barcode reference library based on clonal culturing, increases the detection and identification of diatom species. Thus, our work highlights the complementary aspects of classical taxonomy and DNA metabarcoding (i.e., the importance of their reciprocal illumination). Even with major advances in the development and standardization of molecular tools for diversity assessments and monitoring (Cordier et al. 2018, 2019), the role of morphology in species detection and identification remains central in the ecogenomic era.

ACKNOWLEDGEMENTS

Author contributions: DM, NE, NA, RJ, and JZ conceived the study. DM and JC conducted the field sampling. DM, JC, and OS conducted laboratory work. DM and NA identified the diatoms by microscopy. SP, JHG, and JZ performed the bioinformatic analyses. DM and JZ performed the phylogenetic analyses. DM, SP, JHG, and JZ analyzed the data. DM and JZ drafted the manuscript. All authors contributed to and approved the final version of the manuscript.

The work of DM was funded through doctoral grants from the following Mexican agencies: CONACYT (CVU 367289), CONCYTEQ, and DGRI-SEP. Samples were taken under permit (if required) CONAPESCA PPF/DGOPA-149/14. We thank the Deutsche Forschungsgemeinschaft for Grant INST 130/839-1 FUGG concerning SEM funding, and the Federal Ministry of Education and Research (German Barcode of Life 2 Diatoms [GBOL2], grant number 01LI1501E). We gratefully acknowledge Kirsten Richter, Susan Mbedi, Maja Grubišić, and Verena Deutschmeyer for their help with library preparation for High-Throughput Sequencing. We thank Jana Bansemer for helping with diatom culturing and retrieval of molecular data. Monika Lüchow, Kim Govers, and Juliane Bettig kindly assisted at the SEM. We thank Michael Grünstäudl for his help with the submission of reference sequences to the ENA, and Valentin Vasselon for sharing his calculated CFs for the 18S gene. We thank Verónica Aguilar Zamora for creating the map. We thank Lester Yuan, Christopher Robinson, Charles Hawkins, Katherine Sirianni, and 2 anonymous reviewers whose comments improved our manuscript.

LITERATURE CITED

Abarca, N., R. Jahn, J. Zimmermann, and N. Enke. 2014. Does the cosmopolitan diatom *Gomphonema parvulum* (Kützing) Kützing have a biogeography? PLoS ONE 9:e86885.

Alp, M., I. Keller, A. M. Westram, and C. T. Robinson. 2012. How river structure and biological traits influence gene flow: a population genetic study of two stream invertebrates with differing dispersal abilities. Freshwater Biology 57:969–981.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403–410.

Alverson, A. J., and L. Kolnick. 2005. Intragenomic nucleotide polymorphism among small subunit (18s) rDNA paralogs in the diatom genus *Skeletonema* (Bacillariophyta). Journal of Phycology 41:1248–1257.

Aparicio, J. 2001. Hydrology of the Lerma-Chapala watershed. Pages 3–30 *in* A.M. Hansen and M. van Afferden (editors). The Lerma-Chapala Watershed. Springer, Boston, Massachusetts.

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

Bálint, M., M. Bahram, A. M. Eren, K. Faust, J. A. Fuhrman, B. Lindahl, R. B. O'Hara, M. Öpik, M. L. Sogin, M. Unterseher, and L. Tedersoo. 2016. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiology Reviews 40:686–700.

Bik, H. M., D. Fournier, W. Sung, R. D. Bergeron, and W. K. Thomas. 2013. Intra-genomic variation in the ribosomal repeats of nematodes. PLoS ONE 8:e78230.

Boyer, F., C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac. 2016. OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. Molecular Ecology Resources 16:176–182.

Brown, E. A., F. J. J. Chain, T. J. Crease, H. J. MacIsaac, and M. E. Cristescu. 2015. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecology and Evolution 5:2234–2251.

Chen, X. G., J. Zhang, Y. Huang, and Y. P. Hou. 2013. Diatom taxa identification based on single-cell isolation and rDNA sequencing. Forensic Science International: Genetics Supplement Series 4:e308–e309.

Connon, S. A., and S. J. Giovannoni. 2002. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Applied and Environmental Microbiology 68:3878–3885.

Cordier, T., D. Forster, Y. Dufresne, C. I. M. Martins, T. Stoeck, and J. Pawlowski. 2018. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Molecular Ecology Resources 18:1381–1391.

Cordier, T., A. Lanzén, L. Apothéloz-Perret-Gentil, T. Stoeck, and J. Pawlowski. 2019. Embracing environmental genomics

and machine learning for routine biomonitoring. Trends in Microbiology 27:387–397

Cotler, H., M. Mazari, and J. De Anda (editors). 2006. Atlas de la cuenca Lerma-Chapala: construyendo una visión conjunta. Instituto Nacional de Ecología, Mexico City, Mexico.

Cowart, D. A., M. Pinheiro, O. Mouchel, M. Maguer, J. Grall, J. Miné, and S. Arnaud-Haond. 2015. Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. PLoS ONE 10:e0117562.

del Campo, J., and R. Massana. 2011. Emerging diversity within chrysophytes, choanoflagellates and bicosoecids based on molecular surveys. Protist 162:435–448.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797.

Elbrecht, V., C. K. Feld, M. Gies, D. Hering, M. Sondermann, R. Tollrian, and F. Leese. 2014. Genetic diversity and dispersal potential of the stonefly *Dinocras cephalotes* in a central European low mountain range. Freshwater Science 33:181–192.

Elbrecht, V., and F. Leese. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. PLoS ONE 10:e0130324.

Elbrecht, V., B. Peinert, and F. Leese. 2017a. Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. Ecology and Evolution 7:6918–6926.

Elbrecht, V., E. E. Vamos, K. Meissner, J. Aroviita, and F. Leese. 2017b. Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods in Ecology and Evolution 8:1265–1275.

Ferris, M. J., A. L. Ruff-Roberts, E. D. Kopczynski, M. M. Bateson, and D. M. Ward. 1996. Enrichment culture and microscopy conceal diverse thermophilic *Synechococcus* populations in a single hot spring microbial mat habitat. Applied and Environmental Microbiology 62:1045–1050.

Fonseca, V. G., B. Nichols, D. Lallias, C. Quince, G. R. Carvalho, D. M. Power, and S. Creer. 2012. Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. Nucleic Acids Research 40:e66.

Gillett, N. D., Y. Pan, K. M. Manoylov, and R. J. Stevenson. 2011. The role of live diatoms in bioassessment: a large-scale study of Western US streams. Hydrobiologia 665:79–92.

Gillett, N., Y. Pan, and C. Parker. 2009. Should only live diatoms be used in the bioassessment of small mountain streams? Hydrobiologia 620:135–147.

Godhe, A., M. E. Asplund, K. Härnström, V. Saravanan, A. Tyagi, and I. Karunasagar. 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. Applied and Environmental Microbiology 74:7174–7182.

Gotelli, N. J., and R. K. Colwell. 2011. Estimating species richness. Pages 39–54 *in* A. E. Magurran and B. J. McGill (editors). Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford, UK.

Groendahl, S., M. Kahlert, and P. Fink. 2017. The best of both worlds: a combined approach for analyzing microalgal diversity via metabarcoding and morphology-based methods. PLoS ONE 12:e0172808.

Grossmann, L., C. Bock, M. Schweikert, and J. Boenigk. 2016. Small but manifold – hidden diversity in "*Spumella*-like flagellates". Journal of Eukaryotic Microbiology 63:419–439.

Gruenstaeudl, M., and Y. Hartmaring. 2019. EMBL2checklists: a Python package to facilitate the user-friendly submission of plant and fungal DNA barcoding sequences to ENA. PloS ONE 14:e0210347.

Hamilton, P. B., K. E. Lefebvre, and R. D. Bull. 2015. Single cell PCR amplification of diatoms using fresh and preserved samples. Frontiers in Microbiology 6:1084.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London Series B: Biological Sciences 270:313–321.

Hering, D., R. K. Johnson, S. Kramm, S. Schmutz, K. Szoszkiewicz, and P. F. M. Verdonschot. 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshwater Biology 51:1757–1785.

Ivanova, N. V., T. S. Zemlak, R. H. Hanner, and P. D. N. Hebert. 2007. Universal primer cocktails for fish DNA barcoding. Molecular Ecology Notes 7:544–548.

Jahn, R., N. Abarca, B. Gemeinholzer, D. Mora, O. Skibbe, M. Kulikovskiy, E. Gusev, W. H. Kusber, and J. Zimmermann. 2017b. *Planothidium lanceolatum* and *Planothidium frequentissimum* reinvestigated with molecular methods and morphology: four new species and the taxonomic importance of the sinus and cavum. Diatom Research 32:75–107.

Jahn, R., W. H. Kusber, and C. Cocquyt. 2017a. Differentiating *Iconella* from *Surirella* (Bacillariophyceae): typifying four Ehrenberg names and a preliminary checklist of the African taxa. PhytoKeys 82:73–112.

Kanz, C., P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. 2005. The EMBL nucleotide sequence database. Nucleic Acids Research 33:D29–D33.

Kelly, M., C. Bennett, M. Coste, C. Delgado, F. Delmas, L. Denys, L. Ector, C. Fauville, M. Ferréol, M. Golub, A. Jarlman, M. Kahlert, J. Lucey, B. Ní Chatháin, I. Pardo, P. Pfister, J. Picinska-Faltynowicz, J. Rosebery, C. Schranz, J. Schaumburg, H. van Dam, and S. Vilbaste. 2009. A comparison of national approaches to setting ecological status boundaries in phytobenthos assessment for the European Water Framework Directive: results of an intercalibration exercise. Hydrobiologia 621: 169–182.

Kelly, M. G. 1998. Use of community-based indices to monitor eutrophication in European rivers. Environmental Conservation 25:22–29.

Kermarrec, L., A. Bouchez, F. Rimet, and J. F. Humbert. 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., A. Franc, F. Rimet, P. Chaumeil, J. M. Frigerio, J. F. Humbert, and A. Bouchez. 2014. A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshwater Science 33:349–363.

Kermarrec, L., A. Franc, F. Rimet, P. Chaumeil, J. F. Humbert, and A. Bouchez. 2013b. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Molecular Ecology Resources 13:607–619.

Lang, I., and I. Kaczmarska. 2011. A protocol for a single-cell PCR of diatoms from fixed samples: method validation using *Ditylum brightwellii* (T. West) Grunow. Diatom Research 26:43–49.

Lejzerowicz, F., P. Esling, L. Pillet, T. A. Wilding, K. D. Black, and J. Pawlowski. 2015. High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. Scientific Reports 5:13932.

Mann, D. G., and V. A. Chepurnov. 2004. What have the Romans ever done for us? The past and future contribution of culture studies to diatom systematics. Nova Hedwigia 79:237–291.

Meacham, F., D. Boffelli, J. Dhahbi, D. I. Martin, M. Singer, and L. Pachter. 2011. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics 12:451.

Miller, M. A., W. Pfeiffer, and T. Schwartz. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Pages 1–8 *in* Proceedings of the Gateway Computing Environments Workshop, New Orleans, Louisiana.

Monaghan, M. T., R. Wild, M. Elliot, T. Fujisawa, M. Balke, D. J. G. Inward, D. C. Lees, R. Ranaivosolo, P. Eggleton, T. G. Barraclough, and A. P. Vogler. 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Systematic Biology 58:298–311.

Mora, D., J. Carmona, R. Jahn, J. Zimmermann, and N. Abarca. 2017. Epilithic diatom communities of selected streams from the Lerma-Chapala Basin, Central Mexico, with the description of two new species. PhytoKeys 88:39–69.

Mora Hernández, L. D. 2018. An integrative approach to epilithic diatom diversity analysis in tropical streams from the Lerma-Chapala Basin, Central Mexico (Doctoral dissertation). Freie Universität Berlin, Germany.

Morales, E. A., P. A. Siver, and F. R. Trainor. 2001. Identification of diatoms (Bacillariophyceae) during ecological assessments: comparison between light microscopy and scanning electron microscopy techniques. Proceedings of the Academy of Natural Sciences of Philadelphia 151:95–103.

Moritz, C., and C. Cicero. 2004. DNA barcoding: promise and pitfalls. PLoS Biology 2:e354.

Müller, J., K. Müller, C. Neinhuis, and D. Quandt. 2005. PhyDE-phylogenetic data editor. Version 0.9971. (Available from: http://www.phyde.de)

Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent. 2000. Biodiversity hotspots for conservation priorities. Nature 403:853–858.

Pawlowski, J., S. Audic, S. Adl, D. Bass, L. Belbahri, C. Berney, S. S. Bowser, I. Cepicka, J. Decelle, M. Dunthorn, A. M. Fiore-Donno, G. H. Gile, M. Holzmann, R. Jahn, M. Jirků, P. J. Keeling, M. Kostka, A. Kudryavtsev, E. Lara, J. Lukeš, D. G. Mann, E. A. D. Mitchell, F. Nitsche, M. Romeralo, G. W. Saunders, A. G. B. Simpson, A. V. Smirnov, J. L. Spouge, R. F. Stern, T. Stoeck, J. Zimmermann, D. Schindel, and C. de Vargas. 2012. CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. PLOS Biology 10:e1001419.

Pawlowski, J., M. Kelly-Quinn, F. Altermatt, L. Apothéloz-Perret-Gentil, P. Beja, A. Boggero, A. Borja, A. Bouchez, T. Cordier, I. Domaizon, M. J. Feio, A. F. Filipe, R. Fornaroli, W. Graf, J. Herder, B. van der Hoorn, J. Iwan Jones, M. Sagova-Mareckova, C. Moritz, J. Barquín, J. J. Piggott, M. Pinna, F. Rimet, B. Rinkevich, C. Sousa-Santos, V. Specchia, R. Trobajo, V. Vasselon, S. Vitecek, J. Zimmerman, A. Weigand, F. Leese, and M. Kahlert. 2018. The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Science of The Total Environment 637–638:1295–1310.

Potapova, M., and D. F. Charles. 2005. Choice of substrate in algae-based water-quality assessment. Journal of the North American Benthological Society 24:415–427.

Potapova, M., and D. F. Charles. 2007. Diatom metrics for monitoring eutrophication in rivers of the United States. Ecological Indicators 7:48–70.

Poulíčková, A., J. Špačková, M. G. Kelly, M. Duchoslav, and D. G. Mann. 2008. Ecological variation within *Sellaphora* species complexes (Bacillariophyceae): specialists or generalists? Hydrobiologia 614:373–386.

Proft, S., J. Grau, C. Caswara, C. Mazzoni, D. Mora, and J. Zimmermann. 2017. MetBaN automated pipeline for metabarcoding data using taxonomical/phylogenetical classification of organisms. (Available from: https://github.com/sproft/MetBaN/tree/v0.1.0)

Prygiel, J. 2002. Management of the diatom monitoring networks in France. Journal of Applied Phycology 14:19–26.

Rambaut, A. 2014. FigTree: tree figure drawing tool. Version 1.4.2. Institute of Evolutionary Biology, University of Edinburgh.

Rimet, F., N. Abarca, A. Bouchez, W. H. Kusber, R. Jahn, M. Kahlert, F. Keck, M. G. Kelly, D. G. Mann, A. Piuz, R. Trobajo, K. Tapolczai, V. Vasselon, and J. Zimmermann. 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 18:37–54.

Round, F. E., R. M. Crawford, and D. G. Mann. 1990. Diatoms: biology and morphology of the genera. Cambridge University Press.

Ruck, E. C., T. Nakov, A. J. Alverson, and E. C. Theriot. 2016a. Nomenclatural transfers associated with the phylogenetic reclassification of the Surirellales and Rhopalodiales. Notulae Algarum 10:1–4.

Ruck, E. C., T. Nakov, A. J. Alverson, and E. C. Theriot. 2016b. Phylogeny, ecology, morphological evolution, and reclassification of the diatom orders Surirellales and Rhopalodiales. Molecular Phylogenetics and Evolution 103:155–171.

Ryberg, M. 2015. Molecular operational taxonomic units as approximations of species in the light of evolutionary models and empirical data from Fungi. Molecular Ecology 24:5770–5777.

Rynearson, T. A., and E. V. Armbrust. 2000. DNA fingerprinting reveals extensive genetic diversity in a field population of the centric diatom *Ditylum brightwellii*. Limnology and Oceanography 45:1329–1340.

Sawai, Y. 2001. Distribution of living and dead diatoms in tidal wetlands of northern Japan: relations to taphonomy. Palaeogeography, Palaeoclimatology, Palaeoecology 173:125–141.

Schirmer, M., U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan, and C. Quince. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research 43:e37.

Skibbe, O., J. Zimmermann, W.-H. Kusber, N. Abarca, K. Buczkó and R. Jahn. 2018. *Gomphoneis tegelensis* sp. nov. (Bacillariophyceae): a morphological and molecular investigation based on selected single cells. Diatom Research 33:251–262.

Stachura-Suchoples, K., N. Enke, C. Schlie, I. Schaub, U. Karsten, and R. Jahn. 2016. Contribution towards a morphological and molecular taxonomic reference library of benthic marine diatoms from two Arctic fjords on Svalbard (Norway). Polar Biology 39:1933–1956.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

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

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Systematic Biology 57:758–771.

Sun, D.-L., X. Jiang, Q. L. Wu, and N.-Y. Zhou. 2013. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Applied and Environmental Microbiology 79:5962–5969.

Taberlet, P., E. Coissac, M. Hajibabaei, and L. H. Rieseberg. 2012b. Environmental DNA. Molecular Ecology 21:1789–1793.

Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev. 2012a. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21:2045–2050.

Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution 30:2725–2729.

Tapolczai, K., V. Vasselon, A. Bouchez, C. Stenger-Kovács, J. Padisák, and F. Rimet. 2019. The impact of OTU sequence similarity threshold on diatom-based bioassessment: a case study of the rivers of Mayotte (France, Indian Ocean). Ecology and Evolution 9:166–179.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17:57–86.

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

Trobajo, R., D. G. Mann, E. Clavero, K. M. Evans, P. Vanormelingen, and R. C. McGregor. 2010. The use of partial *cox*1, *rbc*L and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). European Journal of Phycology 45:413–425.

Vasselon, V., A. Bouchez, F. Rimet, S. Jacquet, R. Trobajo, M. Corniquel, K. Tapolczai, and I. Domaizon. 2018. Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring. Methods in Ecology and Evolution 9:1060–1069.

Vasselon, V., I. Domaizon, F. Rimet, M. Kahlert, and A. Bouchez. 2017b. Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: do DNA extraction methods matter? Freshwater Science 36:162–177.

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

Visco, J. A., L. Apothéloz-Perret-Gentil, A. Cordonier, P. Esling, L. Pillet, and J. Pawlowski. 2015. Environmental monitoring: inferring the diatom index from next-generation sequencing data. Environmental Science and Technology 49: 7597–7605.

Von Falkenhayn, L. 2008. An assessment of the use of Bacillariophyceae as biological monitors of heavy metal pollution in Australian tropical streams (Doctoral dissertation). The University of Adelaide, Australia.

Weber, A. A.-T., and J. Pawlowski. 2014. Wide occurrence of SSU rDNA intragenomic polymorphism in foraminifera and its implications for molecular species identification. Protist 165:645–661.

Wester, P., C. A. Scott, and M. Burton. 2005. River basin closure and institutional change in Mexico's Lerma-Chapala Basin. Pages 125–144 *in* M. Svendsen (editor). Irrigation and river basin management: options for governance and institutions. CABI Publishing.

Yu, D. W., Y. Ji, B. C. Emerson, X. Wang, C. Ye, C. Yang, and Z. Ding. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3:613–623.

Zhu, F., R. Massana, F. Not, D. Marie, and D. Vaulot. 2005. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiology Ecology 52:79–92.

Zimmermann, J., N. Abarca, N. Enke, O. Skibbe, W. H. Kusber, and R. Jahn. 2014. Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. PLoS ONE 9:e108793.

Zimmermann, J., G. Glöckner, R. Jahn, N. Enke, and B. Gemeinholzer. 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources 15:526–542.

Zimmermann, J., R. Jahn, and B. Gemeinholzer. 2011. Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Organisms Diversity and Evolution 11:173.