# Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter?

**Valentin Vasselon[1,3], Isabelle Domaizon[1,4], Frédéric Rimet[1,5], Maria Kahlert[2,6], and Agnès Bouchez[1,7]**

[1]CARRTEL, INRA, Université de Savoie Mont Blanc, 74200, Thonon-les-Bains, France
[2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P.O. Box 7050, 75007, Uppsala, Sweden

**Abstract:** Current freshwater biomonitoring with diatoms is based on microscopic examination of the morphology of their silica skeleton. This standardized approach is time consuming and requires a high degree of taxonomic expertise. Metabarcoding combined with high-throughput sequencing (HTS) has great potential for next-generation biomonitoring applications but requires standardization. Molecular inventories are strongly influenced by the DNA extraction method used, but the effect of extraction protocols has not been tested to enable selection of the best DNA extraction method for HTS metabarcoding. We used 5 DNA extraction methods combining various types of cell lysis and DNA purification to extract DNA from 8 pure diatom cultures and 8 samples from streams and lakes with differing water quality. We compared the methods based on: 1) quality and purity of the extracted DNA, 2) community inventories obtained from HTS targeting the ribulose-1, 5-bisphosphate carboxylase (*rbc*L) barcode, and 3) similarity between molecular and microscopy-based inventories of community composition and the Specific Pollution-sensitivity Index [SPI]. A method based on GenElute™-LPA had higher extraction efficiency than the 4 commercial kits but had the highest polymerase chain reaction inhibition level. All 5 methods were efficient for HTS, and method did not affect operational taxonomic unit richness. We observed variations in the relative abundance of some taxa within *Nitzschia*, *Amphora*, *Encyonema*, *Gomphonema*, and *Navicula* between 2 of the 5 methods, but method did not affect global diatom community composition or SPI values. SPI values calculated from microscopy-based inventories and molecular inventories based on all 5 extraction methods were strongly correlated. For convenience purposes (high DNA quantity and low cost), we encourage standardization of HTS diatom biomonitoring based on the SA-Gen method.

**Key words:** next-generation biomonitoring, DNA extraction methods, metabarcoding, high-throughput sequencing, diatom communities

Diatoms are good bioindicators because of their high diversity, short life cycle, high sensitivity to environmental conditions, and widespread distribution in all freshwater ecosystems (Stevenson and Pan 1999). Therefore, diatom communities are used routinely for water-quality assessment in monitoring programs and by environmental agencies in many countries. Well-established guidelines like the Clean Water Act in USA (US CWA) or the Water Framework Directive in Europe (EU WFD) help to standardize methods across countries and laboratories. Classical diatom biomonitoring is based on the composition of environmental communities and relies on morphological identification at the species level with the aid of microscopes and specialized floristic books. Species identification is challenging because of the large diversity of diatoms (Mann and Vanormelingen 2013) and the subtle differences in morphological features of their silica frustule (exoskeleton) used for taxonomy. Quite often, discrepancies in taxonomic inventories occur from one laboratory to another (Kahlert et al. 2012, Werner et al. 2016). Moreover, this approach is time consuming and costly. Increased demand for environmental assessment in recent years implies that the number of samples to be analyzed will increase, a trend that will become untenable if analysis is based on microscopic identification. Thus, fast and cost-effective alternatives must be developed. One promising alternative is application of environmental DNA metabarcoding.

The potential of DNA metabarcoding combined with high-throughput sequencing (HTS) for investigating benthic diatom community structure already has been demonstrated (Kermarrec et al. 2013b, 2014, Zimmermann et al. 2015, Gib-

E-mail addresses: [3]valentin.vasselon@thonon.inra.fr; [4]isabelle.domaizon@thonon.inra.fr; [5]frederic.rimet@thonon.inra.fr; [6]maria.kahlert@slu.se; [7]agnes.bouchez@thonon.inra.fr

DOI: 10.1086/690649. Received 30 August 2016; Accepted 29 October 2016; Published online 9 January 2017.
Freshwater Science. 2017. 36(1):162–177. 

son et al. 2015, Visco et al. 2015), opening the way to "next-generation biomonitoring". However, these pioneer investigators used differing molecular methods and protocols, thereby hampering relevant comparison among studies. Factors ranging from the initial field sampling to the bioinformatics treatment of DNA sequences can affect the final molecular species inventory of diatom communities. These factors include: 1) the DNA marker chosen, which affects species discriminatory power and the availability and completeness of a DNA reference database; 2) the methods used for various steps of molecular analyses (i.e., DNA extraction methods, sequencing technology); and 3) the bioinformatics workflow (i.e., data processing steps, clustering algorithms, and taxonomic assignment methods). HTS metabarcoding is still in its infancy, and guidelines need to be defined at each step to allow its standardization for biomonitoring purposes. Active investigations are under way to find the best DNA marker (Kermarrec et al. 2013b, 2014) or to optimize the HTS data sequence-processing using pipelines (Majaneva et al. 2015, Schmidt et al. 2015), but little attention has been given to the DNA extraction from diatom samples and it requires further study.

Obtaining a molecular inventory relies on extraction of DNA representative of the indigenous diatom community composition. The quality and the quantity of the DNA extracted from environmental samples affect the investigator's ability to obtain a relevant taxonomic list. Several studies have been performed to evaluate the effect of extraction protocols on microbial DNA analysis. The studies have been focused mainly on bacterial communities (Willner et al. 2012, Rubin et al. 2014, Wesolowska-Andersen et al. 2014, Wagner Mackenzie et al. 2015) and freshwater microalgae (Eland et al. 2012) but rarely diatoms (Nguyen et al. 2011). Results of these studies show that the choice of the DNA extraction method, particularly the cell lysis type, affect quality and quantity of extracted DNA and inferences regarding community diversity and structure. However, authors of these studies generally depicted the diversity of the targeted biological groups based on fingerprinting methods (e.g., denaturing gradient gel electrophoresis [DGGE]), which provide only a very coarse view of community diversity.

Our goal was to find the optimal method for DNA extraction when using HTS methods as a step toward standardizing the application of diatom metabarcoding. DNA extraction has 3 main requirements to: 1) obtain good quality DNA and sufficient DNA quantity, 2) obtain inhibitor-free DNA for subsequent molecular biological analyses, and 3) ensure representative lysis of all organisms (in our case, the different diatom species) in the sample. We compared 5 methods of DNA extraction in combination with HTS metabarcoding. These methods combined various types of cell lysis and DNA purification. We tested the 5 methods on 8 pure cultures of diatoms and 8 freshwater samples of benthic diatom communities from streams and lakes with differing water quality and geographical origin. We based our comparison on the following criteria: 1) DNA extraction efficiency (quantity of DNA), DNA quality, and presence of inhibitors in extracted DNA, 2) the diatom community structure as revealed by HTS sequencing of the ribulose-1, 5-bisphosphate carboxylase (*rbc*L) barcode (qualitative and quantitative comparisons were performed at different taxonomic levels), and 3) comparison of molecular and microscopy-based inventories in terms of community composition and inferred water-quality indices.

## METHODS

### Diatom cultures

We selected 8 pure cultures of diatoms from the Thonon Culture Collection (TCC; http://www6.inra.fr/carrtel-collection_eng/) based on their contrasting morphological and phylogenetical features. These strains were cultured in 300 mL sterile DV media, as previously described (Rimet et al. 2014) (Fig. 1A). From each diatom culture, we prepared a 20-mL aliquot containing $10^5$ to $10^6$ cells and froze the aliquot at −80°C until further analysis.

### Environmental community samples

Eight environmental community samples were collected from benthic biofilms at 6 streams and 2 lakes in 3 geographical areas (Sweden, France, and Mayotte, a French Tropical Island) (Fig. 1B). We selected the sampling sites for their contrasting geographic origin, water-quality status (polluted to good quality), and physicochemical characteristics (concentration of organic matter and presence of metals or pesticides). These characteristics were chosen because they can affect DNA extraction from the prevailing diatom assemblages. All environmental samples were collected following the European Water Framework Directive standards (NF EN 13946; AFNOR 2003) by scraping material from the surface of ≥5 submerged stones. The resulting material was transferred to 15-mL Falcon tubes and fixed by immediately adding 99% ethanol to reach a final ethanol concentration of ~70–80%. Ethanol fixation prevents grazing by metazooplankton and allows good preservation of DNA (Motwani and Gorokhova 2013). Fixed environmental samples were stored at room temperature under dark conditions until preparation for morphological analysis and DNA extraction.

We estimated diatom valve concentration in samples based on microscopic counts. Each diatom skeleton is composed of 2 valves. We used the formula:

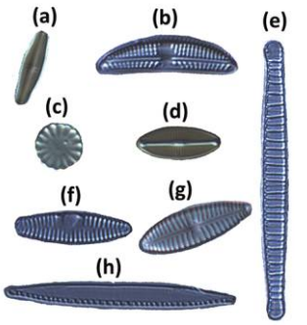
$$N = \text{number of valves counted} \times \frac{R}{M} \quad \text{(Eq. 1)}$$

$$\text{where } R = \frac{\text{cover slip area}\left(\text{mm}^2\right)}{\text{microscopic counting area}\left(\text{mm}^2\right)}, \quad \text{(Eq. 2)}$$

$N$ = number of valves/mg of sample, $R$ = counting ratio, and $M$ = quantity of sample fixed on slide (mg).

A)

| Code | Species | TCC code | Width range (µm) | Length range (µm) | Pictures |
|---|---|---|---|---|---|
| ***AM*** | *Achnanthidium minutissimum* | TCC 667 | 1.5-3.3 | 5.6-20.8 | (a) |
| ***AP*** | *Amphora pediculus* | TCC 702 | 2.5-4 | 6-16 | (b) |
| ***CMEN*** | *Cyclotella menegheniana* | TCC 690 | diameter = 5-43 | | (c) |
| ***CMOL*** | *Craticula molestiformis* | TCC 459 | 3.4-4.9 | 12-15 | (d) |
| ***DT*** | *Diatoma tenuis* | TCC 861 | 2.9-4.9 | 20-85 | (e) |
| ***FP*** | *Fragilaria perminuta* | TCC 753 | 3-4 | 7-40 | (f) |
| ***GP*** | *Gomphonema parvulum* | TCC 492 | 4-8 | 10-46 | (g) |
| ***NP*** | *Nitzschia palea* | TCC 139-1 | 3-4 | 12-42 | (h) |




B)

| Code | Sampling Site | Sampling date | Geographical area | Physico-chemical characteristics | Site quality |
|---|---|---|---|---|---|
| ***Edian*** | Stream Edian | 11-2014 | France | Clear water, thick biofilm | good |
| ***Aire*** | Stream Aire | 11-2014 | France | Copper | polluted |
| ***Lake*** | Lake Geneva | 11-2014 | France | Thick biofilm | good |
| ***767*** | Stream Dammån | 09-2013 | Sweden | Acid pH, humic acid | good |
| ***M36*** | Agricultural stream | 10-2014 | Sweden | Pesticides, nutrients | polluted |
| ***P45*** | Lake Båtkåjaure | 09-2009 | Sweden | Clear water (mountain) | good |
| ***Ref7*** | Stream Dapani | 11-2014 | Mayotte | Clear water, thin biofilm | good |
| ***Pol2*** | Stream Majimbini | 11-2014 | Mayotte | Organic matter, detergent | highly polluted |

Figure 1. A.—Characteristics of the diatom cultures from the Thonon Culture Collection (TCC). B.—Biofilm sampling sites. Pictures transformed from the Rsyst::diatom database (length not to scale).

## DNA extraction

We centrifuged environmental samples and pure culture subsamples at 13,000 rpm for 30 min and removed the supernatant. We used 25 mg of wet pellet as a starter for DNA extraction for each environmental sample. The quantity corresponded to the smallest amount of starting material recommended for the selected DNA extraction methods and is the usual environmental sample amount used for DNA extraction (Fig. 2).

We extracted DNA in triplicate from each diatom culture and each environmental sample with 4 commercial DNA extraction kits (Fig. 2): Macherey–Nagel (Düren, Germany) NucleoSpin® Soil Kit (MN-Soil), Macherey–Nagel NucleoSpin® Plant II Kit (MN-Plant), Stratec (Birkenfeld, Germany) Invisorb® Spin Plant Mini Kit (S-Plant), Qiagen (Hilden, Germany) DNeasy® Blood and Tissue kit (Q-Blood), and 1 non-kit protocol based on Sigma–Aldrich (St Louis, Missouri) GenEluteTM-LPA DNA precipitation (SA-Gen), which was used in previous studies (Kermarrec et al. 2013a, Chonova et al. 2016). These 5 DNA extraction methods have been used or recommended for use to extract DNA from freshwater algae (Nguyen et al. 2011, Eland et al. 2012, Kermarrec et al. 2013a, Zimmermann et al. 2015) and were chosen based on their various types of lysis (mechanical, enzymatic, thermal) and the use or not of columns to remove contaminants/co-extracted molecules (Fig. 2). SA-Gen was the only method that did not include a column purification step. We ran all protocols according to the manufacturer's instructions (Fig. 2) with a single modification for MN-Plant where we changed incubation time at 65°C from 10 to 45 min (manufacturer's recommendation for difficult plant material).

The final elution volume was 40 µL for all DNA extraction methods. We conducted a total of 96 DNA extractions for diatom cultures (MN-Soil method not tested) and 120 for environmental samples (all 5 extraction methods tested).

## Evaluation of DNA extraction efficiency and DNA quality

For all samples, we quantified the extracted DNA with the Life Technologies (Carlsbad, California) Quant-iTTM PicoGreen® dsDNA assay kit using a microplate reader (Fluoroskan AscentTM FL; Thermo Scientific, Waltham, Massachusetts) and following the manufacturer's instructions. To compare DNA extraction efficiency among methods, we normalized DNA concentrations as µg DNA/g wet biofilm for environmental samples and as µg DNA/$10^4$ cells for diatom cultures. We assessed DNA quality by spectrophotometry with 260/280 nm ratio with the NanoDrop®ND-1000 (NanoDrop Technologies, Wilmington, Delaware).

We compared mean values for DNA quantities and qualities based on the Kruskal–Wallis group test followed by the Mann–Whitney pairwise test to evaluate the effect

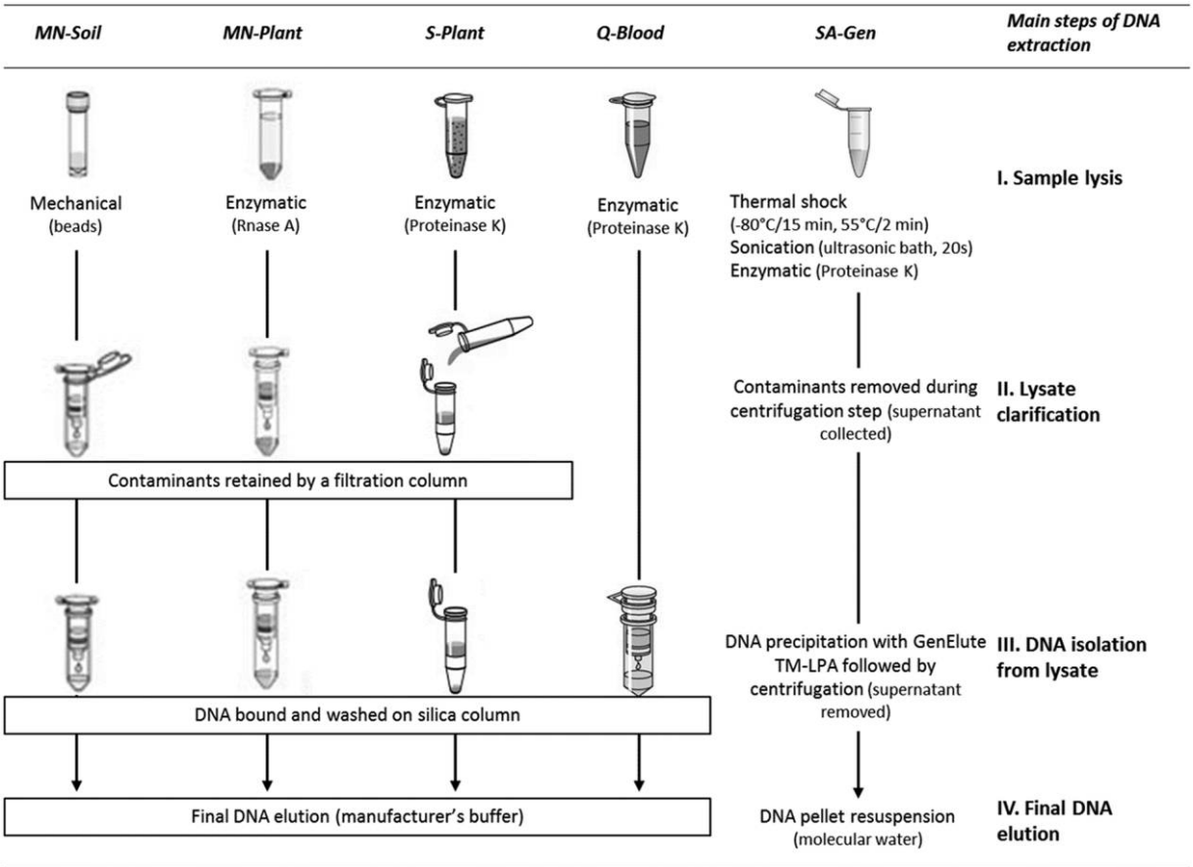



Figure 2. The main steps of DNA extraction for the 5 methods with a focus on sample lysis (I), lysate clarification (II), DNA isolation from lysate (III), and DNA elution (IV). Pictures modified from the manufacturers' web sites.

of the different extraction methods on these parameters. These statistical analyses were performed in R (version 3.0.2; R Project for Statistical Computing, Vienna, Austria).

## Polymerase Chain Reaction (PCR) inhibitor detection (quantitative PCR [qPCR])

We estimated the presence of inhibitors by making serial dilutions of the DNA extracts and estimating *rbc*L copy numbers via qPCR for every dilution (Gallup and Ackermann 2006, Lloyd et al. 2010). In this approach, inhibitors are assumed to be diluted with a log-linear relationship between cycle threshold (Ct) and the dilution factor (DF). Ct values obtained for a 10-fold dilution of the same sample have a theoretical difference of 3.3 cycles when considering 100% PCR efficiency. The presence of PCR inhibitors coextracted with DNA reduces PCR efficiency and affects this expected value of 3.3, allowing detection of these inhibitors. We performed qPCR assay on serial dilutions ($10^0$–$10^{-3}$) of 1 DNA extraction replicate per environmental sample and DNA extraction method (corresponding to 40 environmental DNA extracts). The level of inhibition was estimated by calculating for each dilution level the dilution factor (DF) needed to remove all inhibition effects as $DF = 10^x$, where $x$ = (theoretical Ct − measured Ct)/standard curve slope (transformed from Gibson et al. 2012), measured Ct = Ct obtained during assay for each dilution level, and theoretical Ct = expected Ct value for the dilution without inhibition.

We estimated the theoretical Ct for each assay, and it generally corresponded to the highest dilution ($10^{-3}$). We considered samples with $DF \leq 2$ as not inhibited, values with $2 < DF \leq 10$ as weakly inhibited, $10 < DF \leq 100$ as strongly inhibited, and $DF > 100$ as very strongly inhibited. We conducted qPCR targeting a short region of the *rbc*L plastid gene (312 base pairs [bp]; the same region was used for HTS sequencing) in a Rotor Gene RG-3000 (Corbett Research, Sydney, Australia) with 2 replicates using the QuantiTect SYBR Green PCR Kit (Life Technologies). The mix (25 μL final volume) contained: 12.5 μL of master mix provided by the supplier, 1.25 μL of 10 μM forward primer

Diat_rbcL_708F (AGG TGA AGT TAA AGG TTC ATA CTT DAA) (Stoof-Leichsenring et al. 2012) and reverse primer R3 (CCT TCT AAT TTA CCA ACA ACT G) (Bruder and Medlin 2007), 1.25 μL of 10 g/L bovine serum albumin (BSA), 2 μL of extracted DNA, and 6.75 μL $H_2O$ (molecular biology grade). Reaction conditions were: initial denaturation of DNA at 95°C for 15 min followed by 40 cycles with 45 s denaturation at 95°C, followed by 45 s annealing at 55°C and 45 s extension at 72°C. We used 1 no-template control (NTC) as a negative control.

We standardized qPCR assays by adding serial dilutions (7 points) of standard DNA with known [DNA] and known copy number of the *rbc*L fragment. This reference DNA was prepared with plastid DNA of *Nitzschia palea* following 4 main steps: 1) amplification with Diat_rbcL_708F/R3 primers, 2) insertion of the *rbc*L 312 bp amplicon produced into TOPO plasmid and cloning into *Escherichia coli* bacteria using the TOPO TA cloning kit (Invitrogen, Carlsbad, California), 3) purification and extraction of plasmids with insert from positive clones using the QIAprep Spin Miniprep Kit (Qiagen), 4) evaluation of plasmid DNA concentration using the PicoGreen method (as described above); this concentration was considered as $10^0$ dilution level.

We analyzed the data with Rotor-gene 6 (version 6.1; Corbett Research) with a fluorescence threshold of 0.3 for denoising and determining Ct. The results served both for detection of inhibitions and quantification of *rbc*L genes in environmental samples to provide a quantitative comparison between qPCR estimations and microscopic counts.

## HTS molecular inventories in environmental samples

### *Preparation of the library of amplicons and HTS sequencing*

For all environmental samples, we conducted HTS sequencing of the *rbc*L 312 bp fragment on 2 of the 3 DNA replicates from each extraction method. For each DNA sample, we ran the PCR amplification in triplicate on 1 μL of extracted DNA in a mixture (25 μL final volume) containing: 0.75 U of TaKaRa LA Taq® polymerase (TaKaRa Bio, Sugatsu, Japan), 2.5 μL of 10× buffer, 1.25 μL of 10 μM of primers Diat_rbcL_708F and R3, 1.25 μL of 10 g/L BSA, 2 μL of 2.5 mM deoxynucleotide (dNTP), and completed with 15.6 μL $H_2O$ (molecular biology grade). PCR reaction conditions were the same as those used for qPCR (see above) with 30 cycles. Seventy-eight of the 80 DNA extracts were amplified successfully, and the 2 replicates extracted from Ref 7 sample with SA-Gen method were not amplified.

For each DNA extract, we pooled the 3 replicates of PCR product and then cleaned with Agencourt AMPure beads (Beckman–Coulter, Brea, California) following the manufacturer's instructions except one modification regarding the beads/DNA ratio, which we adjusted to 1.5:1. We assessed purified amplicons for quality and quantified them using the 2200 TapeStation (Agilent Technologies, Santa Clara, California) with D1000 screen tape and reagents. We used the 78 purified amplicons to prepare 78 DNA libraries for HTS with Ion Torrent technology using the NEBNext® Fast DNA Library Prep set for Ion Torrent™ (BioLabs, Ipswich, Massachusetts), following the manufacturer protocol for End repair, PCR amplification of adapter ligated DNA (7 cycles), and cleaning steps. Ligation of library adapters to purified amplicons was done with 2 μL of P1 adapter (NEB kit) and 2 μL of A-X tag adapter provided in Ion Express™ Barcode adapters (Life Technologies) using 1 tag per amplicon.

We checked the quality, size, and concentration of the libraries with the 2200 TapeStation with D1000 High Sensitivity screen tape and reagents. We diluted each library to 100 pM and pooled all of them together in a unique mixture that was sequenced using 1 Ion 318™Chip Kit V2 (Life Technologies) on a PGM Ion Torrent machine by the Plateforme Génome Transcriptome (PGTB, Bordeaux, France).

***Sequence data processing (Fig. S1)*** Demultiplexing and adapter-removal steps were made by the Sequencing Platform, which provided a single fastq file for each of the 78 libraries (fastq files available at: http://doi.org/10.5281/zenodo.166859). DNA reads were filtered for length and quality using Mothur software (Schloss et al. 2009) in every fastq file with the following settings: minimum length = 250 bp, Phred quality score >23 over a moving window of 25 bp, maximum of 1 mismatch in forward primer sequence, homopolymers <8 bp, and absence of ambiguous base. Reads that were not fully aligned with the *rbc*L barcode were removed. The 78 resulting files were analyzed together. Denoising of sequencing error was performed with the Precluster command by creating read clusters, allowing 1 nucleotide difference between DNA reads. Chimera removal was done using the Uchime algorithm (Edgar et al. 2011). The potential effect of the DNA extraction method or the sampling site on read abundances was assessed with 2-way analysis of variance.

We used the Rsyst::diatom database (Rimet et al. 2016, version updated in January 2015, http://www.rsyst.inra.fr/en) restricted to our 312-bp *rbc*L barcode as a reference database. Taxonomic assignment of DNA reads at the 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 Bacillariophyta (diatoms) were used in further analysis.

We conducted a dereplication step and calculated uncorrected pairwise distances between aligned reads (alignment performed using the align.seqs command in Mothur with the algorithm proposed by Needleman and Wunsch 1970 and default setting) to generate a similarity distance matrix. Based on this distance matrix, reads were clustered in operational taxonomic units (OTUs) using the furthest-

neighbor algorithm at a 95% similarity level. This similarity level was reported as a relevant cut-off threshold for OTU delineation that limits artificial inflation of eukaryote OTUs (following recommendations by Mangot et al. 2013). Singletons were removed, and all samples were normalized to the smallest read abundance obtained among the 78 libraries for further analysis (Fig. S1).

Taxonomy was assigned to OTUs on the basis of the consensus taxonomy of reads (application of the classify.otu command from Mothur) (Schloss et al. 2009) with a stringent consensus confidence threshold (>80%) (Fig. S1). OTU α diversity was estimated in Mothur with the Chao1 estimator as a global richness estimator and Shannon index as diversity estimator.

***Statistical analysis on community structure as revealed by HTS*** We used the Kruskal–Wallis test to evaluate the effect of extraction method on Chao and Shannon indices. We compared community compositions of the 78 DNA extracts at the OTU and species levels. The OTU list represents the whole DNA reads that were clustered at 95% similarity level, whereas the species list takes into account only OTUs for which the taxonomic assignment was good enough to provide identification at the species level. We used the OTU or species lists to compute Bray–Curtis dissimilarity indices, which we visualized using nonmetric multidimensional scaling (NMDS). We used permutational ANOVA (PERMANOVA) (PRIMER-E, Plymouth, UK) to compare similarity between DNA extraction methods within and between the 8 environmental samples and similarity percentage (SIMPER) (PRIMER-E) analyses to detect which OTUs were the main contributors to the dissimilarity.

### Comparison between molecular and morphological taxonomic inventories

We based morphological taxonomic inventories of environmental samples on diatom valves according to the European Committee for Standardization (NF EN 14407; AFNOR 2004). We counted a minimum of 400 valves with the aid of a light microscope with 1000× magnification and identified them based on classical European floras for French and Swedish samples (e.g., Krammer and Lange-Bertalot 1986, 1988, 1991a, b; Krammer 2000, 2001, 2002, 2003) and specific literature for Mayotte tropical samples (e.g., Bourrelly and Manguin 1952, Metzeltin and Lange-Bertalot 1998, 2007, Tudesque et al. 2008).

We compared diatom taxonomic inventories obtained by the molecular approach to those obtained by the morphological approach at the species and genus levels. We used OMNIDIA 5 software (Lecointe et al. 1993, library 5.3 2015) to calculate and compare the Specific Pollution-sensitivity Index (SPI) (Cemagref 1982) based on species lists (or genus if species level was not reached) obtained by PGM sequencing or by microscopy for each environmental sample.

We also calculated valve and *rbc*L gene copy numbers per mg of wet biofilm from microscopic count and qPCR assay and calculated the ratio [valve]/[*rbc*L copy].

## RESULTS

### DNA extraction efficiency, quality, and PCR inhibition

DNA extraction efficiency differed significantly among methods for diatom cultures ($p < 0.001$) and environmental samples ($p < 0.001$). The SA-Gen method yielded the highest quantity of DNA for diatom cultures and environmental samples, whereas the lowest DNA quantities were obtained with the S-Plant method for diatom cultures and the MN-soil method for environmental samples (Table 1). All methods yielded good DNA quality (260/280 ratios: 1.7–2.0) with diatom cultures and environmental samples, but environmental samples extracted with MN-Plant method had a slightly lower value (1.5) (Table 1).

Inhibition levels for each environmental sample and DNA extraction method were estimated from qPCR assays (Table S1). DNA samples extracted with the SA-Gen method presented the highest level of PCR inhibition (Table 2), whereas DNA obtained with the MN-Soil method was easily amplified without DNA dilution and was free of inhibition. MN-Plant, S-Plant, and Q-Blood extracts were slightly inhibited, and a 10-fold dilution was sufficient to remove inhibition (with 1 exception for the MN-plant method on the Ref 7 sample).

### Comparison of diatom quantification: microscopy vs qPCR

*rbc*L copy number/mg sample was calculated during qPCR assay for all the environmental samples and DNA

Table 1. Mean (SD, $n = 72$) DNA extraction efficiency and 260/280 DNA ratios obtained for the 5 extraction methods with the pure culture and environmental samples. ND = not determined (out of range), Ø = missing values. Values sharing the same letter are not statistically different.

| | DNA yield | | 260/280 ratio | |
|---|---|---|---|---|
| | Pure culture (μg/$10^4$ cells) | Environmental sample (μg/15 mg biofilm) | Pure culture | Environmental sample |
| MN-Soil | Ø | 1.6 (1.5)$^{ab}$ | Ø | 1.8 (0.5)$^{a}$ |
| MN-Plant | 3.9 (5.7)$^{a}$ | 2.1 (3.1)$^{a}$ | 1.8 (0.2)$^{a}$ | 1.5 (0.4)$^{b}$ |
| S-Plant | 3.8 (4.9)$^{ab}$ | 5.6 (6)$^{c}$ | 2 (0.1)$^{b}$ | 1.8 (0.3)$^{a}$ |
| Q-Blood | 5.9 (7.1)$^{bc}$ | 3.7 (4.6)$^{b}$ | 1.7 (0.1)$^{c}$ | 1.9 (0.4)$^{a}$ |
| SA-Gen | 8.5 (16.8)$^{c}$ | 20.8 (22.5)$^{d}$ | 1.9 (0.1)$^{d}$ | ND |

Table 2. Mean (SD, $n = 72$) estimation of the inhibition level for DNA extracted from the environmental samples. – = not inhibited, + = weakly inhibited, ++ = strongly inhibited, +++ = very strongly inhibited, ND = not determined (out of range) (see Table S1 for corresponding details).

| | MN-Soil | MN-Plant | S-Plant | Q-Blood | SA-Gen |
|---|---|---|---|---|---|
| Edian | – | + | – | – | +++ |
| Aire | – | + | – | + | ++ |
| Lake | – | + | + | + | ++ |
| 767 | – | ND | ND | ND | ND |
| M36 | – | – | – | – | +++ |
| P45 | ND | ND | ND | ND | ND |
| Ref 7 | – | ++ | + | + | +++ |
| Pol2 | – | + | + | – | ++ |

extraction methods and compared to valve number/mg sample (except for 767 and P45 samples, which were out of range for qPCR assay) (Table S2). The ratio between *rbc*L copy and valve numbers showed that the *rbc*L concentration was mostly (26 of 30 cases) below the valve concentration. *rbc*L concentrations obtained with the SA-Gen method provided the best correspondence compared to the valve concentration (Table S2).

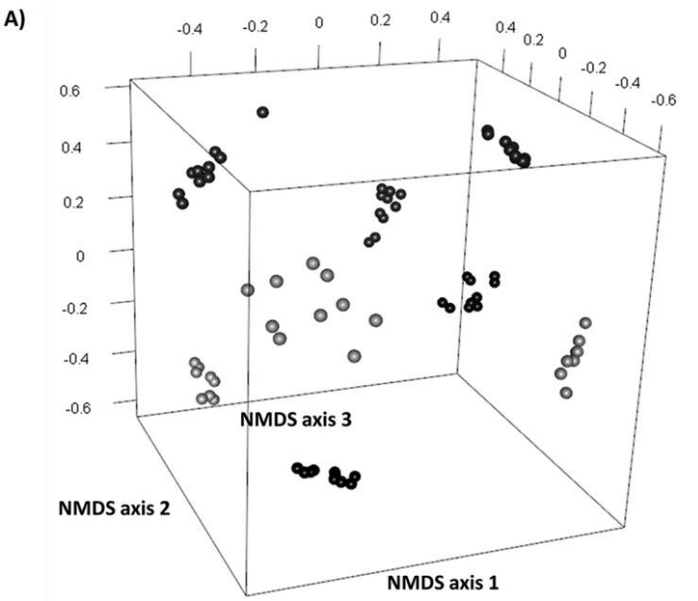
## Effect of extraction methods on richness, composition, and structure of diatom community

After DNA sequencing of the 78 libraries, a total of 4,711,673 of DNA reads was obtained with an average read length of 271 bp. After trimming and removing singletons, 967,089 DNA reads (20.5% of the initial number) were conserved and clustered into 3293 OTUs at a 95% similarity level (Table S3). DNA extraction method did not affect the total number of reads obtained after the bioinformatics process (2-way ANOVA, $p = 0.1$), but sampling site did ($p = 0.008$). The smallest average number of reads was obtained for Pol2 (5183 reads) and the maximum was obtained for Ref7 (16,831 reads). We normalized read abundances for each sample to 4180 (lowest read number obtained for sample Pol2 with kit MN-plant; Table S3).

The values obtained for the Chao global richness estimator varied widely among environmental samples (94–436 OTUs). The Chao global richness estimator and Shannon diversity index values did not differ among DNA extraction methods (Table S4, S5).

The NMDS based on OTU similarity showed that the 78 DNA samples were discriminated mainly according to sampling site (Fig. 3A). Sampling site explained 90% of the total variance (PERMANOVA, $R^2 = 0.90$, $p < 0.001$), whereas DNA extraction method explained only 1.2% ($R^2 = 0.012$, $p < 0.001$). Site-by-site analysis with PERMANOVA showed that the DNA extraction method explained 72 to 91% of the total variance in 6 environmental samples, whereas no significant effect was assessed for 2 samples (Fig. 3B).

We conducted further analyses focused on the SA-Gen and MN-Soil methods, which provided the most different results in terms of quantity, quality of DNA, and community structure (Fig. 4A). We used SIMPER to identify the OTUs that were the main contributors (contribution > 1%) to the




B)

| | Variance explained | p-value |
|---|---|---|
| Edian | $R^2 = 0.81$ | **$p < 0.001$** |
| Aire | $R^2 = 0.90$ | **$p < 0.001$** |
| Lake | $R^2 = 0.86$ | **$p < 0.001$** |
| 767 | $R^2 = 0.72$ | **$p < 0.002$** |
| M36 | $R^2 = 0.91$ | **$p < 0.004$** |
| P45 | $R^2 = 0.42$ | $p = 0.63$ |
| Ref7 | $R^2 = 0.61$ | $p = 0.25$ |
| Pol2 | $R^2 = 0.72$ | **$p < 0.001$** |

Figure 3. A.—Three-dimensional nonmetric multidimensional scaling (NMDS) plot of Bray–Curtis dissimilarity based on operational taxonomic unit (OTU) composition of the DNA extracts obtained from the 8 environmental samples (extraction performed in duplicate for the 5 extraction methods), stress value = 0.18. B.—Results obtained by permutational analysis of variance (PERMANOVA) to reveal the effect of the DNA extraction method on dissimilarity values within sites.

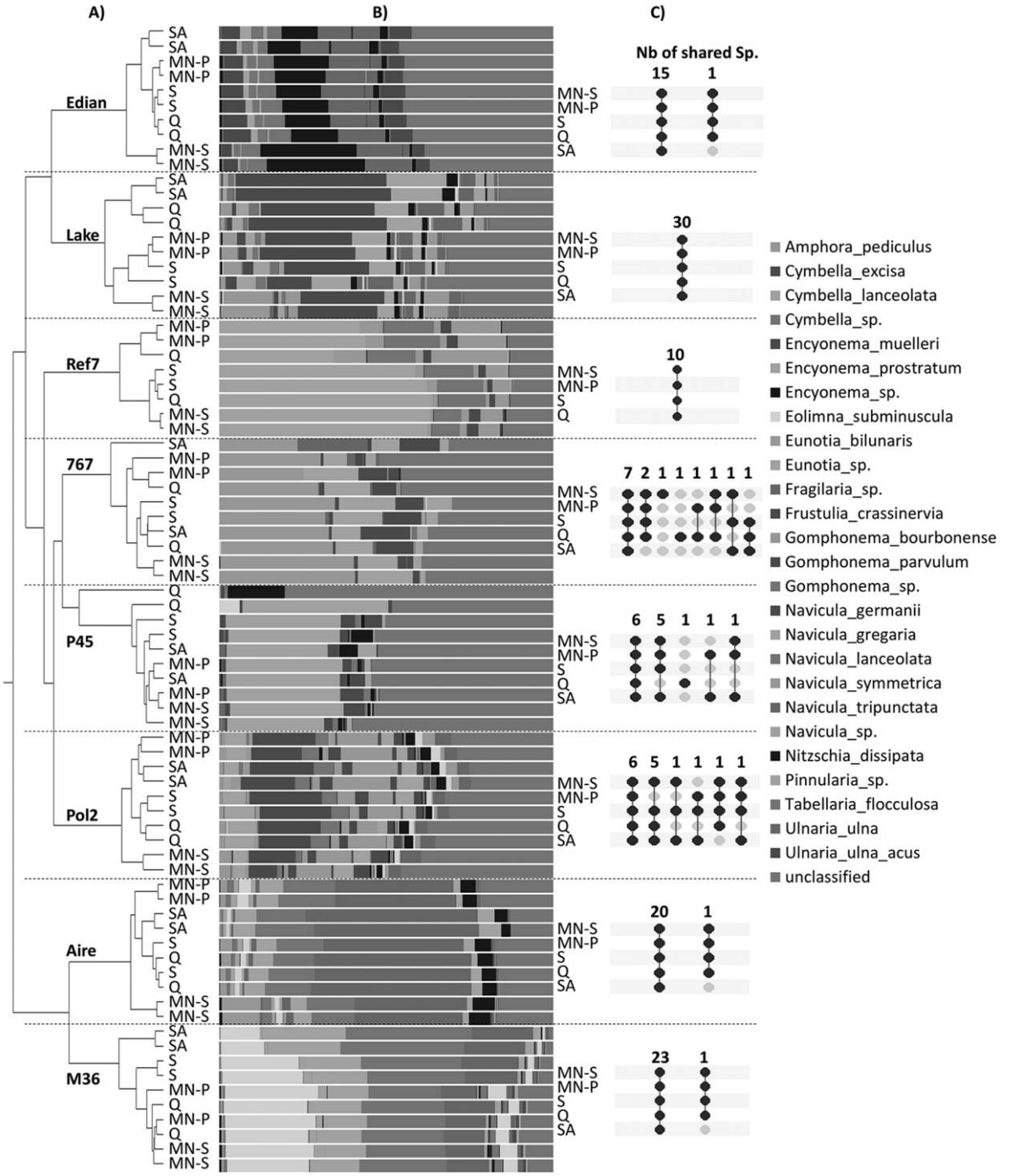



Figure 4. Diatom community structure as revealed by molecular inventory for the 5 DNA extraction methods (each performed in duplicate) at 8 sampling sites. A.—Hierarchical clustering tree based on Bray–Curtis dissimilarity computed from species lists (proportion of reads). B.—Histograms presenting relative abundances of species (legend shows only species with relative abundance > 1%). C.—Number of species detected and shared among DNA extraction methods (Nb. of shared sp.). Only species represented by >10 reads at each sampling site were used. MN-S = MN-Soil, MN-P = MN-Plant, S = S-Plant, Q = Q-Blood, SA = SA-Gen.

dissimilarity between communities assessed using the SA-Gen and MN-Soil methods (Table 3). Among the 6 environmental samples, 38 OTUs were identified as main contributors. Differences between the 2 methods were mostly (in 98% of cases) a result of variations in the read abundances of common OTUs rather than the presence or absence of OTUs detected by only 1 of the 2 methods. OTUs assigned to the genera *Nitzschia* and *Amphora* were represented by

Table 3. Results of the similarity percentages (SIMPER) analysis performed to identify the operational taxonomic units (OTUs) contributing to >1% of the dissimilarity between diatom communities obtained from the SA-Gen (SA) and MN-Soil extraction methods (MN). The list of contributors is presented for each environmental sample. Read abundances obtained with the 2 methods (SA/MN reads) and the ratios of read abundances for each OTU are presented. – = no contribution was found or the OTU was absent from the sample.

| | | | Edian | Aire | Lake | 767 | M36 | Pol2 |
|---|---|---|---|---|---|---|---|---|
| Family | Genus | OTU | SA/MN reads (ratio) | SA/MN reads (ratio) | SA/MN reads (ratio) | SA/MN reads (ratio) | SA/MN reads (ratio) | SA/MN reads (ratio) |
| Achnanthidiaceae | *Planothidium* | n°043 | – | – | – | – | 31/163 (0.2) | – |
| Amphipleuraceae | *Frustulia* | n°031 | – | – | – | 382/52 (7.3) | – | – |
| Bacillariaceae | *Nitzschia* | n°021 | – | 115/226 (0.5) | – | – | – | – |
| | | n°088 | – | – | – | – | – | 42/141 (0.3) |
| | Unclassified | n°188 | – | – | – | – | – | 2/102 (0.2) |
| Catenulaceae | *Amphora* | n°027 | – | – | 21/564 (0.04) | – | – | – |
| | | n°039 | – | 19/374 (0.05) | – | – | – | – |
| | | n°113 | – | – | – | – | – | 22/146 (0.2) |
| Cymbellaceae | *Encyonema* | n°004 | – | – | 1840/992 (1.9) | – | – | – |
| | | n°013 | – | – | 553/141 (3.9) | – | – | – |
| | | n°065 | 53/163 (0.3) | – | – | – | – | – |
| | | n°070 | – | – | 104/20 (5.2) | – | – | – |
| | | n°128 | – | – | 88/1 (88.0) | – | – | – |
| Eunotiaceae | *Eunotia* | n°010 | – | – | – | 804/1199 (0.7) | – | – |
| | | n°036 | – | – | – | 81/408 (0.2) | – | – |
| | | n°038 | – | – | – | 217/68 (3.2) | – | – |
| | | n°048 | – | – | – | 193/280 (0.7) | – | – |
| | | n°083 | – | – | – | 345/0 | – | – |
| Fragilariaceae | *Fragilaria* | n°017 | 298/423 (0.7) | – | – | – | – | – |
| | *Pseudostaurosira* | n°047 | – | – | 81/264 (0.3) | – | – | – |
| Gomphonemataceae | *Gomphonema* | n°012 | – | – | – | – | – | 129/38 (3.4) |
| | | n°041 | – | – | – | – | – | 365/229 (1.6) |
| | | n°044 | – | – | – | – | – | 323/214 (1.5) |
| | Unclassified | n°025 | – | – | – | 542/826 (0.7) | – | – |
| Naviculaceae | *Navicula* | n°001 | 386/71 (5.4) | 2008/1321 (1.5) | – | – | 733/480 (1.5) | – |
| | | n°003 | – | 581/454 (1.3) | – | – | 1298/816 (1.6) | – |
| | | n°008 | – | – | – | – | 623/361 (1.7) | – |
| | | n°009 | – | – | – | – | – | 613/414 (1.5) |
| | | n°018 | – | – | – | – | 243/122 (2.0) | – |
| | | n°046 | – | 195/15 (13.0) | – | – | – | – |
| Pinnulariaceae | *Caloneis* | n°063 | – | 22/114 (0.2) | – | – | – | – |
| Sellaphoraceae | *Eolimna* | n°005 | – | – | – | – | 467/1029 (0.5) | – |
| Skeletonemataceae | *Discostella* | n°085 | – | – | 21/151 (0.1) | – | – | – |
| Unclassified | Unclassified | n°022 | – | – | 90/298 (0.31) | – | – | – |
| | | n°024 | – | – | – | – | 31/227 (0.1) | – |
| | | n°034 | – | – | – | – | – | 171/567 (0.3) |
| | | n°055 | – | – | 39/160 (0.2) | – | – | – |
| | | n°067 | - | – | – | – | – | 69/257 (0.3) |

more reads when we used the MN-Soil method, whereas OTUs belonging to the genera *Encyonema*, *Gomphonema*, and *Navicula* were represented by more reads when we used the SA-Gen method (Table 3).

Some OTUs could not be assigned at the species level. The proportion of DNA sequences that remained unclassified at the species level varied from 1.5% (sample M36) to 78% (sample P45). On average, considering all sampling site and extraction methods, 71% of the reads were assigned to the species level. The comparison of the species inventories from the 78 libraries (Fig. 4A) showed, as previously observed for the OTUs, that samples clustered primarily by sampling site. Community structures based on species composition were similar among methods for each sample except samples 767 and P45 (Fig. 4B, C) for which we suspected potential bias during the initial subsampling (small sample volume [52µL] used for DNA extraction with sample 767 and difficulty homogenizing sample P45).

### Morphology vs molecular diatom community composition

The taxonomic lists obtained with the HTS approach for each environmental sample with each of the extraction methods were compared at genus and species levels to those obtained with the classical microscopy-based approach (Fig. 5). In general, 43% of the genera (maximum = 61.5% for Edian sample) and 18% of the species (maximum = 34.5% for sample M36) were detected by both approaches. Sixty-three percent of species were detected only by microscopy (on average for all samples), whereas only 19% of specific-HTS species were observed. However, a very high number of OTUs could not be assigned to a precise species because the reference database was incomplete (68% of species detected only by microscopy were not represented in the database).

SPI values calculated based on diatom lists identified by HTS and by microscopy were compared (Fig. 6). SPI values were consistent with expected water-quality status (Fig. 1B) for both French and Swedish sites. The SPI has not been adapted for Mayotte Island yet and cannot be used to infer quality status there. Different DNA extraction methods provided similar SPI values, which were close to SPI values obtained by microscopy except for Mayotte samples (Pol2 and Ref 7).

## DISCUSSION

### DNA extraction method affects quantity and quality of extracted DNA

The highest DNA extraction efficiency was observed for both diatom pure cultures and biofilm samples with the SA-Gen method, which outperformed the 4 commercial DNA extraction kits in terms of DNA quantity. Elution parameters (e.g., elution volume, temperature) are known to affect DNA yield when commercial kits are used, and yield loss can range from 20 to 30% (according to the manufacturer's specifications). However, the difference of efficiency between the commercial kits and the SA-Gen method is much higher than this % variation, and elution conditions alone fail to explain the low DNA concentrations obtained with the 4 commercial kits compared to the SA-Gen method.

We think the lysis method particularly affected DNA extraction efficiency, as previously suggested by Deiner et al. (2015) for eubacteria and freshwater eukaryotes. Various lysis methods including freezing–thawing (Fuhrman et al. 1988), enzymes (Somerville et al. 1989), liquid N (Bruckner et al. 2008), sonication (Chung et al. 2005), or bead beating (Yuan et al. 2015) can be used to disrupt the cell wall prior

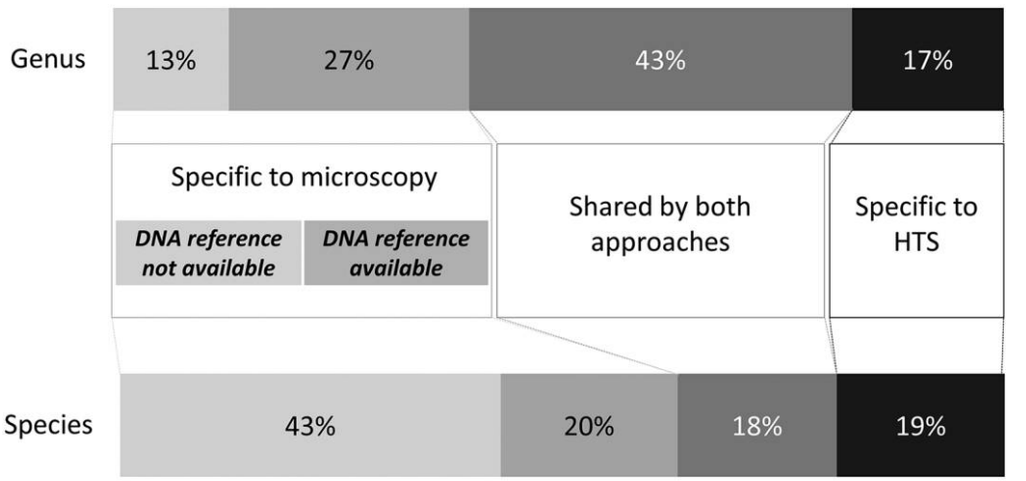



Figure 5. Mean percentage of diatom genera and species detected by microscopy, by molecular inventory, or by both methods for the 8 sampling sites. For all genera and species detected by only microscopy, the presence/absence of their DNA reference in the molecular database is specified. Unclassified operational taxonomic units were not used. HTS = high-throughput sequencing.

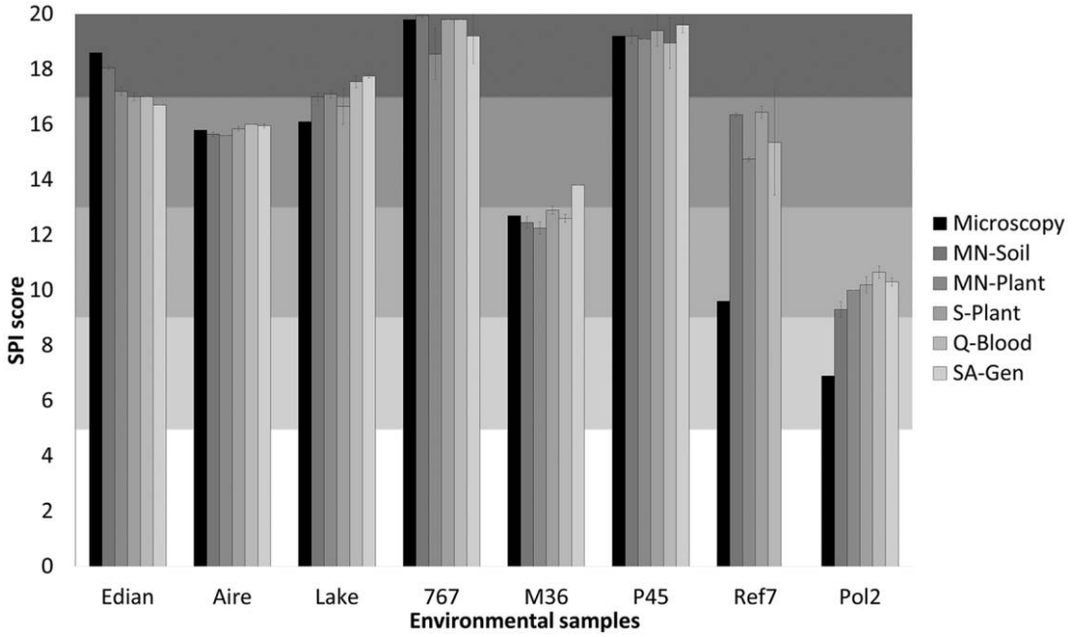



Figure 6. Specific pollution-sensitivity index (SPI) values based on morphological inventories (counts obtained from microscopic observations of diatom valves) and on molecular inventories (reads based on relative abundances estimated by high throughput sequencing [HTS]). When species taxonomic level was not reached, ecological values at the genus level were used for the calculation. Shading/colors correspond to the water-quality thresholds with low values indicating poor quality and high values indicating good quality.

to DNA extraction. Bead beating, as presented in MN-Soil, has been used in metabarcoding studies of benthic eukaryotic communities, including diatoms, because it saves time and works with complex environmental matrices like sediments or biofilms (Chariton et al. 2015, Zimmermann et al. 2015). However, diatom cells are protected by a robust silica valve that limits the ability of these classic lysis methods, even bead beating (Eland et al. 2012), to disrupt the diatom cell and release DNA. The SA-Gen method combines different lysis mechanisms (sonication, enzyme, temperature variation) to recover high quantities of DNA from both pure cultures and environmental samples. For pure cultures, the quantity of DNA collected was >2× higher with the SA-Gen protocol than with the other methods tested. We also observed higher efficiency of SA-Gen for environmental samples, but we could not assess precisely which part of this total DNA was from diatoms because DNA from other organisms (bacteria or other microbes present in the biofilms) was co-extracted. However, DNA extracted by SA-Gen from environmental samples provided the highest quantity of *rbc*L copy (per mg of wet biofilm), as revealed by qPCR with our diatom-specific primers, and the best quantitative correlation to diatom valves counted by microscopy.

SA-Gen is an in-house method that does not include a silica column during the purification step. All the other protocols applied in our study include a silica column, so we assume that part of the DNA could have been lost by remaining fixed to the purification column. The effect of DNA purification methods on DNA recovery has been studied for soils and sediment samples (Miller et al. 1999). In these studies, use of a column reduced recovery to 80%, and in some cases (e.g., sediment samples), as low as 40% of initial DNA concentrations. In our study, the 4 methods that included column purification (MN-Soil, MN-Plant, S-Plant, and Q-Blood) reached a maximum efficiency of 69% with diatom pure cultures and only 27% with environmental samples relative to the SA-Gen method.

The DNA purification steps could be useful for DNA originating from environmental matrices that may contain a number of compounds that inhibit or decrease the sensitivity of PCR. PCR-inhibitor molecules in extracted DNA can come from residual compounds of the extraction process (e.g., ethanol) or from molecules that are co-extracted with DNA (e.g., protein, polysaccharides, humic acids) (Schrader et al. 2012). Based on the 260/280 ratios, we assume that the quality of our DNA extracts was not affected by protein contamination (for both pure culture and environmental samples) or by residual contamination of the extraction (for pure culture), regardless of which extraction method was used. The main source of inhibition was co-extracted environmental compounds. Based on real-time PCR results, we were able to estimate the level of inhibition present in all the DNA extracts from environmental samples. The SA-Gen method produced extracts with the highest inhibition level, whereas extracts produced with methods that included a column purification step were only slightly inhibited. Making serial dilutions of DNA ex-

tracts to reach a concentration of inhibitors that is low enough not to inhibit PCR reactions is an efficient strategy to overcome the problem of PCR inhibitors, but this approach requires a large initial amount of DNA so that diluting it does not affect its representativeness. SA-Gen produced a large-enough quantity of DNA to permit improvement of its quality by dilution. When a low concentration of DNA is extracted, another option for overcoming the problem of inhibition is to use a column purification step to complete the SA-Gen method. We were able to purify DNA extracted with the SA-Gen method from Ref 7 samples that could not be amplified because of inhibitors with the aid of a DNA purification column (NucleoSpin® gDNA Clean-up, Macherey–Nagel). Despite the high loss of DNA during the purification (minimum = 40% loss), purified DNA quantities exceeded the quantities obtained with the MN-Soil kit, and we were able to use PCR amplification on the purified DNA without the need for dilution (data not shown). This result suggests that adding a column-purification step to the SA-Gen method could be a good solution for low DNA-concentration samples (<3 ng DNA/μL) containing very high levels of PCR inhibitors.

### Diatom community composition is unchanged whatever the extraction method

No effect of DNA extraction methods on OTU richness and diversity was found, and the sample origin appeared to be the main source of variation in our study. This inter-sample variation is consistent with the contrasting characteristics of our environmental samples (origin, quality status), which harbor different diatom community composition. Despite the presence of PCR inhibitors, SA-Gen provided a picture of diatom community similar to that provided by the other methods.

Regardless of taxonomic level (OTU or species), the taxonomic composition of the community represented in the extracts was not affected by DNA extraction methods. We observed 81.5% of shared species between the 5 methods, and when we removed the 2 samples with initial subsampling bias (767 and P45), this value increased to 93.8%. However, proportional reads did differ among extraction methods for some taxa. The observed intrasample variation (~27%) is consistent with variation observed in studies of bacterial community structure in water (Staley et al. 2015) or salivary samples (Lazarevic et al. 2013). These investigators found variations in relative abundance at the order and phylum/genus levels that were related to DNA extraction methods. Such variations are usually attributed to biases within the extraction process. Some diatom genera appeared to be preferentially detected with MN-Soil (*Nitzschia*, *Amphora*) or with SA-Gen methods (*Encyonema*, *Gomphonema*, *Navicula*), indicating that all methods did not extract DNA equally from all taxa. Depending on the diatom species, the skeleton can display different features (e.g., shape, size, thickness) and different proportions of silica (Barker 1992). Mechanical resistance of diatom skeletons can vary from one species to another depending on factors, such as porosity or shape (Hamm et al. 2003, Moreno et al. 2015). Some diatom species are more resistant than others to mechanical lysis (bead beating) during the DNA extraction, and this resistance affects the relative abundances obtained (Koid et al. 2012, Manoylov et al. 2016). We hypothesize that diatom species with long and thin skeletons may be more easily broken by mechanical lysis than small species with thick skeletons, thereby affecting their relative representation in the molecular inventory. Considering the samples for which we obtained an efficient taxonomic assignment we can verify that small species (<20-μm length) were proportionally less represented in the molecular inventories than in the morphological ones, whereas the species >50 μm long appeared to be proportionally more abundant in the molecular inventories (data not shown).

### Accuracy of molecular inventories and downstream quality indices

Diatom taxonomic inventories, based on assigned OTUs, were compared with taxonomic inventories based on morphological data for each DNA extraction method. Slight deviation between molecular and morphological diatom taxa was observed, but none of the DNA extraction methods provided a better match with microscopy than the others because they all shared similar diatom taxa. We found only 2 exceptional deviations in taxonomic composition data (from samples 767 and P45), and both were most probably caused by initial subsampling bias. We considered all observed taxa from the morphological inventories but only diatom taxa with robust taxonomic assignment from the molecular inventories when we compared molecular and morphological inventories. The molecular inventories were especially incomplete for Swedish and tropical samples because they encompassed taxonomic diversity that is not well represented in the Rsyst::diatom database. This problem illustrates the need to continue updating the barcode reference database to provide more complete coverage of diatom diversity. The quality and completeness of the molecular reference database used to make the link between molecular data and diatom references is crucial in metabarcoding studies, as already pointed by Zimmermann et al. (2014) for diatoms. Rsyst::diatom, the molecular reference database we used, was created to provide a reliable database curated by diatom taxonomist experts. The unsolved problem is the difficulty of enriching the molecular reference database with new taxa that can be identified unambiguously based on morphology. This requirement carries with it the need for the capacity to sample, isolate, accurately identify, and sequence new species. Single-cell PCR technique can help investigators obtain sequences from uncultured diatoms combined with their morphological identification (Hamilton et al. 2015). However, the efficiency of this approach is still low for diatoms and giving a

precise taxonomic identification at the species level is often impossible with only one cell.

Despite the partial match between molecular and morphological data, molecular SPI values were highly correlated to morphological SPI values for the 6 European samples (Edian, Aire, Lake, 767, M36, and P45). The Rsyst::diatom database was mainly populated with the DNA sequence of diatoms isolated from temperate regions, so it provided better molecular coverage for our European samples than the 2 tropical samples (Pol2 and Ref7). One of the mismatching taxa in the tropical samples was very abundant (*Nitzschia inconspicua*, 33.5 and 29.5% of total valves counts) only in the microscopy-based inventories, which added to the large differences between molecular and microscopy SPI values for these samples. This species is represented by many sequences in the Rsyst::diatom database but is a paraphyletic species and a "taxonomic mess" (Rovira et al. 2015), which yields incomplete taxonomic assignment at genus/species level. *Nitzschia inconspicua* has a medium-sensitivity indicator value and is usually present at sites with medium or poor quality status. Consequently, its absence from the molecular SPI calculation tends to give higher SPI values with metabarcoding than with microscopy.

SPI values obtained were consistent with quality status of all environmental sites, even for Mayotte sites for which the SPI calculation is not yet adapted to infer quality status. SPI calculation is driven mainly by species with abundances >5% (Bigler et al. 2010), and we were able to detect most abundant genera (75.9%) and species (33.6%) in our DNA inventories. This feature of the SPI can explain why molecular and morphological SPI values were highly correlated for European samples despite some deviations in diatom taxonomic lists.

Taxa must be quantified before quality indices can be calculated. The correlation between relative abundances of sequences obtained by HTS and diatom specimens observed by microscopy was not constant and varied from one taxon to another in our data. For many biological groups, DNA metabarcoding for quantifying relative abundances is limited by biological and technical biases that might influence sequence read counts (Thomas et al. 2015). These biases can have multiple origins and include DNA extraction efficiency; variation of copy number of the targeted genes, primer specificity, and PCR amplification efficiency that may differ among species; DNA sequencing errors and bioinformatics filtering that may affect DNA sequence reliability (Jeon et al. 2008, Amend et al. 2010, Bragg et al. 2013, Weber and Pawlowski 2013, Deiner et al. 2015, Elbrecht and Leese 2015, Schmidt et al. 2015). In the case of diatoms, the number of copies of *rbc*L per genome, the number of genomes per chloroplast, and the number of chloroplasts per cell may influence the correlation between DNA sequence counts and morphological counts. However, we assume that the number of copies of *rbc*L per genome and the number of chloroplasts per cell probably do not introduce major biases in DNA sequence counts. The sequenced plastid genomes currently available (e.g., Ruck et al. 2014) reveal that the *rbc*L gene is present as 1 copy per plastid genome. The number of plastids per cell varies from 1 to 4 for benthic diatoms (Round et al. 1990) indicating that a correction factor could be developed based on plastid genome number variation, as proposed by Angly et al. (2014) to correct bacterial quantification based on 16S ribosomal RNA amplicons. Additional technical biases linked to primer efficiency, PCR amplification, and sequencing errors are not easily estimated and corrected unless some control material can be introduced as an internal standard, as proposed for quantification in fish (Thomas et al. 2015) or alien DNA to estimate the fraction of amplicons captured by the sequence library (Gifford et al. 2011, Mangot et al. 2013). As a follow-up to our comparison of DNA extraction methods, further investigation could be done to estimate the importance of technical/biological biases and to test the feasibility of potential correction factors to improve quantification of HTS and to adapt calculations of quality indices.

## Conclusion

Our results show that all of the DNA extraction methods tested provide DNA of sufficient quality and quantity to perform benthic diatom community analysis based on HTS to obtain reliable molecular inventories of diatoms. The composition of diatom assemblages obtained was not affected by the choice of DNA extraction method. The relative abundances of some taxa can vary with the efficiency of lysis methods to disrupt diatom cells, but this variability did not affect the SPI value.

The operating cost of following propositions of Kermarrec et al. (2014) to implement next-generation biomonitoring with diatom metabarcoding as an alternative to classical morphological approach has to be considered. The cost per sample may vary depending on the HTS technologies used (Loman et al. 2012), but Stein et al. (2014) showed that DNA metabarcoding can be a valid economic solution for biomonitoring programs at the national scale. They showed for algal indicators (including diatoms) that the costs of molecular and classical methods for sampling and analysis are similar. The SA-Gen method, which is 24 to 39× times cheaper than other DNA extraction methods (including analysis and equipment costs), is an attractive choice to decrease the cost of next-generation biomonitoring. Moreover, this method provides a large quantity of DNA from environmental samples and the best correlation between *rbc*L copy number and valve observed by microscopy. The low quality DNA and the presence of PCR inhibitors in SA-Gen extracts did not affect diatom composition and SPI calculation, so we encourage the use of the SA-Gen method to perform DNA extraction for HTS diatom biomonitoring purposes.

Use of metabarcoding for biomonitoring is a complex workflow that requires standardization. We have provided a benchmark for the first step of this workflow. Further work is required for standardization of the full process, including reference database update, quantification, bioinformatics workflow, and adaptation of methods for calculating indices.

## ACKNOWLEDGEMENTS

Author contributions: AB, ID, FR, and VV contributed to the study conception and design. MK, FR, and VV were involved in acquisition of data. VV performed bioinformatics treatments, and AB, ID, MK, FR, and VV were involved in analysis and interpretation of data. AB, ID, MK, FR, and VV participated in drafting the article and revising it critically.

We thank Kalman Tapolczai for participating in the morphological identification and counting the diatoms in our environmental samples. We also thank Franck Salin, Christophe Boury, and Erwan Guichoux from the "Plateforme Génome Transcriptome" (PGTB, Bordeaux, France) who performed HTS sequencing and provided fastq files containing DNA reads. We extend special thanks to Alain Franc and Philippe Chaumeil from the "Biodiversité, Gènes et communautés" (INRA Biogeco) scientific team for helpful discussions. We thank the laboratory technical support of Cécile Chardon and Louis Jacas for their help and advice, and people from the INRA Rsyst network from which the Rsyst::diatom database was initiated. This work was funded by the French National Agency for Water and Aquatic Environments (ONEMA). The Swedish contribution was funded by the Swedish Agency for Marine and Water Management and by the SLU Environmental Monitoring and Assessment program Lakes and Watercourses.

## LITERATURE CITED

AFNOR (Agence Française de Normalisation). 2003. Water quality guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers. European Standard EN 13946:1–15.

AFNOR (Agence Française de Normalisation). 2004. Water quality–Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. European Standard EN 14407:1–13.

Amend, A. S., K. A. Seifert, and T. D. Bruns. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Molecular Ecology 19:5555–5565.

Angly, F. E., P. G. Dennis, A. Skarshewski, I. Vanwonterghem, P. Hugenholtz, and G. W. Tyson. 2014. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome 2:11.

Barker, P. 1992. Growth and reproductive strategies of freshwater phytoplankton. Regulated Rivers: Research and Management 7:308–309.

Bigler, C., V. Gälman, and I. Renberg. 2010. Numerical simulations suggest that counting sums and taxonomic resolution of diatom analyses to determine IPS pollution and ACID acidity indices can be reduced. Journal of Applied Phycology 22: 541–548.

Bourrelly, P., and E. Manguin. 1952. Algues d'eau douce de la Guadeloupe et dépendances. SEDES, Paris, France.

Bragg, L. M., G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson. 2013. Shining a light on dark sequencing: characterising errors in ion torrent PGM data. PLoS Computational Biology 9:e1003031.

Bruckner, C. G., R. Bahulikar, M. Rahalkar, B. Schink, and P. G. Kroth. 2008. Bacteria associated with benthic diatoms from Lake Constance: phylogeny and influences on diatom growth and secretion of extracellular polymeric substances. Applied and Environmental Microbiology 74:7740–7749.

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

Cemagref. 1982. Étude des méthodes biologiques quantitative d'appréciation de la qualité des eaux. Bassin Rhône-Méditerranée-Corse. Rapport Division Qualité des Eaux Lyon, Agence financière de Bassin Rhône–Méditerranée–Corse, Pierre-Bénite, France.

Chariton, A. A., S. Stephenson, M. J. Morgan, A. D. L. Steven, M. J. Colloff, L. N. Court, and C. M. Hardy. 2015. Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. Environmental Pollution 203:165–174.

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

Chung, C.-C., S.-P. L. Hwang, and J. Chang. 2005. Cooccurrence of ScDSP gene expression, cell death, and DNA fragmentation in a marine diatom, *Skeletonema costatum*. Applied and Environmental Microbiology 71:8744–8751.

Deiner, K., J. C. Walser, E. Mächler, and F. Altermatt. 2015. Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. Biological Conservation 183:53–63.

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

Eland, L. E., R. Davenport, and C. R. Mota. 2012. Evaluation of DNA extraction methods for freshwater eukaryotic microalgae. Water Research 46:5355–5364.

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.

Fuhrman, J. A., D. E. Comeau, A. Hagström, and A. M. Chan. 1988. Extraction from natural planktonic microorganisms of DNA suitable for molecular biological studies. Applied and Environmental Microbiology 54:1426–1429.

Gallup, J. M., and M. R. Ackermann. 2006. Addressing fluorogenic real-time qPCR inhibition using the novel custom Excel file system "FocusField2-6GallupqPCRSet-upTool-001" to attain consistently high fidelity qPCR reactions. Biological Procedures Online 8:87–153.

Gibson, J. F., S. Shokralla, C. Curry, D. J. Baird, W. A. Monk, I. King, and M. Hajibabaei. 2015. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. PLoS ONE 10:e0138432.

Gibson, K. E., K. J. Schwab, S. K. Spencer, and M. A. Borchardt. 2012. Measuring and mitigating inhibition during quantitative real time PCR analysis of viral nucleic acid extracts from large-volume environmental water samples. Water Research 46: 4281–4291.

Gifford, S. M., S. Sharma, J. M. Rinta-Kanto, and M. A. Moran. 2011. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. ISME Journal 5:461–472.

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.

Hamm, C. E., R. Merkel, O. Springer, P. Jurkojc, C. Maier, K. Prechtel, and V. Smetacek. 2003. Architecture and material properties of diatom shells provide effective mechanical protection. Nature 421:841–843.

Jeon, S., J. Bunge, C. Leslin, T. Stoeck, S. Hong, and S. S. Epstein. 2008. Environmental rRNA inventories miss over half of protistan diversity. BMC Microbiology 8:222.

Kahlert, M., M. Kelly, R.-L. Albert, S. F. P. Almeida, T. Bešta, S. Blanco, M. Coste, L. Denys, L. Ector, M. Fránková, D. Hlúbiková, P. Ivanov, B. Kennedy, P. Marvan, A. Mertens, J. Miettinen, J. Picinska-Fałtynowicz, J. Rosebery, E. Tornés, S. Vilbaste, and A. Vogel. 2012. Identification versus counting protocols as sources of uncertainty in diatom-based ecological status assessments. Hydrobiologia 695:109–124.

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.

Koid, A., W. C. Nelson, A. Mraz, and K. B. Heidelberg. 2012. Comparative analysis of eukaryotic marine microbial assemblages from 18S rRNA gene and gene transcript clone libraries by using different methods of extraction. Applied and Environmental Microbiology 78:3958–3965.

Krammer, K. 2000. The genus *Pinnularia*. *In* H. Lange-Bertalot (editor). Diatoms of Europe. Diatoms of the European inland waters and comparable habitats. Volume 1. Gantner Verlag, Ruggell, Germany.

Krammer, K. 2001. *Navicula* sensu stricto, 10 genera separated from *Navicula* sensu stricto, *Frustulia*. Gantner Verlag, Ruggell, Germany.

Krammer, K. 2002. *Cymbella*. *Cymbella*. *In* H. Lange-Bertalot (editor). Diatoms of Europe. Diatoms of the European inland waters and comparable habitats. Volume 3. Gantner Verlag, Ruggell, Germany.

Krammer, K. 2003. *Cymbopleura*, *Delicata*, *Navicymbula*, *Gomphocymbellopsis*, *Afrocymbella*. *In* H. Lange-Bertalot (editor). Diatoms of Europe. Diatoms of the European inland waters and comparable habitats. Volume 4. Gantner Verlag, Ruggell, Germany.

Krammer, K., and H. Lange-Bertalot. 1986. Bacillariophyceae 1. Teil: Naviculaceae. Süßwasserflora von Mitteleuropa. Gustav Fischer Verlag, Stuttgart, Germany.

Krammer, K., and H. Lange-Bertalot. 1988. Bacillariophyceae 2. Teil: Bacillariaceae, Epithemiaceae, Surirellaceae. Süßwasserflora von Mitteleuropa. Gustav Fischer Verlag, Stuttgart, Germany.

Krammer, K., and H. Lange-Bertalot. 1991a. Bacillariophyceae 3. Teil: Centrales, Fragilariaceae, Eunotiaceae. Süßwasserflora von Mitteleuropa. Gustav Fischer Verlag, Stuttgart, Germany.

Krammer, K., and H. Lange-Bertalot. 1991b. Bacillariophyceae 4. Teil: Achnanthaceae. Kritische Ergänzungen zu *Navicula* (Lineolatae) und *Gomphonema*. Gesamtliteraturverzeichnis Teil 4. Süßwasserflora von Mitteleuropa. Gustav Fischer Verlag, Stuttgart, Germany.

Lazarevic, V., N. Gaïa, M. Girard, P. François, and J. Schrenzel. 2013. Comparison of DNA extraction methods in analysis of salivary bacterial communities. PLoS ONE 8:e67699.

Lecointe, C., M. Coste, and J. Prygiel. 1993. Omnidia: software for taxonomy, calculation of diatom indices and inventories management. Hydrobiologia 269–270:509–513.

Lloyd, K. G., B. J. MacGregor, and A. Teske. 2010. Quantitative PCR methods for RNA and DNA in marine sediments: maximizing yield while overcoming inhibition. FEMS Microbiology Ecology 72:143–151.

Loman, N. J., R. V Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen. 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology 30:434–439.

Majaneva, M., K. Hyytiäinen, S. L. Varvio, S. Nagai, and J. Blomster. 2015. Bioinformatic amplicon read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of communities. PLoS ONE 10:e0130035.

Mangot, J.-F., I. Domaizon, N. Taib, N. Marouni, E. Duffaud, G. Bronner, and D. Debroas. 2013. Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. Environmental Microbiology 15: 1745–1758.

Mann, D. G., and P. Vanormelingen. 2013. An inordinate fondness? The number, distributions, and origins of diatom species. Journal of Eukaryotic Microbiology 60:414–420.

Manoylov, K., France, Y., Geletu, A., and J. Dominy. 2016. Algal community membership of estuarine mudflats from the Savannah River, United States. Journal of Marine Science and Engineering 4(1):11.

Metzeltin, D., and H. Lange-Bertalot. 1998. Tropical diatoms of South America I. Iconographia Diatomologica 5:1–695.

Metzeltin, D., and H. Lange-Bertalot. 2007. Tropical diatoms of South America II. Iconographia Diatomologica 18:1–877.

Miller, D. N., J. E. Bryant, E. L. Madsen, and W. C. Ghiorse. 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. Applied and Environmental Microbiology 65:4715–4724.

Moreno, M. D., K. Ma, J. Schoenung, and L. P. Dávila. 2015. An integrated approach for probing the structure and mechanical properties of diatoms: toward engineered nanotemplates. Acta Biomaterialia 25:313–324.

Motwani, N. H., and E. Gorokhova. 2013. Mesozooplankton grazing on picocyanobacteria in the Baltic Sea as inferred from molecular diet analysis. PLoS ONE 8:e79230.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48: 443–453.

Nguyen, T. N. M., M. Berzano, C. O. Gualerzi, and R. Spurio. 2011. Development of molecular tools for the detection of freshwater diatoms. Journal of Microbiological Methods 84:33–40.

Rimet, F., P. Chaumeil, F. Keck, L. Kermarrec, V. Vasselon, M. Kahlert, A. Franc, and A. Bouchez. 2016. R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. Database 2016:baw016.

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

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

Rovira, L., R. Trobajo, S. Sato, C. Ibáñez, and D. G. Mann. 2015. Genetic and physiological diversity in the diatom *Nitzschia inconspicua*. Journal of Eukaryotic Microbiology 62:815–832.

Rubin, B. E. R., J. G. Sanders, J. Hampton-Marcell, S. M. Owens, J. A. Gilbert, and C. S. Moreau. 2014. DNA extraction protocols cause differences in 16S rRNA amplicon sequencing efficiency but not in community profile composition or structure. Microbiology Open 3:910–921.

Ruck, E. C., T. Nakov, R. K. Jansen, E. C. Theriot, and A. J. Alverson. 2014. Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms. Genome Biology and Evolution 6:644–654.

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

Schmidt, T. S. B., J. F. Matias Rodrigues, and C. von Mering. 2015. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. Environmental Microbiology 17: 1689–1706.

Schrader, C., A. Schielke, L. Ellerbroek, and R. Johne. 2012. PCR inhibitors—occurrence, properties and removal. Journal of Applied Microbiology 113:1014–1026.

Somerville, C. C., I. T. Knight, W. L. Straube, and R. R. Colwell. 1989. Simple, rapid method for direct isolation of nucleic acids from aquatic environments. Applied and Environmental Microbiology 55:548–554.

Staley, C., T. J. Gould, P. Wang, J. Phillips, J. B. Cotner, and M. J. Sadowsky. 2015. Evaluation of water sampling methodologies for amplicon-based characterization of bacterial community structure. Journal of Microbiological Methods 114:43–50.

Stein, E. D., M. C. Martinez, S. Stiles, P. E. Miller, and E. V. Zakharov. 2014. Is DNA barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the United States? PLoS ONE 9:e95525.

Stevenson, R. J., and Y. Pan. 1999. Assessing environmental conditions in rivers and streams with diatoms. The Diatoms: Applications for the Environmental and Earth Sciences 1:4.

Stoof-Leichsenring, K. R., L. S. Epp, M. H. Trauth, and R. Tiedemann. 2012. Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. Molecular Ecology 21:1918–1930.

Thomas, A. C., B. E. Deagle, J. P. Eveson, C. H. Harsch, and A. W. Trites. 2015. Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. Molecular Ecology Resources 16:714–726.

Tudesque, L., F. Rimet, and L. Ector. 2008. A new taxon of the section *Nitzschiae lanceolatae* Grunow: *Nitzschia costei* sp. nov. compared to *N. fonticola* Grunow, *N. macedonica* Hustedt, N. tropica Hustedt and related species. Diatom Research 23: 483–501.

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.

Wagner Mackenzie, B., D. W. Waite, and M. W. Taylor. 2015. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Frontiers in Microbiology 6:1–11.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73:5261–5267.

Weber, A. A.-T., and J. Pawlowski. 2013. Can abundance of protists be inferred from sequence data: a case study of foraminifera. PloS ONE 8:e56739.

Werner, P., S. Adler, and M. Dreßler. 2016. Effects of counting variances on water quality assessments: implications from four benthic diatom samples, each counted by 40 diatomists. Journal of Applied Phycology 28:2287–2297.

Wesolowska-Andersen, A., M. Bahl, V. Carvalho, K. Kristiansen, T. Sicheritz-Pontén, R. Gupta, and T. Licht. 2014. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome 2:19.

Willner, D., J. Daly, D. Whiley, K. Grimwood, C. E. Wainwright, and P. Hugenholtz. 2012. Comparison of DNA extraction methods for microbial community profiling with an application to pediatric bronchoalveolar lavage samples. PLoS ONE 7:e34605.

Yuan, J., M. Li, and S. Lin. 2015. An improved DNA extraction method for efficient and quantitative recovery of phytoplankton diversity in natural assemblages. PLoS ONE 10: e0133060.

Zimmermann, J., N. Abarca, N. Enke, N. Enk, 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.