Environmental Science & Technology

Article

pubs.acs.org/est

Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data

Joana Amorim Visco,[†] Laure Apothéloz-Perret-Gentil,[†] Arielle Cordonier,[‡] Philippe Esling,[†,§] Loïc Pillet,[†,∥] and Jan Pawlowski*[,†]

[†]Department of Genetics and Evolution, University of Geneva, Boulevard d'Yvoy 4, CH 1205 Geneva, Switzerland

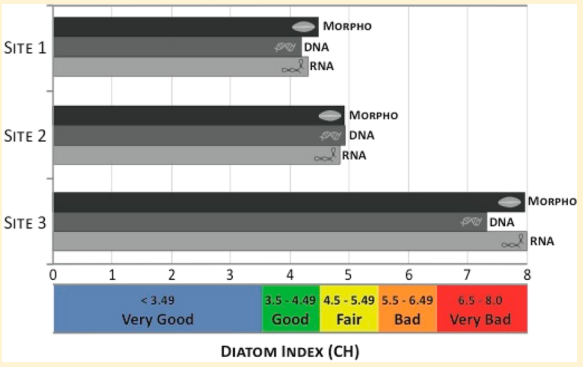
[‡]Water Ecology Service, Department of Environment, Transports and Agriculture, Canton of Geneva, CH 1205 Geneva, Switzerland

[§]IRCAM, UMR 9912, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

[∥]CNRS, UMR 7144, Laboratoire Adaptation et Diversité en Milieu Marin, Place Georges Teissier, CS90074, 29688 Roscoff, France

Supporting Information

ABSTRACT: Diatoms are widely used as bioindicators for the assessment of water quality in rivers and streams. Classically, the diatom biotic indices are based on the relative abundance of morphologically identified species weighted by their autoecological value. Obtaining such indices is time-consuming, costly, and requires excellent taxonomic expertise, which is not always available. Here we tested the possibility to overcome these limitations using a next-generation sequencing (NGS) approach to identify and quantify diatoms found in environmental DNA and RNA samples. We analyzed 27 river sites in the Geneva area (Switzerland), in order to compare the values of the Swiss Diatom Index (DI-CH) computed either by microscopic quantification of diatom species or directly from NGS data. Despite gaps in the reference database and variations in relative abundance of analyzed species, the diatom index shows a significant correlation between morphological and molecular data indicating similar biological quality status for the majority of sites. This proof-of-concept study demonstrates the potential of the NGS approach for identification and quantification of diatoms in environmental samples, opening new avenues toward the routine application of genetic tools for bioassessment and biomonitoring of aquatic ecosystems.



INTRODUCTION

Diatoms are phototrophic protists common in all aquatic ecosystems and widely used as bioindicators of environmental conditions, particularly in rivers and streams.[1,2] The applications of diatoms as bioindicators range from routine monitoring of water quality to the assessment of industrial pollution impact.[3−6] Because diatoms are highly sensitive to environmental conditions and grow rapidly, they respond quickly to changes in chemical, physical, or biological factors. Hence, analyzing the composition of their communities provides an easy method to detect environmental changes due to natural or anthropogenic causes.

Various biotic indices have been developed to assess environmental impact using diatoms.[7] Most of these indices are based on the relative frequency of species weighted by their autoecological value and eventually other index-specific factors. In Europe, the Water Framework Directive[8] recommends using diatoms to assess water quality, but the computation of diatom indices vary from one country to another.[2] In Switzerland, the Swiss Diatom Index (DI-CH) was proposed in order to characterize the biological status of rivers and streams using the frequencies and distributions of more than 400 diatom species and morphological varieties.[9] The DI-CH classifies watercourses into 5 categories, corresponding to *very good*, *good*, *average*, *poor*, and *bad* degree of pollution, as established by the Swiss Federal Council in the Waters Protection Ordinance.[10]

The DI-CH is calculated as follows

$$\text{DI-CH} = \frac{\sum_{i=1}^{n} D_i G_i H_i}{\sum_{i=1}^{n} G_i H_i}$$

where D_i is the factor based on the autoecological value for taxon i, G_i is the weighting factor for taxon i, H_i is the relative frequency of taxon i in a studied sample (number of valves found for the taxon i divided by the total number of valves counted), and n is the total number of taxa found in a sample.

The main limitation of all other diatom indices is related to the species identification being based on morphology. Indeed, diatoms constitute one of the most specious groups of protists, with the number of species estimated at nearly 200 000.[11] However, most freshwater diatoms are small (usually <50 um),

Received: December 18, 2014
Revised: June 2, 2015
Accepted: June 8, 2015
Published: June 8, 2015

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

and their microscopic identification requires special sample preparation methods and expert taxonomic knowledge. The size, shape, and design of diatom valves are the main features used for taxonomic identification of diatom species. Yet, intraspecific variability can be very high, and some morphological characters can become indistinct as a result of size reduction during the life cycle. In some cases, the morphological differences between species are so subtle that even trained taxonomists may come to different conclusions.[12]

Over the past decade, molecular barcoding has become widely recognized as an efficient tool for species identification. This approach is based on the assumption that a short DNA sequence (DNA barcode) contains enough information to distinguish species. The main advantage of using DNA barcodes in applied studies is that standardization and automation of the protocols is easier than that in the traditional morphology-based approach. Several diatom barcoding studies have been performed based mainly on the analysis of five genes: cox1,[13,14] the rbcL gene,[15,16] the ITS region,[17,18] the V4 region of the 18S rDNA,[19,20] and the D2/D3 region of the LSU rRNA gene.[15] Although there is no consensus on the ideal diatom DNA barcode, it has been proposed that some highly discriminating barcodes (ITS, cox1) are more suitable for taxonomic studies, whereas those that are less variable but more universal (18S, rbcL) are more appropriate for applied studies.[12]

Recent developments of next-generation sequencing (NGS) technologies offer the possibility to use molecular barcoding for fast and reliable diversity surveys based on environmental samples. NGS-based environmental monitoring has been proposed as a time and cost-effective alternative to the traditional morphology-based approaches.[21−23] Several experimental studies have been conducted on NGS-based inventories of freshwater benthic macroinvertebrates.[24−26] The major gaps highlighted by these studies include the incompleteness of the database, the technical biases, and the irrelevance of NGS quantitative data as compared to the abundance of specimens. Previous studies focusing specifically on diatoms completed their taxonomic reference database, evaluated different DNA barcodes, and compared the composition of diatom communities inferred from microscopic and NGS data.[27−30] One of these studies also briefly compared the diatom indices computed from morphological and molecular data,[28] although presently this aspect has still not been thoroughly examined.

Here, we test the hypothesis that the use of NGS could lead to a similar assessment of the water quality as the morphological study. To do so, we analyze the diatom communities in 27 watercourses of the Geneva basin, using the hypervariable region V4 of 18S rDNA as the diatom DNA barcode and the Illumina Miseq platform for high-throughput sequencing. Assuming that the RNA provides a better proxy for active cells, we compare the DNA and RNA data for the relative abundance of each taxon in order to test which ones fit better to the morphological data. Finally, we compute the DI-CH values for each site and compare them with the values inferred from microscopic study. We analyze the congruence between NGS and morphological analyses and discuss the current limitations of NGS approach that should be overcome to reduce the divergence between molecular and morphological indices.

MATERIALS AND METHODS

Sampling. The samples were collected in 2013−14 as part of a routine bioassessment campaign performed by the Service of Water Ecology (SECOE) of the Department of Environment, Transport, and Agriculture in Geneva, Switzerland.[31] The biofilm containing epilithic diatoms was collected from 27 sites located in shallow waterways of the Geneva basin following the directives established by the Swiss Federal Office for the Environment[9] (Supporting Information, SI, Table S1). Between three to five stones were selected at each sampling site. The periphyton taken by scratching the stones with diatom-scraping devices was resuspended with freshwater taken from the river and then transferred to sampling bottles. Each sample was homogenized and divided into two subsamples, one for morphological analysis by SECOE and the other for molecular analysis. Morphological samples were preserved in a concentrated (37%) formaldehyde solution, while molecular samples were kept cold (ca. 0 °C) during sampling (max. Four hours). Upon arrival to the laboratory, 1 mL of homogenized periphyton suspension was transferred to 1.5 mL tubes and centrifuged at 8000*g* for 10 min. The supernatant was discarded and the pellets stored at −80 °C until DNA/RNA extractions.

Morphological Analysis. Sample preparation, species identification, counting, and DI-CH calculations were performed as recommended by the Swiss Federal Office for the Environment.[9] Periphyton suspensions were sorted, and undesirable material was discarded. A decarbonation step using hydrochloric acid was performed, followed by the elimination of organic material by calcination combined with a treatment with hydrogen peroxide. Diatoms were then washed and mounted in Naphrax. Diatoms slides were observed using an Olympus light microscope with Nomarski differential interference contrast optics at a magnification of 1000x. Species identification was performed with the bibliographic support of The Flora of Diatoms,[32] Diatoms of Europe,[33] Iconographia Diatomologica,[34,35] and Diatomeen im Süsswasser-Benthos von Mitteleuropa.[36]

DNA/RNA Extraction. DNA and RNA were extracted with PowerBiofilm DNA and RNA isolation kits (MO BIO Laboratories Inc.) following the manufacturer instructions. RNA was purified from carried-over DNA molecules with TURBO DNase kit Ambion (Life Technologies) and cDNA obtained by reverse transcription using SuperScript III Reverse Transcriptase kit (Invitrogen). A total of 27 DNA and 27 cDNA (RNA) samples were obtained for this study.

For the extraction of cultured diatoms, pelleted cells were prepared by centrifuging 1 mL of fresh diatoms cultures at 8000*g* for 10 min. The extractions were then performed with DNeasy Plant Mini Kit (Qiagen) or PowerBiofilm DNA isolation (MO BIO).

Reference Database. We built a reference database of the V4 region composed of 460 unique diatom sequences. First, we downloaded from the GenBank database all sequences corresponding to the species and genera found in the morphological analyses of Geneva samples and also those commonly found in Switzerland.[9] The alignment was performed with the Seaview program.[37] Sequences were analyzed by Maximum Likelihood (ML) phylogenetic inference, and those showing incorrect identification were discarded. A total of 298 unique sequences from GenBank were kept.

To extend our reference database, we sequenced 10 diatom species obtained from culture collections: *Fragilaria pinnata* and

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

Nitzschia ovalis from the CCAP (Culture Collection of Algae and Protozoa, SAMS Research Services Ltd., Scottish Marine Institute, Oban, U.K., http://www.ccap.ac.uk), *Achnanthidium minutissimum*, *Achnanthidium pyrenaicum*, *Achnanthidium straubianum*, *Amphora pediculus*, *Cocconeis placentula*, *Encyonema silesiacum*, *Nitzschia palea*, and *Sellaphora seminulum* from the TCC (Thonon Culture Collection, INRA-UMR Carrtel, Thonon-les-Bains, France, http://www6.inra.fr/carrtel-collection). We also added 152 Sanger sequences from other eDNA analyses of Geneva watercourses. The sequences were submitted to the Genbank database (KR089906-KR090057, KR150668-KR150677).

PCR Amplification, Cloning, and Sanger Sequencing. To complete the reference database and to test the specificity of PCR primers, the diatom cultures and environmental samples cited above were examined. The hypervariable region V4 of the 18S rRNA gene was amplified using primers modified after Zimmermann[19] DIV4for: 5′-GCGGTAATTCCAGCTCCA-ATAG-3′, DIV4rev3:5′-CTCTGACAATGGAATACGAATA-3′. PCR amplifications were performed in a total volume of 25 μL using Taq DNA Polymerase by Roche Applied Science. PCR regime included an initial denaturation at 94 °C for 2 min, then 35 cycles of denaturation at 94 °C for 45 s, annealing at 50 °C for 45 s, elongation at 72 °C for 1 min, and a final elongation at 72 °C for 10 min. PCR amplicons were purified with a High Pure PCR Product Purification kit (Roche Applied Science) and cloned using a TOPO TA Cloning kit for sequencing (Invitrogen). Sequence reactions were performed with BigDye Terminator (Applied Biosystems), and sequences were obtained by Sanger sequencing on ABI PRISM 3130XL Genetic Analyzer System (Applied Biosystems/Hitachi).

PCR Amplification for Next-Generation Sequencing. PCR were performed on DNA and RNA (cDNA) isolated from periphyton samples using unique combinations of forward and reverse tagged primers. Individual tags are composed of 8 nucleotides attached at each primer's 5′-extremity. A total of 20 different forward and reverse tagged primers were designed to enable multiplexing of all PCR products in a unique sequencing library. PCRs were performed as described above. Purified PCR products were quantified by fluorometric method using QuBit HS dsDNA kit (Invitrogen). Concentrations were then calculated and normalized for all samples. Approximately 50 ng of amplicons of each DNA and RNA sample from the SECOE 2013 (DIATOM 2013) and 2014 (DIATOM 2014) campaigns were pooled. An amount of 100 ng of pooled amplicons was used for the Illumina library preparation.

Illumina Library Preparation and Sequencing. Indexed paired-end libraries of pooled amplicons for consecutive cluster generation and DNA sequencing were constructed using an Illumina TruSeq Nano DNA Sample Preparation Kit—Low Throughput. Libraries were prepared following the manufacturer instructions. The fragment sizes of each library were verified by loading 3 μL of the final product in a 1.5% agarose gel with 1x SYBRSafe (Invitrogen) and quantified by a fluorometric method using a QuBit HS dsDNA kit (Invitrogen). An MiSeq Reagent Nano kit v2, with 500 cycles with nano (2 tiles) flow cells was used to run libraries on the MiSeq System. Two 250 cycles were used for an expected output of 500 Mb and an expected number of 1 million reads per library.

NGS Data Analysis. Operational Taxonomic Units (OTUs) were obtained and assigned following the method described in Pawlowski et al.[38] using the diatoms reference database described above. Raw FASTQ reads were quality-filtered by removing any sequence with a mean quality score of 30, and also removing all sequences with ambiguous bases or any mismatch in the tagged primer or contig region. These extremely stringent parameters ensure that we keep only high-quality reads. Then, paired-end reads were assembled by aligning them into a contiguous sequence with highest similarity. In case of mismatching bases, we kept in the final contig the closest base from the read 5′-extremity, based on the fact that the probability of miscalls increases toward the 3′-extremity. These sequences were then demultiplexed (assigned to their corresponding sample) depending on the tagged primers found at each end. Dereplication of the data set obtained after assembly was necessary in order to obtain unique sequences, called Independent Sequence Units (ISUs). An abundance threshold of 10 was used for the minimum number of replicates found for each ISU, and this abundance was recorded for further analyses. Subsequently, ISUs were assigned by performing a pairwise Needleman—Wunsch global alignment against our entire reference database. For the ISUs that were not assigned at the end of this procedure, we relied on a BLAST filtering procedure. We removed the ISUs that did not match any Bacillariophyceae sequences in the NCBI database with at least 99% coverage and 97% identity.

Phylogenetic Analyses. The taxonomic assignment of OTUs was checked by phylogenetic analyses. A tree was built with all the sequences from the database and the OTUs from the NGS analysis. The most abundant ISU was used as the representative sequence for each OTU. The ML phylogeny was constructed using RAxML v.7.4.2,[39] with GTR + G as model of evolution and 1000 replicates for the bootstrap analysis. The OTUs were assigned to the reference morphospecies if they formed a clade supported by bootstrap values >60 (following Zimmermann et al.[29] and references cited therein).

RESULTS

NGS Data Statistics. For DIATOM 2013, we obtained 1 176 424 reads from Illumina sequencing (SI Table S2). The filtering process rejected 169 841 reads with low mean quality, 61 508 reads with low base quality, 2205 reads with not enough matching bases in the contig region and 177 325 reads with errors or mismatches in the primers. Hence, a total of 765 545 reads remained after filtering and were available for further analysis. For DIATOM 2014, we obtained 1 055 387 reads. The filtering process rejected 296 799 reads with low mean quality, 17 095 reads with low base quality, 152 394 reads with not enough matching bases in the contig region, 247 694 reads with errors or mismatches in the primers and 23 222 with insufficient sequence lengths. Hence, a total of 318 183 good reads remained for further analysis.

Morphological Data and DI-CH Calculation. For each sampling site, about 400 valves were observed and identified with light microscopy at SECOE. Morphospecies were counted, and the relative abundance of each taxon was calculated for each site (SI Table S3). A total of 96 species was found by morphological identification. The number of taxa per site varied from 5 (AMB) to 37 (HEB). One species (*Amphora pediculus*) was found at every site and represented the most abundant taxon counted for all sites together. The values of DI-CH were calculated using the formula presented previously. The DI-CH values varied from 3.64 (NAM) to 7.98 (AMB). Highest DI-CH values were obtained for sites with larger numbers of diatoms with high autoecological values, such as *Nitzschia*

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

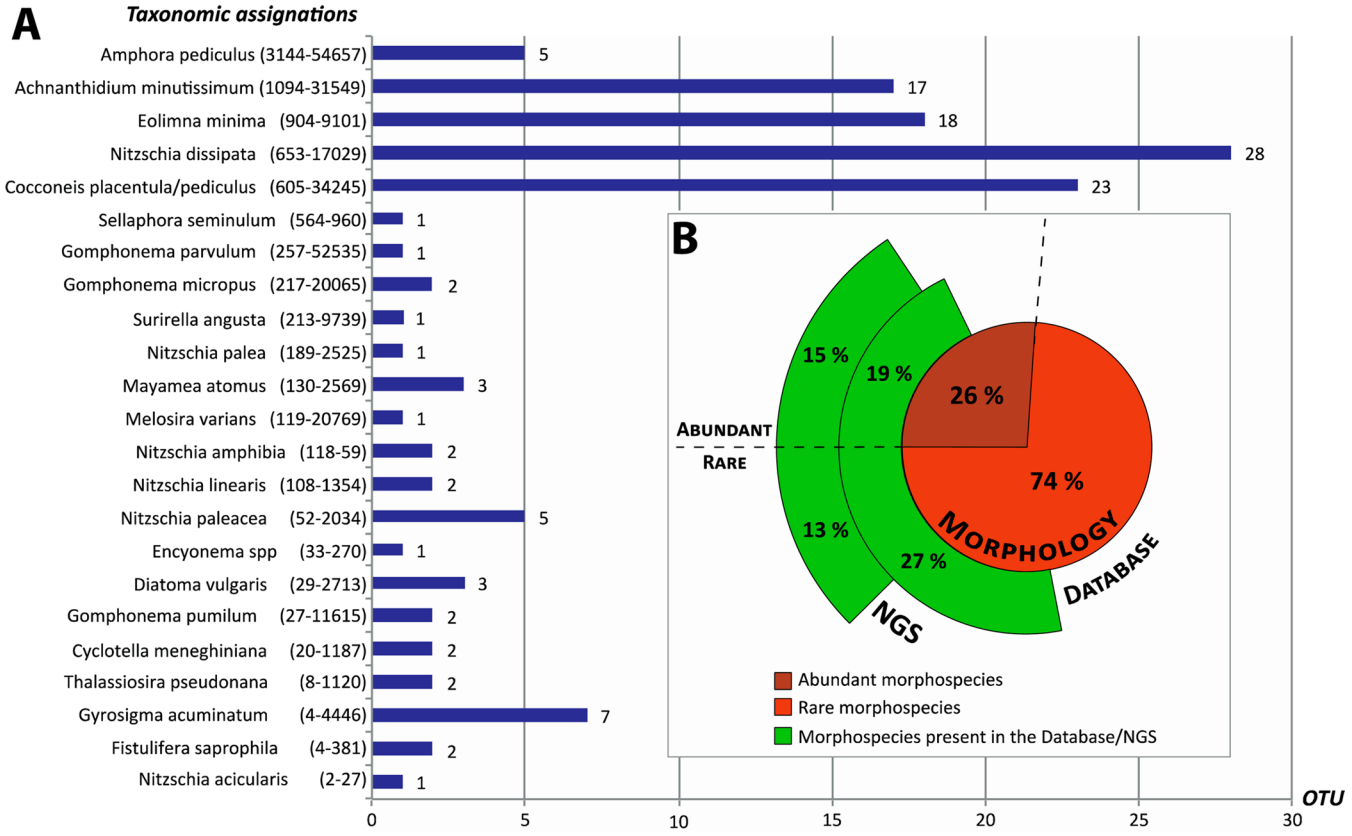


Figure 1. (A) Taxonomic assignments in common with morphospecies sorted by the number of counts in the morphologic analysis (in parentheses). The bar plot represents the number of OTU in each taxonomic assignation. (B) Pie chart of abundant (brown) and rare (orange) morphospecies found in morphologic analysis. Arcs in green represent the morphospecies present in the database (internal one) and in the NGS assignations (external one). Each arc is divided between abundant and rare species by a dashed line.

amphibia, *Sellaphora seminulum*, *Eolimna minima*, *Gomphonema micropus*, *Gomphonema parvulum*, *Eolimna subminuscula*, *Navicula veneta*, and *Nitzschia acicularis*.

Taxonomic Assignment of NGS data. Analysis of the NGS data grouped the reads into 242 OTU for the DIATOM 2013 and 103 for the DIATOM 2014 runs. In order to assign those OTUs to morphological taxa, an ML tree with all OTUs and our reference database was built. After phylogenetic analysis, we removed 128 OTUs for the DIATOM 2013 run and 60 OTUs for the DIATOM 2014 run because they could not be univocally assigned to any morphological clade. In total, 144 OTUs remained and were assigned to 30 taxa. Twenty-three of these taxa corresponded to the morphospecies found in microscopic analyses, while seven matched to species in the reference database that were not evidently found with the morphology-based approach.

Among the 23 assigned species (Figure 1A), 15 were confidently identified, i.e., they formed well-supported clades (BV > 60) including reference sequences assigned to a single morphospecies. *Encyonema* spp. was a special case since the only GenBank reference sequence of the clade was not identified beyond the genus level. Five species formed clades with reference sequences assigned to two different species of the same genus. These species were *Amphora pediculus*, *Achnanthidium minutissimum*, *Cocconeis placentula/pediculus*, *Mayamea atomus*, and *Fistulifera saprophila*.

Two assignments were particularly problematic. The OTUs assigned to *Cyclotella meneghiniana* formed a well-supported clade (BV 78) with 8 other *Cyclotella* species, half of which were marine species. We assigned these OTUs to *C. meneghiniana* because it was the only species present in the morphological list with an autoecological value. In the second case, the two OTUs assigned to the morphospecies *Thalassiosira pseudonana* formed a well-supported clade (BV 88) with 13 other *Thalassiosira* species and with the species *Stephanodiscus minutulus*. As both *S. minutulus* and *T.pseudonana* have the same autoecological value, we kept them together using the name of *T. pseudonana* as in morphological analyses.

In total, the number of morphospecies recognized in the NGS data amount to only 28% of all those identified in this study microscopically. However, it should be noted that the GenBank database only covers 46% of the morphospecies found in microscopic analyses (Figure 1B). The difference between these two percentages is accounted for by morphospecies (i.e., genus *Navicula*) that could not be identified unambiguously due to the lack of resolution of the V4 region. However, it is important to notice that most species not found in NGS were rare (below 100 counts in the morphologic analysis), as shown by Figure 1B. The list of the morphospecies with their count in the morphologic analysis and their presence in the database and in the NGS assignation are reported in SI Table S4.

Abundance of Assigned Species. As the calculation of diatom indices includes the relative abundance of species, we analyzed the variations in morphological counts and the number of reads inferred from DNA and RNA data for each assigned species. As can be seen in the SI (Table S5 and Figure

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

S1), the relative abundance of species per site varies considerably depending on the type of data. In particular, the proportion of a species in DNA samples is often lower than in morphological counts and RNA samples. We checked whether this could be a consequence of the higher abundance of undetermined sequences in the DNA data, by reanalyzing the data with assigned OTUs only. However, the proportions between DNA, RNA, and morphological abundances remain the same in most of the cases.

The correlation between the number of reads and individuals for the most ubiquitous and abundant species is significant for both DNA and RNA of *A. pediculus* and DNA of *A.minutissima* (Figure 2). The relative abundance of some species (*A. pediculus*, *E. minima*) is higher in morphocounts than in NGS data. However, among the assigned morphospecies, there are very few sites where the species was found in microscopic preparations but not in the NGS data. This deviation is more obvious in less common taxa, with species such as *Nitzschia amphibia* being found almost exclusively in morphological analyses, while some species (e.g., *Gyrosigma acuminatum*) or genera (e.g., *Gomphonema*) are overrepresented in NGS data (SI Figure S1).

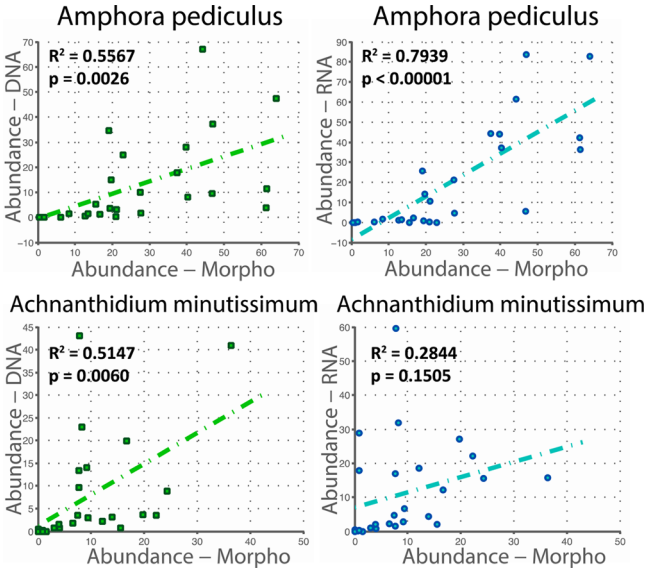


Figure 2. Relationships between the relative abundance of the two most abundant species *Amphora pediculus* (upper) and *Achnanthidium minutissimum* (lower). This information is displayed separately for DNA (left) and RNA (right) where each point shows the relationship between the relative abundance found in morphological (*x*-axis) or molecular (*y*-axis) counts. The dotted lines represent the results of model II regression with a least-squares fitting for the relative abundances of all samples. The R^2 and *p*-value are indicated for each regression axis.

Diatom Index. The NGS DI-CH index was calculated with the 23 taxa, for which the D and G values were available. When those values were different for a variety or subspecies of the same species, the values of the most abundant and frequent taxa were retained. All the DI-CH values for morphology, DNA, and RNA per site are presented in SI Table S6.

The variations in diatom indices inferred from morphological and molecular (DNA/RNA) data for 27 sites are illustrated in Figure 3. For the majority of sites (25 out of 27), the deviation between the morphological and at least one of the molecular indices (DNA or RNA) was less than 1 unit, and the biological quality status inferred from the two types of data was identical. For 17 sites (63%), the morphological index indicated the same level of water quality as at least one type of molecular data. Both DNA and RNA data were congruent with the morphological index in 7 out of 27 sites. When considered separately, the same level was indicated in 10 and 12 sites for DNA and RNA, respectively. The values of the morphological index exceeded those inferred from DNA and RNA in 16 sites (20 in the case of RNA). As we can see, the correlation between morphological and molecular indices is significant for DNA (Figure 4A) with $R^2 = 0.59$ and *p*-value = 0.0013 and becomes strongly supported in the case of RNA (Figure 4B) with $R^2 = 0.85$ and *p*-value <0.0001.

DISCUSSION

By exhibiting the strong similarity between the DI-CH values inferred from microscopic and NGS analyses of diatom communities, our proof-of-concept study clearly demonstrates the usefulness of NGS diatom data to evaluate water conditions. Our results confirm the previously reported similarity between values of the Specific Pollution Sensitivity biotic index obtained by microscopy and by NGS (pyrosequencing) analysis of SSU and rbcL barcodes.[28] Both studies

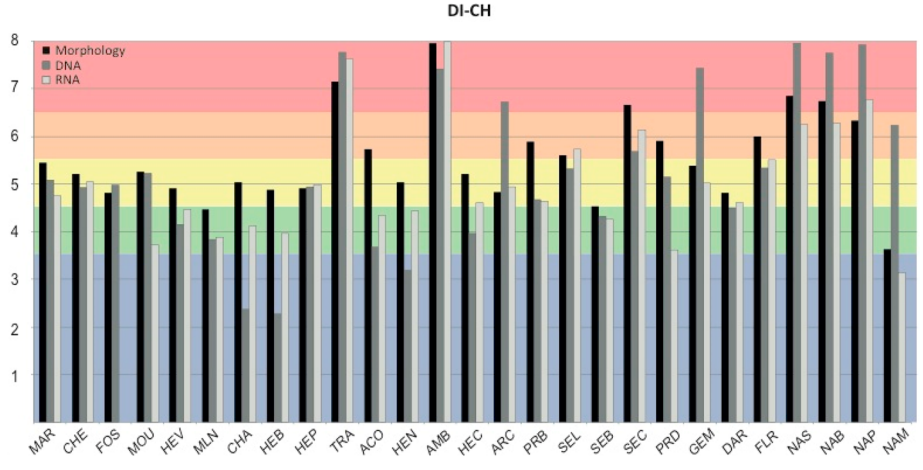


Figure 3. DI-CH values for morphologic analysis (black), DNA (dark gray), and RNA (light gray) per sites. Colors represent the threshold for water quality given by the DI-CH index.

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

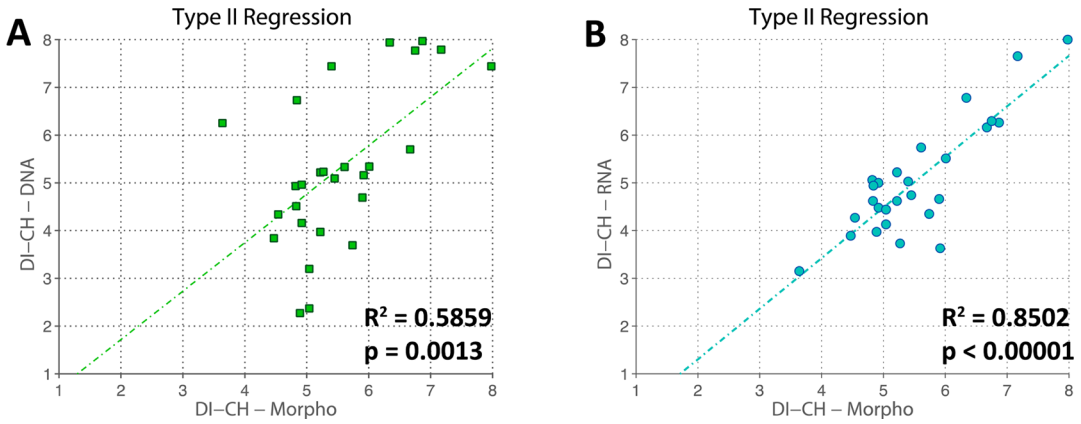


Figure 4. Relationships between the DI-CH inferred from morphological and DNA (A) or RNA (B) abundances per sites. Each point shows the relationship between the DI-CH found in morphological (*x*-axis) or molecular (*y*-axis) counts over all sites. The dotted lines represent the results of model II regression with a least-squares fitting for the relative abundances of all samples. The R^2 and *p*-value are indicated for each regression axis.

fully support the growing evidence that NGS environmental studies have the potential to become new tools for the assessment of aquatic ecosystems health, based on analysis of benthic macroinvertebrates,[24,25] diatoms,[27,29] and other protists.[38]

The congruence between diatom indices inferred either from morphological or NGS data is remarkable, given the poor database coverage and various technical biases. The correlation is especially strong for RNA (Figure 4B), likely because it provides a better depiction of the living diatom community composition. The DNA, however, can be preserved in water for a certain period of time and even carried over long distances.[40] Interestingly, the correlation between NGS and morphology in species relative abundances seems to have limited impact on the correlation between indices. This could be due to the fact that the index is calculated as the sum of a set of species with their respective weighting factors, which tends to reduce the effect of variations for individual species. In fact, a large number of species is assigned to the same set of weights, which means that the abundance of any given species can be replaced by the abundance of a set of several other species. Noticeably, the index correlates better in the sites with lower species richness, which might be related to the reduction of technical or biological biases in low complexity samples.

Although the results of our study are promising, there is still a wide potential to reduce the divergences between molecular and morphological results by addressing the current limitations of NGS data analysis. Some technical biases related to the DNA extraction, PCR conditions, primer specificity, library preparation, and sequence analysis have been extensively discussed in previous studies.[27,41,42] We discuss here the limitations that concern specifically the present study: (1) database incompleteness and inaccuracy, (2) inconsistencies between molecular and morphological taxonomy, and (3) biases in the quantitative analysis of NGS data.

Incompleteness and Inaccuracy of Databases. Gaps and misidentifications in reference databases are commonly believed to be the main hindrance to assigning taxonomy to environmental sequences. In fact, the diatom database is probably more exhaustive than that of any other groups of protists, especially those that cannot be cultivated.[43] The proportion of genetically characterized species in our study (46%) is slightly lower than in other studies targeting well-studied temperate regions (53−78%) but remains higher than those conducted in tropical regions (30−38%).[28] The development of comprehensive databases, like that of Zimmermann et al.,[30] which provided molecular (V4, rbcL) and morphological (LM, SEM) data for 70 cultured diatom strains, is an important step toward filling the gaps in diatom inventories. However, establishing cultures of diatom species for every eco-region could be extremely time-consuming and might not always be successful. An alternative approach could be based on single-cell PCR followed or preceded by LM or SEM study.[44] The success rate of these methods is still very low, but further developments in the field of single-cell genomics might rapidly improve their efficiency.

It should be noted that, although completing the database is important, it does not imply that the sequencing of all morphospecies is necessary. In our study, we assigned species according to very stringent criteria by removing all uncertain cases. Once the reference database is completed for common species such as *Achnanthes lanceolata*, and the identification of *Navicula* species is improved by using more rapidly evolving marker, the correlation between NGS and morphological indices might become even stronger. In fact, the vast majority of species currently missing from the database are rare, with less than 100 specimens per species counted in all samples. Their relative importance in the computation of diatom indices depends on the autoecological value associated with each species. However, it might be sufficient to correctly assign all common species and those rare species with high autoecological value to obtain a perfect match.

Molecular vs Morphological Taxonomy. Another potential source of conflict lies in the divergence between the morphological and molecular (phylogenetic) determination of diatom species. On the one hand, almost all morphospecies are represented by several genetically distinctive types. On the other hand, some morphospecies are subdivided into subspecies or morphological varieties, each with their own specific autoecological values. In the first case, the cryptic diversity may constitute a considerable advantage for biomonitoring, particularly if the cryptic species are associated with some specific ecological conditions. The second case is more problematic because the subspecific taxa are generally uncharacterized genetically.

In this study, we combined all subspecies and morphotypes belonging to the same species because it was impossible to distinguish them genetically. We also combined two species of

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

Cocconeis, to avoid a possible misidentification of numerous phylotypes forming the clade of *C. placentula*, among which *C. pediculus* branches. In our approach, we followed the principle that the species can be grouped if they share the same ecologies and morphologies[45] and if they form a clade in phylogenetic analysis. Grouping at the generic level[46] may be useful, as in the case of *Encyonema*, but it is not necessary and may even be inappropriate in the case of polyphyletic genera.

Taxonomic resolution largely depends on the choice of the DNA barcode. Until now, only the chloroplastic rbcL and nuclear ribosomal 18S V4 region have been used in NGS diatom studies. Here, we chose the V4 region because its amplification from eDNA samples is easier and its size better fits the sequencing length of Illumina Miseq. It has been shown that the taxonomic resolution of V4 (and 18S in general) is lower than rbcL.[27] However, the interspecies variation of a given barcode may change between genera, and its efficiency will depend on the taxonomic composition of diatom community.[29] For example, in our study, the resolution of V4 was too low to unambiguously assign *Navicula* species, but it was sufficient to distinguish most of the species of *Nitzschia* and *Gomphonema*. Ideally, as both V4 and rbcL barcodes are complementary they should be used together in NGS analyses.

Relative abundance. Undoubtedly, the quantitative analysis of NGS data presents the greatest challenge in efforts to alleviate biases in the calculation of diatom indices. Indeed, numerous NGS environmental surveys exhibited discrepancies between the number of sequences assigned to a given species and the number of specimens of the same species in microscopic preparations[47,48] or even mock communities.[49] This lack of correlation between the abundance of reads and individuals could be explained either by technical biases introduced during DNA extraction, PCR amplification or sequencing,[50] or by biological factors such as the variations of rRNA gene copies,[51] which may depend on genome size,[52] number of nuclei,[53] or differences in cell size.[54]

Our study shows that molecular and morphological counts are well correlated in some species, but differ significantly in others (Figure 2). These variations seem taxon-specific and could be explained by variation in the numbers of rRNA gene copies in different diatom species. However, the ground-truth biological data necessary to test such a hypothesis are not available for diatoms. In fact, the correlation between molecular and morphological abundance data was previously observed in the NGS study of changes in foraminiferal[38] and metazoan (unpublished data) communities associated with the environmental impact of fish-farming, as well as in the study of the seasonal abundance in some species of ciliates and chrysophytes.[55] As the match between microscopic and molecular abundances concerns mainly the abundant species, this could explain why the impact of abundance variations on the final computation of the diatom index is relatively moderate.

Future Perspectives. The results presented in this pilot study will require validation by further NGS-based surveys of diatom diversity. In particular, substantial efforts will need to be done by diatom taxonomists and biologists to complete the DNA barcoding reference database and to determine the range of genetic and morphological variation in diatom species. Better knowledge of diatom genomes, especially the quantification of nuclear and chloroplast gene copies, will help in improving the estimation of species abundance from molecular data. Additional NGS studies of diatom communities in different ecological settings are also needed in order to optimize the molecular protocols and improve the accuracy of NGS data analysis, in particular to use the correction factors that would help overcoming the biases in relative abundance estimations.

All these efforts are worthwhile considering the tremendous benefits that the routine application of NGS approaches would bring to diatom-based monitoring. First, the use of DNA barcodes will allow standardization of species identification, which will help in overcoming the recurrent problems of misidentification and will facilitate the comparison of species inventories. Second, the molecular approach will provide more accurate real-time assessment of living communities, especially if RNA is analyzed rather than DNA. Third, the use of NGS technology coupled with the automation of molecular protocols will considerably reduce the time for sample processing, which will, in turn, allow an increase in the number of monitored sites. Finally, given the rapidly diminishing costs of NGS technologies, the application of these new tools will allow important savings.

■ ASSOCIATED CONTENT

Supporting Information

Table S1, Site locations, geographic references, and sampling dates performed along the Geneva basin (Switzerland) in collaboration with SECOE-DETA and used for the study. Table S2, Showing the filtering process on libraries DIATOM 2013 and DIATOM 2014. Table S3, Relative abundance and DI-CH values of morphological data per site location. Table S4, List and counting of species found during the morphological analysis of the two campaigns and their presence in the database (DN) and in the molecular assignation (NGS). Table S5, Relative abundance of morphologic, DNA and RNA data per sites. Table S6, DI-CH values for morphologic, DNA, and RNA data per sites. Figure S1, Relative abundance of 23 assigned taxa inferred for morphology (red), DNA (light green), and RNA (blue). The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/es506158m.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +41 22 3793069; fax: +41 22 379 33 40; e-mail: Jan.Pawlowski@unige.ch (J.P.).

Author Contributions

J.A.V. and L.A.P.G. contributed equally to this work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

Swiss National Science Foundation, Swiss Federal Office for the Environment, G&L Claraz Donation

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Frédérique Rimet and Agnez Bouchez for the cultures of diatoms and helpful discussion, and Andrew Gooday for comments on the manuscript. We also thank François Pasquini from Water Ecology Service of the canton of Geneva for providing the infrastructure and equipment. Financial support was provided by the Swiss National Science Foundation (Grants 316030_150817 and 31003A-140766)

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

and G & L Claraz Donation. This study is a part of the SwissBOL program supported by the Swiss Federal Office for the Environment.

■ ABBREVIATIONS

NGS next generation sequencing
eDNA environmental DNA
DI-CH Swiss Diatom Index

■ REFERENCES

(1) Stevenson, R. J.; Pan, Y.; van Dam, H. Assessing environmental conditions in rivers and streams with diatoms. In *The Diatoms: Applications of the Environmental and Earth Sciences*; Stoermer, E. F., Smol, J. P., Eds.; Cambridge University Press: Cambridge UK, 2010; 57 p.

(2) Rimet, F. Recent views on river pollution and diatoms. *Hydrobiologia* **2012**, *683*, 1−24.

(3) Belore, L. M.; Winter, J. G.; Duthie, H. C. Use of diatoms and macroinvertebrates as bioindicators of water quality in southern Ontario rivers. *Can. Water Resour. J.* **2002**, *27*, 457−484.

(4) Lobo, E. A.; Callegaro, V. L. M.; Hermany, G.; Bes, D.; Wetzel, C. A.; Oliveira, M. A. Use of epilithic diatoms as bioindicators from lotic systems in southern Brazil, with special emphasis in eutrophication. *Acta Limnol. Bras.* **2004**, *16*, 25−40.

(5) Poulickova, A.; Duchoslav, M.; Dokulil, M. Littoral diatom assemblages as bioindicators of lake trophic status: a case study from perialpine lakes in Austria. *Eur. J. Phycol.* **2004**, *39*, 143−152.

(6) Martin, G.; Fernandez, M. R. Diatoms and indicators of water quality and ecological status: sampling, analysis and some ecological remarks. In *Ecological Water Quality—Water Treatment and Reuse*; Voudouris, K., Ed.; InTech: North Canton, OH, 2012; pp 182−203.

(7) Kelly, M. G.; Bennett, C.; Coste, M.; Delgado, C.; Delmas, F.; et al. 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* **2009**, *621*, 169−182.

(8) Directive 2000/60/EC of the European Parliament and of the Council of 23 October, 2000, establishing a framework for Community action in the field of water policy. *Official Journal L 327*, **22/12/2000** pp 0001−0073.

(9) Hürlimann, J.; Niederhauser, P. *Méthodes d'Analyse et d'Appréciation des Cours d'Eau. Diatomées Niveau R (region)*; Etat de l'environnement n° 0740. Office Fédéral de l'Environnement: Berne, 2007, 132p.

(10) *WPO—Water Protection Ordinance 814.201; 1998*. The Swiss Federal Council, based on Articles 9, 14 paragraph 7, 16, 19 paragraph 1, 27 paragraph 2, 36a paragraph 2, 46 paragraph 2, 47 paragraph 1, and 57 paragraph 4 of the Waters Protection Act of 24 January 1991 (WPA).

(11) Mann, D. G.; Droop, S. J. M.; Kristiansen, J. Biodiversity, biogeography and conservation of diatoms. Biogeography of freshwater algae. *Hydrobiologia.* **1996**, *336*, 19−32.

(12) Mann, D. G.; Sato, S.; Trobajo, R.; Vanormelingen, P.; Souffreau, C. DNA barcoding for species identification and discovery in diatoms. *Cryptogam.: Algol.* **2010**, *31*, 557−577.

(13) Evans, K. M.; Wortley, A. H.; Mann, D. G. An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist* **2007**, *158*, 349−364.

(14) Evans, K. M.; Mann, D. G. A proposed protocol for nomenclaturally effective DNA barcoding of microalgae. *Phycologia* **2009**, *48* (1), 70−74.

(15) Hamsher, S. E.; Evans, K. M.; Mann, D. G.; Poulickova, A.; Saunders, G. W. Barcoding diatoms: exploring alternatives to COI-5P. *Protist* **2011**, *162*, 405−422.

(16) Macgillivary, M. L.; Kaczmarska, I. Survey of the efficacy of a short fragment of the rbcL gene as a supplemental DNA barcode for diatoms. *J. Euk. Microbiol.* **2011**, *58*, 529−536.

(17) Moniz, M. B. J.; Kaczmarska, I. Barcoding micro- and meso-fauna. Barcoding diatoms: is there a good marker? *Mol. Ecol. Res.* **2009**, *9*, 65−74.

(18) Moniz, M. B. J.; Kaczmarska, I. Barcoding of diatoms: nuclear encoded ITS revisited. *Protist* **2010**, *161*, 7−34.

(19) Zimmermann, J.; Jahn, R.; Gemeinholzer, B. Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Org. Divers. Evol.* **2011**, *11*, 173−192.

(20) Luddington, I. A.; Kaczmarska, I.; Lovejoy, C. Distance and character-based evaluation of the V4 region of the 18S rRNA gene for the identification of diatoms (Bacillariophyceae). *PLoS One* **2012**, *7*, 1−11.

(21) Baird, D. J.; Hajibabaei, M. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* **2012**, *21*, 2039−2044.

(22) Bohmann, K.; Evans, A.; Gilbert, T. M. P.; Carvalho, G. R.; Creer, S.; Knapp, M.; Yu, D. W.; de Bruyn, M. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.* **2014**, *29* (6), 358−367.

(23) Taberlet, P.; Coissac, E.; Pompanon, F.; Brochmann, C.; Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **2012**, *21*, 2045−2050.

(24) Hajibabaei, M.; Shokralla, S.; Zhou, X.; Singer, G. A. C.; Baird, D. J. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One* **2011**, *6*, e17497.

(25) Hajibabaei, M.; Spall, J. F.; Shokralla, S.; van Konynenburg, S. Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecol.* **2012**, *12*, 28.

(26) Carew, M. E.; Pettigrove, V. J.; Metzeling, L.; Hoffmann, A. A. Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Front. Zool.* **2013**, *10*, 45.

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

(28) Kermarrec, L.; Franc, A.; Rimet, F.; Chaumeil, P.; Frigerio, J. M.; Jean-François Humbert, J. F.; Bouchez, A. A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Sci.* **2014**, *33*, 349−363.

(29) Zimmermann, J.; Glöckner, G.; Jahn, R.; Enke, N.; Gemeinholzer, B. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Mol. Ecol. Res.* **2014**, *15*, 526−542.

(30) Zimmermann, J.; Abarca, N.; Enk, N.; Skibbe, O.; Kusber, W.-H.; Jahn, R. Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. *PLoS One* **2014**, *9*, e108793.

(31) Cordonier, A.; Gallina, N.; Nirel, P. M. Essay on the characterization of environmental factors structuring communities of epilithic diatoms in the major rivers of the canton of Geneva, Switzerland. *Vie Milieu/Life Environ.* **2010**, *60*, 223−231.

(32) Krammer, K.; Lange-Bertalot, H. Bacillariophyceae. In *Süsswasserflora von Mitteleuropa*; Gustav Fischer Verlag: Stuttgart, Germany. 1986−1991a,b. Teil 1−4.

(33) Lange-Bertalot, H., Ed. *Diatoms of the European Inland Waters and Comparable Habitats*; ARG Gantner Verlag KG: Ruggell, Lichtenstein, 2001−2003, Vols. *2*−*4*.

(34) Lange-Bertalot, H.; Metzeltin, D. *Indicators of Oligotrophy*; Koeltz Scientific Books: Königstein im Taunus, Germany, 1996, 390 p.

(35) Reichardt, E. *Zur Revision der Gattung Gomphonema*; Koeltz Scientific Books: Königstein im Taunus, Germany, 1999, 203 p.

(36) Hofmann, G.; Werum, M.; Lange-Bertalot, H. *Diatomeen im Süßwasser—Benthos von Mitteleuropa*; Koeltz Scientific Books: Königstein im Taunus, Germany, 2011.

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605

(37) Gouy, M.; Guindon, S.; Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **2010**, *27*, 221−224.
(38) Pawlowski, J.; Esling, P.; Lejzerowicz, F.; Cedhagen, T.; Wilding, T. A. Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Mol. Ecol. Resour.* **2014**, *14*, 1129−40.
(39) Stamatakis, A.; Aberer, A. J.; Goll, C.; Smith, S. A.; Berger, S. A.; Izquierdo-Carrasco, F. RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics.* **2012**, *28*, 2064−6.
(40) Deiner, K.; Altermatt, F. Transport distance of invertebrate environmental DNA in a natural river. *PLoS One* **2014**, *9*, e88786.
(41) Lee, C. K.; Herbold, C. W.; Polson, S. W.; Wommack, K. E.; Williamson, S. J.; McDonald, I. R.; Cary, S. C. Groundtruthing next-gen sequencing for microbial ecology−biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One* **2012**, *7*, e44224.
(42) Esling, P.; Lejzerowicz, F.; Pawlowski, J. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res.* **2015**, *43*, 2513−2524.
(43) Pawlowski, J.; Audic, S.; Adl, S.; Bass, D.; Belbahri, L.; et al. CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biol.* **2012**, *10*, 1−5.
(44) Lang, I.; Kaczmarska, I. A protocol for a single-cell PCR of diatoms from fixed samples: method validation using *Ditylium brightwellii* (T. West) Grunow. *Diatom Res.* **2013**, *26*, 43−49.
(45) DeNicola, D. M. A review of diatoms found in highly acidic environments. *Hydrobiologia* **2000**, *433*, 111−122.
(46) Rimet, R.; Bouchez, A. Biomonitoring river diatoms: Implications of taxonomic resolution. *Ecol. Indic.* **2012**, *15*, 92.
(47) Nolte, V.; Pandey, R. V.; Jost, S.; Medinger, R.; Ottenwälder, B.; Boenigk, J.; Schlötterer, C. Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol. Ecol.* **2010**, *19*, 2908−2915.
(48) Stoeck, T.; Breiner, H.-W.; Filker, S.; Ostermaier, V.; Kammerlander, B.; Sonntag, B. A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology. *Environ. Microbiol.* **2014**, *16*, 430−444.
(49) Amend, A.; Seifert, K. A.; Bruns, T. D. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol. Ecol.* **2010**, *19*, 5555−5565.
(50) Pawlowski, J.; F. rowicz, F.; Esling, P. Next-generation environmental diversity surveys of Foraminifera: preparing the future. *Biol. Bull.* **2014**, *227*, 93−106.
(51) Weber, A.; Pawlowski, J. Can abundance of protists be inferred from sequence data? A case study of cultured Foraminifera. *PLoS One* **2013**, *8*, e56739.
(52) Prokopowich, C. D.; Gregory, T. R.; Crease, T. J. The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **2003**, *46*, 48−50.
(53) Heyse, G.; Jönsson, F.; Chang, W.-J.; Lipps, H. J. RNA-dependent control of gene amplification. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 22134−22139.
(54) Godhe, A.; Asplund, M. E.; Höm, K.; Saravanan, V.; Tyagi, A.; Karunasagar, I. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **2008**, *74*, 7174−7182.
(55) Medinger, R.; Nolte, V.; Pandey, R. V.; Jost, S.; Ottenwälder, B.; Schlötterer, C.; Boenigk, J. Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol. Ecol.* **2010**, *19*, 32−40.

DOI: 10.1021/es506158m
Environ. Sci. Technol. 2015, 49, 7597−7605