MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) doi: 10.1111/1755-0998.12668

Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring

LAURE APOTHÉLOZ-PERRET-GENTIL,* ARIELLE CORDONIER,† FRANÇOIS STRAUB,‡ JENNIFER ISELI,‡ PHILIPPE ESLING§ and JAN PAWLOWSKI*

*Department of Genetics and Evolution, University of Geneva, boulevard d'Yvoy 4, 1205, Geneva, Switzerland, †Water Ecology Service, Department of Territorial Management, Canton of Geneva, avenue de Sainte-Clotilde 23, 1211, Geneva, Switzerland, ‡PhycoEco, Rue des XXII-Cantons 39, 2300, La Chaux-de-Fonds, Switzerland, §IRCAM, UMR 9912, Université Pierre et Marie Curie, place Igor Stravinsky 1, 75004, Paris, France

Abstract

Current biodiversity assessment and biomonitoring are largely based on the morphological identification of selected bioindicator taxa. Recently, several attempts have been made to use eDNA metabarcoding as an alternative tool. However, until now, most applied metabarcoding studies have been based on the taxonomic assignment of sequences that provides reference to morphospecies ecology. Usually, only a small portion of metabarcoding data can be used due to a limited reference database and a lack of phylogenetic resolution. Here, we investigate the possibility to overcome these limitations using a taxonomy-free approach that allows the computing of a molecular index directly from eDNA data without any reference to morphotaxonomy. As a case study, we use the benthic diatoms index, commonly used for monitoring the biological quality of rivers and streams. We analysed 87 epilithic samples from Swiss rivers, the ecological status of which was established based on the microscopic identification of diatom species. We compared the diatom index derived from eDNA data obtained with or without taxonomic assignment. Our taxonomy-free approach yields promising results by providing a correct assessment for 77% of examined sites. The main advantage of this method is that almost 95% of OTUs could be used for index calculation, compared to 35% in the case of the taxonomic assignment approach. Its main limitations are under-sampling and the need to calibrate the index based on the microscopic assessment of diatoms communities. However, once calibrated, the taxonomy-free molecular index can be easily standardized and applied in routine biomonitoring, as a complementary tool allowing fast and cost-effective assessment of the biological quality of watercourses.

Keywords: bioindication, environmental DNA, metabarcoding, water quality

Received 28 July 2016; revision received 6 March 2017; accepted 7 March 2017

Introduction

Various biotic indices are widely used for the assessment of water quality. Traditionally, the indices are calculated based on the diversity of selected bioindicator taxa identified morphologically (Borja & Dauer 2008; Poikane *et al.* 2011). Recently, several attempts have been made to use eDNA data to infer the community structure of bio-indicator species (Baird & Hajibabaei 2012; Chariton *et al.* 2015). Several factors have been identified that may potentially impede the correct assignment of sequences to morphospecies and therefore the calculation of accurate indices. In particular, the incompleteness of the genetic database, the lack of resolution of phylogenetic markers and cryptic diversity (Yu *et al.* 2012; Carew *et al.* 2013; Eiler *et al.* 2013) have been highlighted as major issues. To overcome these limitations, we examine here whether it is possible to infer a molecular index directly from eDNA data without referring to the morphotaxonomy.

As a case study, we chose benthic diatoms, which are widely used as bioindicators of rivers and streams because of their high sensitivity to environmental changes and well-established taxon-specific ecological tolerances and preferences (Stevenson *et al.* 2010). In 2000, the European Union published a directive, the Water Framework Directive (Directive 2000/60/EC), that commits all member states to evaluate the status of their water bodies and to achieve a good status for them by a set deadline, recommending diatoms as one of the ideal bioindicators for river assessment. Different biotic indices are used across the different countries (Kelly

Correspondence: Pawlowski Jan, E-mail: jan.pawlowski@unige.ch

et al. 2008). In Switzerland, two biological indices are used to comply with the concomitant ecological objectives specified by the Swiss decree on water protection (Swiss Federal Council 1998), the IB-CH using macrozoobenthos and the DI-CH, using diatoms. The Swiss Diatom Index (DI-CH) is based on chemical parameters indicating anthropogenic pollution and classifies the water quality into five different ecological classes on a scale from 1 to 8 (1–3.5: very good; 3.5–4.5: good; 4.5–5.5: average; 5.5–6.5: bad; 6.5–8: very bad). The calculation follows the weighted average equation of Zelinka & Marvan (1961) and is defined as

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

This equation involves an autecological value D and a weighting factor G, which are specific to each species. It also uses an additional parameter H, which corresponds to the relative frequency of a particular taxon in the sample.

Like other diatom indices (Kelly *et al.* 2001; Coste *et al.* 2009), the DI-CH requires a morphologic determination to the species level. This requirement is a major weakness of the currently used system. Indeed, diatoms are a highly diverse group of protists and the identification of their tiny frustules requires special sample preparation, high-quality microscopes and in-depth taxonomic expertise. Inter-calibration exercises among specialists are organized to validate the robustness of the indices. These time-consuming limiting factors contrast with the need for the fast routine assessment of water quality required by Water Framework Directive and the Swiss Federal Office for the Environment.

The development of high-throughput sequencing (HTS) technologies applied to diversity surveys of microbial eukaryotes communities provided a possibility to overcome some of these limitations (Pawlowski *et al.* 2016). Several attempts have been made to use HTS eDNA metabarcoding as a tool for identifying diatom species either in mock communities (Kermarrec *et al.* 2013, 2014) or in environmental samples (Kermarrec *et al.* 2014; Zimmermann *et al.* 2014, 2015; Visco *et al.* 2015). Some authors attempted to infer diatom indices from metabarcoding data (Kermarrec *et al.* 2014; Visco *et al.* 2015; Keck *et al.* 2016). However, the results of these studies were not entirely satisfactory due to uncertainties concerning the correct assignment of sequences to morphospecies and various biases involved in qualitative and quantitative analyses of molecular data.

Here, we propose a taxonomy-free approach to calculate the Swiss Diatom Index values directly from sequence data. To test this new approach, we analyse 87 epilithic samples from Swiss rivers, mostly located in the Geneva basin, using the hypervariable region V4 of 18S rDNA as the diatom DNA barcode and the Illumina Miseq platform for sequencing. As illustrated in Fig. 1, we calculate the DI-CH values inferred from molecular data with two methods. First, by phylogenetic assignment of OTUs to morphospecies (DI-MOLTAXASSIGN – pathway 2), as previously described in Visco *et al.* 2015. Second, by assigning OTUs directly to ecological classes (DI-MOLTAXFREE – pathway 3). Finally, we compare those values with the ones derived from traditional microscopic studies (DI-CH – pathway 1).

Material and methods

Sampling

In total, 87 samples were collected during the 2013–2015 period in the Geneva and Neuchâtel cantons in Switzerland (Table S1, Fig. S1, Supporting information). This number includes 27 samples already published in (Visco *et al.* 2015). All the samples were collected as part of the monitoring program for water quality performed by the Service of Water Ecology (SECOE) of the Department of Environment, Transport and Agriculture of the Geneva canton and the Service of Energy and Environment of the Neuchâtel canton. The biofilm containing epilithic diatoms was collected following the directives established by the Swiss Federal Office for the Environment (Hürlimann & Niederhauser 2007). Each sample was divided into two subsamples for morphological and molecular analyses. Morphological samples were preserved with a final concentration of at least 4% of formaldehyde, while molecular samples were kept cold (ca. 0 °C) during sampling. In the laboratory, about 1 mL of each sample suspension was centrifuged and pellets were stored at −80 °C until further investigations.

Morphological analysis

The preparation of diatoms slides for microscopic observation was performed as recommended by the Swiss Federal Office for the Environment (Hürlimann & Niederhauser 2007). About 500 valves per sample were counted and identified mainly with the bibliographic support of The Flora of Diatoms (Krammer & Lange-Bertalot 1986-1992), Diatoms of Europe (Lange-Bertalot 2001) and Iconographia Diatomologica (Lange-Bertalot & Metzeltin 1996; Reichardt 1999), and Diatomeen im Süsswasser-Benthos von Mitteleuropa (Hofmann *et al.* 2011). In the case of the samples from Neuchâtel, after the 500 valves had been counted, the preparations were scanned for 20 min to

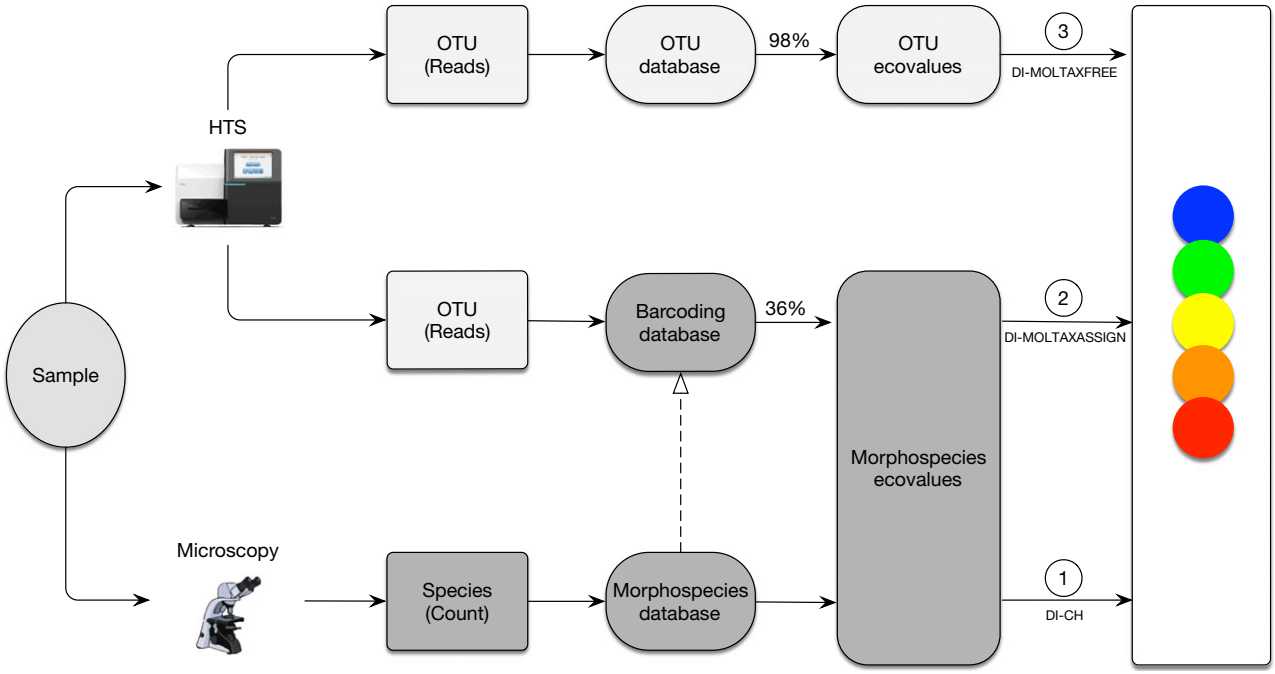

Fig. 1 Workflow illustrating the different methods used in this paper.

find rare species. Finally, the DI-CH values for each site were calculated following the equation described above.

Reference database

We chose the V4 region following the work of (Zimmermann *et al.* 2011) and our previous study (Visco *et al.* 2015). Although alternative diatom barcodes, such as rbcL, seem to offer better taxonomic resolution, we favour the V4 region because its amplification from eDNA samples is easier and its size better fits the sequencing length of the Illumina Miseq platform.

We built a reference database of the 18S V4 region of diatoms using online databases GenBank Release 212 and R-syst::diatom v5 (Rimet *et al.* 2016) and Sanger sequences from previous environmental studies in the Geneva basin (Visco *et al.* 2015). The region of interest was cut from downloaded sequences and aligned using the Seaview program (Gouy *et al.* 2010). The alignment was checked manually. Environmental sequences were screened using Uchime for chimeras (Edgar *et al.* 2011), which were then removed. The remaining sequences were analysed by Maximum Likelihood (ML) phylogenetic inference and those that did not branch in the clade corresponding to their morphological identification were discarded. After filtering, 1297 unique diatom sequences were kept, including 155 environmental sequences coming from the same geographic area as the study (Table S2, Supporting information).

Molecular analysis

DNA was extracted with the PowerBiofilm® DNA Isolation kit (MO BIO Laboratories Inc.) according to the manufacturer instructions. Three extraction replicates were performed for each sample. The hypervariable region V4 of the 18S rRNA gene of diatoms was then enriched by PCR amplification using specific diatom primers modified after (Zimmermann *et al.* 2011). Following previous studies, PCRs were performed as described in Visco *et al.* (2015), using unique combinations of forward and reverse primers tagged with individual tags composed of eight nucleotides attached at each primers 5′-extremities (Esling *et al.* 2015). A total of 20 different forward and reverse tagged primers were designed to enable multiplexing of all PCR products in a unique sequencing library. The sequences of tags and primers are provided in Table S3 (Supporting information).

Two PCR replicates were performed for each extraction and were then pooled for purification with High Pure PCR Cleanup Micro kit (Roche Diagnostics). In total, six PCR replicates were pooled for each sample. Purified PCR products were quantified with QuBit HS ds DNA kit (Invitrogen) and pooled in equimolar quantities. Two libraries were prepared (DIATOM03 for 2014 samples and DIATOM05 for 2015 samples, containing 24 and 36 samples, respectively) using Illumina TruSeq® DNA PCR-Free Library Preparation Kit following the manufacturer's instructions. The libraries were

then quantified with qPCR using KAPA Library Quantification Kit and sequenced on a MiSeq instrument using paired-end sequencing for 500 cycles with NANO KIT v2.

HTS data analysis

Quality filtering and assembly were performed according to the method described in Visco *et al.* 2015;. The two runs from our previous study and the two from this study were combined, and this complete data set was de-replicated; that is, the identical sequences were grouped together to obtain unique sequences, called Independent Sequence Units (ISUs). An abundance threshold of 10 was used for the minimum number of reads required for each ISU (Bokulich *et al.* 2013). We removed the ISUs that did not match any diatom sequences in the NCBI database with at least 99% coverage and 97% identity. ISUs were then grouped at 99% using complete-linkage clustering method. Finally, we removed chimeric sequences found with manual inspection of Uchime (Edgar *et al.* 2011) candidates.

Phylogenetic analyses

Taxonomic assignment of the operational taxonomic units (OTUs) was checked by phylogenetic analyses. The most abundant ISUs were used as the representative sequence for each OTU and were aligned to the reference database. The Maximum Likelihood (ML) phylogeny was constructed using RAxML v.7.2.8 (Stamatakis 2014) with GTR + G as model of evolution and 1000 replicates for the bootstrap analysis. The OTUs were then assigned to a morphospecies if they formed a clade supported by bootstrap values >60, following our previous study (Visco *et al.* 2015) and that of (Zimmermann *et al.* 2015). After the OTUs were assigned, DI-CHMOLTAXASSIGN scores were calculated based on the molecular data, using the D and G values given by the assigned species and the relative frequency of reads for the H factor.

Calculation of ecological values

To calculate the autecological value D and the weighting factor G for each OTU, we rely on an approach similar to that used to create the DI-CH index itself (Hürlimann & Niederhauser 2007). For the calibration, the reference status for each site was given by the DI-CH values. For the calculation, only the OTUs with a relative frequency >1% in at least one sample were kept. To find the autecological value D, the samples were grouped into 15 classes from 1 to 8 with a step of 0.5 according to their ecological status. For each OTU, the class with the highest 80th percentile of relative frequencies was then kept as the D value. For the weighting factor G, the samples were grouped into eight ecological classes. For each OTU, the distribution of 80% of its total abundance across the eight classes was used to determine the weighting factor, using the following thresholds. 8: OTUs present in classes 1–3 and 7–8, corresponding to extreme ecological status. 4: OTUs present in 1 class only. 2: OTUs present in 2 classes. 1: OTUs present in 3 classes. 0.5: abundant OTUs present in a minimum of 4 classes or representing at least 3% in 3 classes. The workflow for this computation is summarized in Fig. S2 (Supporting information). This calculation was first done with the complete data set to compare the values given by the species assigned with the ones inferred from the DI-MOLTAXFREE approach.

Inference of the molecular index and cross-validation

The molecular index was inferred from HTS data based either on those OTUs that could be assigned to morphospecies (DI-MOLTAXASSIGN) or all OTUs having a relative abundance of more than 1% in at least one sample of the data set (DI-MOLTAXFREE). In the second case, the ecological values D and G were calculated as described above, while the H values were equal to the relative number of sequences (reads) for each OTU.

To evaluate the status of the taxonomy-free index (DI-MOLTAXFREE), two cross-validation tests were performed. In each case, the D and G values were recalculated without the tested samples. First, we used a leave-one-out cross-validation. To do so, one sample was removed from the data set for the calculation of the value D and the factor G. Then, these D and G values were used to calculate the DI-MOLTAXFREE index of the removed sample. This process was repeated for each sample. Second, we performed a 25/75 cross-validation in which the D and G values were calculated for 65 sites and the evaluation of the index on the 22 remaining sample. The sites were randomly chosen, and the validation was repeated for 1000 trials. The formula used to calculate the DI-MOLTAXFREE was the same as for the calculation of the morphological DI-CH presented in the introduction.

Results

HTS data

The samples were sequenced in four independent Illumina runs. A total number of 2 206 456 good reads distributed across the 87 samples remained after filtering. The details for each run are described in Table S4 (Supporting information). The reads from all runs were dereplicated, resulting in 3079 ISUs. The ISUs were

clustered into 663 OTUs. After chimera removal, a final number of 440 OTUs was used for further analyses. The distribution of these OTUs and the number of reads per site are detailed in Table S5 (Supporting information). The number of OTUs per site varied from 1 (FOS) to 77 (VXB) with a median value of 27 (Table S6, Supporting information).

Morphological analysis

Morphospecies were counted, and the relative abundance of each taxon was calculated for each site (Table S7, Supporting information). A total of 269 morphospecies was identified across the 87 sites. The number of taxa per site varied from 5 (AMB) to 72 (PTH) with a median value of 24 (Table S6, Supporting information). The ecological status values ranged between 1.61 (VXD) and 7.98 (AMB). The different ecological classes (very good, good, average, bad and very bad) were represented by 15, 26, 25, 12 and 9 sites, respectively (Table S8, Supporting information). These DI-CH values were used as references for the molecular analysis.

Taxonomic assignment

We built a ML tree with our reference database and all OTUs (Fig. S3, Supporting information). After analysis, 152 OTUs (35%) were assigned to 43 morphospecies, of which 28 were found in the morphological analyses, while 15 matched to morphospecies not found microscopically in our samples. Figure S4 (Supporting information) shows the number of morphospecies recognized through morphological analysis, and in the genetic database and our HTS data set after phylogenetic assignment. Almost 70% of the morphospecies (185/269) found in the morphological counts were not represented in the database, leaving 84 morphospecies that were represented in the database. However, among these only 28 species were assigned in the molecular data set.

Ecological values comparison

In this section, we compare the D and G values provided by the morphological database with those inferred from molecular data (DI-MOLTAXFREE). To do so, we selected 78 of 152 taxonomically assigned OTUs that

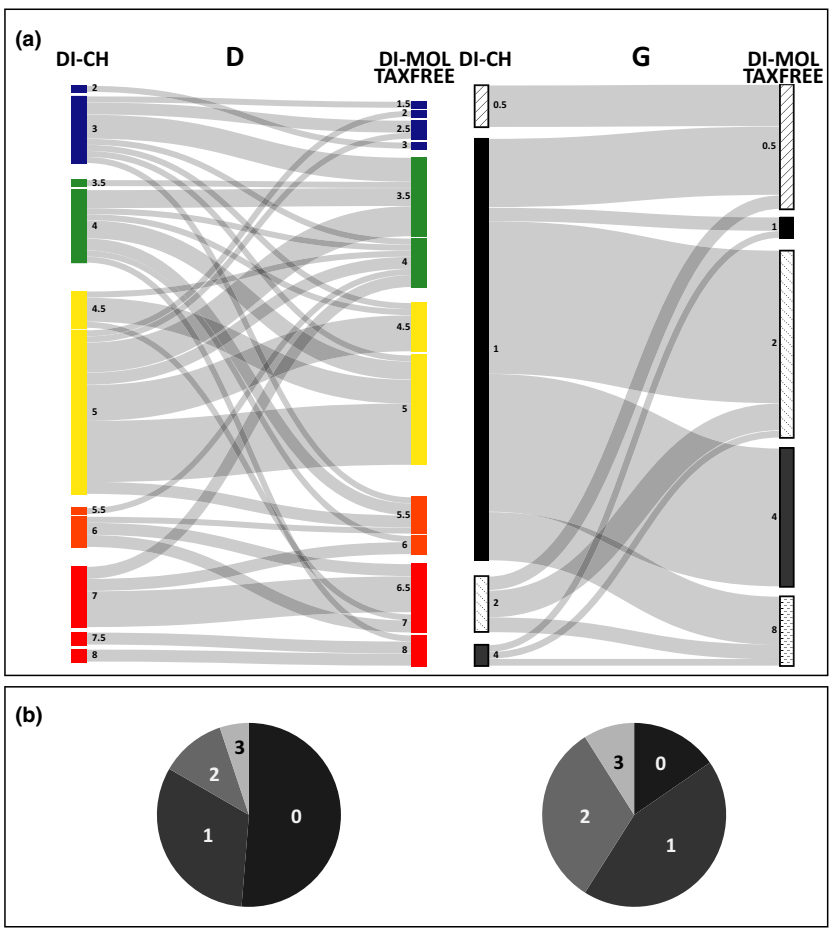

Fig. 2 Comparison of DG values for 78 assigned OTUs. The figure is separated into two parts: D values on the left and G values on the right. For each value, the sankey diagram (a) represents the relationship between the values inferred from morphology (DI_CH), and those inferred by the molecular index (DI-MOLTAXFREE). The links represent the assigned OTUs. Pie charts (b) represent the proportion of assigned OTUs as a function of the number of classes that change between their two values. No class changes are indicated in black, one class changes in dark grey, two classes change in medium grey and three classes change in light grey. For the D value, the class are separated as follows: 1–3.5: very good; 3.5–4.5: good; 4.5–5.5: average; 5.5–6.5: bad; 6.5–8: very bad and the scale of the G value is 0.5, 1, 2, 4 and 8.

could be given the D and G values of the related morphospecies and represented more than 1% of the total number of sequences in at least one sample of the data set. The selected OTUs were assigned to 23 different morphospecies. Their D and G values obtained from the morphotaxonomic database were compared to the values obtained by the taxonomy-free approach (Fig. 2).

More than half of the 78 OTUs show a morphological and a molecular D value indicating the same ecological status and 15% of the OTUs show exactly the same G values. These numbers increase to 83% and 59% with a maximum of one change for the D value and G value, respectively. For both values, <10% show a drastic change of three categories difference. The D and G values are given for each assigned OTUs in Table S9 (Supporting information).

Relative abundance

Besides the ecological values D and G, we also compared the relative abundance of each species based on microscopic counts of specimens found at a particular site to the relative abundance of the corresponding OTU represented by the number of HTS reads (sequences). In Fig. S5 (Supporting information), we provide the results of this comparison for the 23 assigned morphospecies. In the majority of cases, we observed that the relative abundance of sequences is higher compared to the abundance of specimens (circles are located above the triangles). However, in few cases (e.g. *Sellaphora seminulum*), the opposite is observed. We calculated the correlation between the morphological and the molecular abundance for the four most abundant species. As shown in Fig. S6 (Supporting information), three species (*Cocconeis placentula*, *Eolimna minima*, *Planothidium lanceolatum*) showed a strong correlation (R^2 = 0.79, 0.76 and 0.90, respectively, with *P*-values < 0.0001), whereas *Achnanthidium minutissimum* did not (R^2 = 0.41).

Diatom index

The molecular scores inferred using the taxonomic assignment (DI-MOLTAXASSIGN) and the taxonomy-free method (DI-MOLTAXFREE) were compared to examine the coverage of the HTS data set by each of those two approaches. The range of the values calculated by the DI-MOLTAXASSIGN was 3.00–7.98 compared with 2.7–6.93 for the DI-MOLTAXFREE method. As illustrated in Fig. 3, the taxonomic assignment method utilized 36% of the reads, whereas the taxonomy-free approach utilized 98% of the data set. Similar proportions were found in the number of OTUs, with 38% and 85% of OTUs included in the taxonomic assignment and taxonomy-free approaches, respectively. For only one site (HEP), the number of OTUs used in the taxonomic assignment method was greater than in the taxonomy-free approach. This particular site shows a huge genetic diversity in *Cocconeis placentula* (17 different OTUs), although six of them were very rare and therefore were removed from the taxonomy-free analysis.

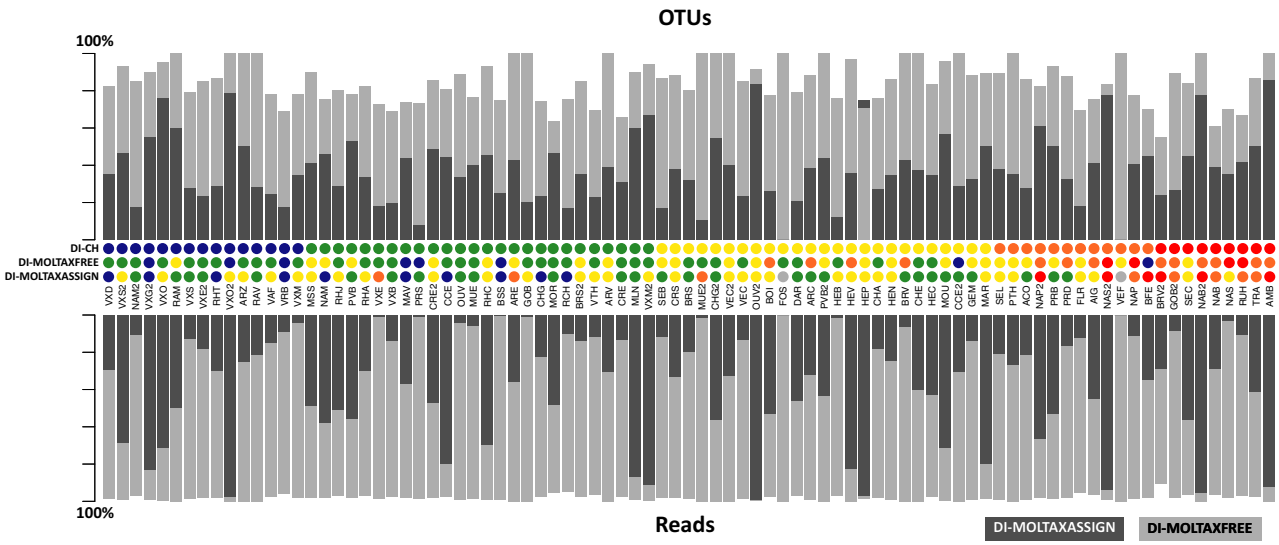

Fig. 3 Percentage of the HTS data set used by the taxonomic assignment (dark grey) and the molecular index (light grey) methods for each site. The OTUs are illustrated at the top and the reads at the bottom. In the middle, the coloured dots represent the ecological status given by the calculation of DI-CH values with Morphology (DI-CH), Molecular index (DI-MOLTAXFREE) or Taxonomic assignment (DI_MOLASSIGN). For the molecular index, the results of the leave-one-out cross-validation are used. The very good, good, average, bad and very bad statuses are represented with blue, green, yellow, orange and red colour, respectively.

The central part of Fig. 3 indicates the ecological status inferred by each approach. The two molecular methods (taxonomic assignment and taxonomy-free) give the same ecological status for 45% (38/85) of the samples; 14 of them are congruent with the morphological evaluation. For 38% (33/87) of the samples, the DI-MOLTAXFREE gave the same class as the DI-CH compared with 30% (26/85) for the DI-MOLTAXASSIGN. For two sites (FOS and VEF), no sequences could be assigned and, therefore, no taxonomic assignment evaluation was possible.

The taxonomic assignment and taxonomy-free molecular indices are compared further in Fig. 4, which shows the correlations of each index with the values of the morphological index (DI-CH) and indicates the difference compared to the values of DI-CH. The correlation between DI-MOLTAXASSIGN and the DI-CH (R^2 = 0.57 and P-value < 0.0001) is lower than the correlation between DI-MOLTAXFREE and the DI-CH (R^2 = 0.67 and P-value < 0.0001). The values of the indexes differ by <1 in 77% of the samples for the DI-MOLTAXFREE, compared to 52% for the DI-MOLTAXASSIGN. The proportion of sites correctly assessed with the DI-MOLTAXASSIGN increases to 88% for the most sampled sites belonging to the good and average classes, which are the best represented in our data set. The under-sampled classes show less good results, with 75%, 67% and 46% of correctly assessed sites for the bad, very bad and very good classes, respectively (Fig. S7, Supporting information).

In the case of DI-MOLTAXFREE, the leave-one-out cross-validation test was used to better illustrate the comparison with DI-MOLTAXASSIGN (Fig. 4). However, similar results were obtained using the 25/75 cross-validation tests (shown as stars in Fig. 4 and illustrated in Figs S8 and S9, Supporting information). The seven most problematic sites remain the same in the two cross-validation tests. In those cases, the difference compared to the DI-CH is >1.5 for the leave-one-out analysis and for the 25/75 cross-validation, <6% of the trials show a difference below 1. Four of these seven sites belong to the very good quality class.

Discussion

Overcoming the taxonomic assignment issue

The main objective of this study was to test whether the step of taxonomic assignment is necessary to calculate a molecular diatom index with eDNA data. Previous studies highlighted various biases introduced by this step but still kept it as an integral part of their analyses (Kermarrec *et al.* 2014; Visco *et al.* 2015; Zimmermann *et al.* 2015). The present study shows that the molecular index computed with (DI-MOLTAXASSIGN) or without (DI-

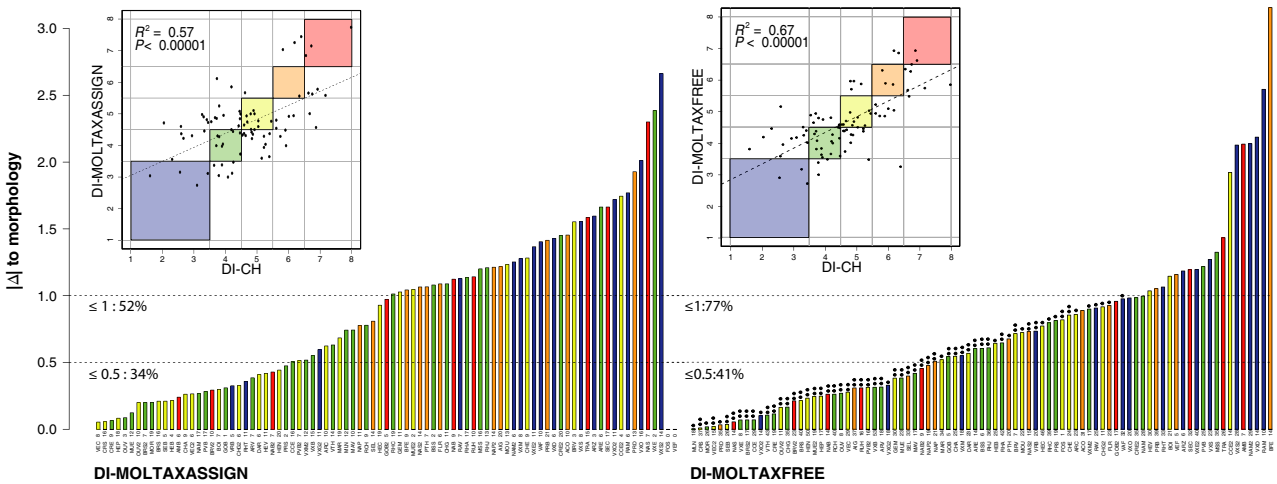

Fig. 4 Comparison between the DI-CH values given by morphology and molecular methods (taxonomic assignment on the left and leave-one-out cross-validation on the right). For each method, two types of graphics are represented. The scatter plots show the relationships between the DI-CH inferred from morphological (x-axis) and the molecular methods (y-axis). Coloured boxes represent the ecological status given by the DI-CH (blue: very good, green: good, yellow: average, orange: poor, red: bad). The regression line for all samples is represented by dashed line, and the R^2 and P-value are indicated for each graph. The bar plots show for each site the absolute difference between the DI-CH values given by the morphology and the molecular methods. Sites are coloured in function of their DI-CH value (blue: very good, green: good, yellow: average, orange: poor, red: bad). Above each site name, the number of OTUs used to calculate the index is indicated. Dashed lines are drawn at the 0.5 and 1 difference thresholds. Percentages of sites below these thresholds are indicated in the graphs. For each site, the black dots show the results of the 25/75 cross-validation. One dot is used if at least 70% of the replicates gave an absolute difference with the morphological DI-CH below 1 and two dots are used if this percentage is above 90%.

MOLTAXFREE) taxonomic assignment is not significantly different. Moreover, we observe a higher correlation between morphological and molecular indices in the case of the taxonomy-free approach (Fig. 4), suggesting that taxonomic assignment may not be essential for eDNA-based diatom monitoring.

Our results suggest that the main benefit of taxonomy-free approach lies in its much higher data coverage compared to the use of taxonomic assignment. The latter step considerably reduces the amount of available data due to the incompleteness of genetic reference databases, which comprise only 31% of the morphospecies identified in this study. This small number is reduced further to 10%, as 56 morphospecies present in genetic database (many belonging to the genus *Navicula*) could not be correctly assigned because of the lack of resolution of the 18S V4 marker. The selection of another marker (e.g. *rbc*L proposed by Kermarrec *et al.* 2013 and MacGillivary & Kaczmarska 2011) could probably improve the phylogenetic assignment for some species. However, it is uncertain whether the global data coverage would be much better.

Even if all morphospecies were sequenced with a more highly resolving marker, the taxonomic assignment will still be compromised by the issue of cryptic genetic diversity. It is well known that, in common with many other protists, the majority of diatom morphospecies are represented by large numbers of OTUs that are not always monophyletic (Amato *et al.* 2007; Beszteri *et al.* 2007; Rimet *et al.* 2014; Rovira *et al.* 2015; Van den Wyngaert *et al.* 2015). For example, *Cocconeis placentula* is represented in our data by 17 OTUs. Although this species complex has been split morphologically into several subspecies, their correspondence to numerous OTUs branching within the *C. placentula* clade is not well established. As a result, it is not possible to use different ecological values assigned to these subspecies and, conversely, to take advantage of ecological values assigned to *C. placentula* OTUs by the taxonomy-free approach. Regarding the practical application of the diatom index, the main problem with the taxonomic assignment approach is not so much the lack of correspondence between OTUs and morphospecies, but the difficulty of avoiding the errors introduced by the direct translation of ecological values associated with morphospecies to corresponding OTUs.

Accuracy of ecological values

By overcoming the step of taxonomic assignment, our method provides an independent assessment of ecological values. These values have been estimated directly from the HTS data, using morphological analyses as a reference to establish the ecological status of each site. As such estimations have never been attempted before, we examine the difference between these newly calculated values and those given by morphological observations. Although this comparison could only be performed on a few assigned OTUs and a limited number of sites, the results shown in Fig. 2 and Fig. S6 (Supporting information) are promising.

In the case of the autecological D values, the same ecological status was obtained for most of the OTUs. On the contrary, the variations of the weighting factor G are wider, with most of the OTUs having G values more or less equally distributed between 0.5 and 8, while most morphospecies are characterized by a G value of 1 (Fig. 2). As the G value reflects the occurrence of species/OTU across the sites, it is possible that these wider variations are related to the presence of extracellular DNA that can be dispersed over large distances (Deiner & Altermatt 2014). Alternatively, it is possible that the G values are affected by low amplification efficiency, which artificially reduces the range of occurrence, making the ecological tolerance of a given OTU appear narrower than in morphological surveys.

The accuracy of the DI-MOLTAXFREE also depends on the stability of D and G values during cross-validation. As illustrated in Fig. S10 (Supporting information), the values of the weighting factor G are relatively stable, with 83% of 228 analysed OTUs changing less than one category. In the case of D values, the variations are greater, although they rarely exceed two points. These large variations can be an effect of under-sampling, limiting the number of sites where an OTU occurs. This probably applies in the case of OTU 427, which is responsible for the highest difference between the DI-MOLTAXFREE and the DI-CH index found at the site BFE. Another possibility is that morphological misidentification leads to an erroneous assessment of some sites where an OTU is present. Such misidentifications can occur when the samples are processed routinely without a detailed scanning electron microscope examination of each specimen. To avoid such errors, it is necessary to stabilize the D and G values by increasing the number of sites and adapting the D and G values to the specificities of molecular data.

The issue of relative abundance

The third factor that influences the molecular index is relative abundance. This is also examined here. It is widely accepted that different technical and biological biases impact the relative abundance of specimens and sequences, making impossible the use of quantitative data in HTS surveys (Elbrecht & Leese 2015). However, this was not confirmed by the present study, at least as far as the most abundant species are concerned (Fig. S6, Supporting information). The same tendency was observed in other protists, such as foraminifera, where

the same species dominated morphological and molecular assemblages (Pawlowski *et al.* 2014). We could speculate that this relatively good match between the numbers of specimens and sequences of abundant species is reinforced by the exponential character of PCR amplification. As shown in the case of *C. placentula* and *E. minima* (Fig. S6, Supporting information), when a species is very abundant in microscopic counts, it is often even more abundant in HTS reads. However, this is not always true. For example, at some sites, the relative abundance of specimens of *Sellaphora seminulum* exceeds the abundance of reads (Fig. S7, Supporting information), suggesting that the PCR amplification may not be very efficient in this species.

In general, the importance of quantitative biases seems to be reduced in the case of small, single-cell organisms such as diatoms or foraminifera. However, the biomass of protistan cells can also vary considerably and the variability of rRNA copy numbers has been demonstrated in some diatoms (Alverson & Kolnick 2005; Godhe *et al.* 2008) and other protists (Gong *et al.* 2013; Weber & Pawlowski 2013). The taxonomy-free approach avoids this problem, because it does not involve the direct comparison of the relative abundance of specimens and sequences. Assuming that the PCR and other technical biases are the same across the samples for a given OTU (as long as the experimental conditions remain unchanged), the impact of these biases on the accuracy of taxonomy-free molecular index will be less important and easier to control than in the case of taxonomic assignment approach. Nevertheless, the formulae on which current indices are based are not adapted specifically for quantitative HTS data; a special effort will be required to address this issue in future studies.

Limitations of taxonomy-free approach

Although, as mentioned above, the taxonomy-free index has many advantages, it also has some important limitations that have to be overcome before the index can be used routinely. In view of our results, the most important factor causing incongruence between molecular and morphological indices is the lack of comprehensive sampling. As illustrated in Fig. 4, the DI-MOLTAXFREE approach considerably reduced the number of incorrectly assigned sites compared to the DI_MOLTAXASSIGN method. Yet, there are still sites that differ significantly from their status according to the morphological DI-CH method, and remarkably, most of them belonging to the under-sampled classes of very bad, bad and very good water quality.

The effect of under-sampling is particularly dramatic in the case of very good (blue) sites, half of which lie outside the 1-point limit (Fig. S5, Supporting information). This can be explained by the fact that these very good water quality sites are not characterized by specific indicator species but rather by different species-rich communities (Whitton *et al.* 1991; Hürlimann & Niederhauser 2007), which might be difficult to reconstruct without an extensive sampling. Conversely, the lack of congruence observed in the case of the bad and very bad quality sites can be explained the fact that these sites are usually characterized by high abundances of a few indicator species (Hill *et al.* 2001; Stevenso *et al.* 2010). When these sites appear rarely in the data set because of under-sampling, the absence of indicator species/OTUs in cross-validation studies may lead to the totally wrong assignment of a given site, as possibly happened in the case of sites AMB and BFE in our analyses (Fig. 4).

These few examples highlight the importance of sampling effort to ensure the accuracy of ecological values associated with OTUs in the taxonomy-free approach. However, even the most extensive eDNA sampling will not be able to alleviate all limitations of using OTUs rather than morphospecies to evaluate the quality of the environment. In particular, the metabarcoding data are unable to provide the kind of ecological information that is available through microscopic observations. For example, the list of OTUs and their relative frequencies says nothing about the physiological state of species, which can be measured by the proportion of teratological morphotypes in microscopic analyses (reviewed in Falasco *et al.* 2009). In general, the extensive knowledge of the taxonomy, biology and ecology of diatoms that can be derived from microscopic observations cannot be easily applied to the interpretation of molecular data. Therefore, the taxonomy-free index should be considered as a complementary tool rather than as a replacement for morphology-based studies.

Future challenges and perspectives

Our study raises several questions concerning the applicability of taxonomy-free approach in routine biomonitoring. Some of these questions, concerning the geographic range of OTUs and their ecological preferences, can hardly be answered without extensive sampling. Therefore, to further test the taxonomy-free index, the most important challenge is to obtain data from a much broader geographic area and from more diverse habitats. As shown by our results, the assessment of water quality is relatively good in the case of sites of average and good ecological status that dominate in our sampling. On the contrary, the diatom communities of the very good and very bad quality sites are not yet sufficiently represented in our data sets and, therefore, the inferred ecological values are not accurate enough. This highlights the importance of having not only numerous

sites but also sufficiently varied sampling habitats to cover the widest diversity possible.

Another important challenge is the calibration of the taxonomy-free index. In the present study, we relied on a well-established diatom index that is routinely used to characterize water quality in Swiss rivers and streams. The Swiss index and other diatom indices currently available are based on decades of microscopic data collection that has provided comprehensive information about diatom species ecology and distribution. These morphological data are essential to calibrate the taxonomy-free index and ensure its accuracy and robustness. However, where morpho-taxonomic data are not available due to a lack of taxonomic expertise, other types of data, such as chemical parameters or macro-invertebrate surveys, could serve as alternative calibration options. The most readily available data are chemical parameters. Yet, to be useful for diatom index calibration, the chemical analyses have to be conducted over longer periods of time. Depending on the diversity and geographic ranges of diatom OTUs, calibration of the taxonomy-free index would be necessary for different habitats and geographic localities. However, once the index is properly calibrated, the ecological values for each OTU will be more stable and the values of diatom index will be more reliable.

To conclude, our study demonstrates the great potential of the taxonomy-free molecular index for environmental biomonitoring. Although our work focuses on diatoms and the specific case of the Swiss diatom index, the taxonomy-free approach could easily be applied to other groups of single-cell bioindicators, such as ciliates (Lee *et al.* 2004; Chen *et al.* 2008; Jiang *et al.* 2011), and foraminifera (Schönfeld *et al.* 2012; Vidovic *et al.* 2014; Alve *et al.* 2016). New molecular indices could also be tested for microbial and meiofaunal taxa that are not currently used as bioindicators. The implementation of these new indices would help to extend the range of monitored sites and increase the frequency of monitoring. Once established, molecular indices could provide a fast, easily standardized and highly sensitive tool that complements the current morphology-based methods available for the water quality assessment.

Acknowledgements

We thank Francois Pasquini from Water Ecology Service of the Canton of Geneva and Isabelle Butty from the division of water and soil of the Canton of Neuchâtel for their permission to use the samples and the morphological data. We also than Andrew Gooday from National Oceanographi Centre, Southampton for very careful editing of the manuscript and discussion. Financial support was provided by the Swiss National Science Foundation (grants 316030_150817 and 31003A-140766) and G & L Claraz Donation. This study is a part of the SwissBOL program supported by the Swiss Federal Office for the Environment.

References

Alve E, Korsun S, Schönfeld J *et al.* (2016) Foram-AMBI: a sensitivity index based on benthic foraminiferal faunas from North-East Atlantic and Arctic fjords, continental shelves and slopes. *Marine Micropaleontology*, **122**, 1–12.

Alverson AJ, Kolnick L (2005) Intragenomic nucleotide polymorphism among small subunit (18s) Rdna paralogs in the diatom genus skeletonema (bacillariophyta)1. *Journal of Phycology*, **41**, 1248–1257.

Amato A, Kooistra WHCF, Ghiron JHL *et al.* (2007) Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist*, **158**, 193–207.

Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Beszteri B, John U, Medlin LK (2007) An assessment of cryptic genetic diversity within the *Cyclotella meneghiniana* species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphisms. *European Journal of Phycology*, **42**, 47–60.

Bokulich NA, Subramanian S, Faith J *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature methods*, **10**, 57–59.

Borja A, Dauer DM (2008) Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. *Ecological Indicators*, **8**, 331–337.

Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA (2013) Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, **10**, 45.

Chariton AA, Stephenson S, Morgan MJ *et al.* (2015) Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental Pollution (Barking, Essex: 1987)*, **203**, 165–174.

Chen Q-H, Xu R-L, Tam NFY, Cheung SG, Shin PKS (2008) Use of ciliates (Protozoa: Ciliophora) as bioindicator to assess sediment quality of two constructed mangrove sewage treatment belts in Southern China. *Marine Pollution Bulletin*, **57**, 689–694.

Coste M, Boutry S, Tison-Rosebery J, Delmas F (2009) Improvements of the Biological Diatom Index (BDI): description and efficiency of the new version (BDI-2006). *Ecological Indicators*, **9**, 621–650.

Deiner K, Altermatt F (2014) Transport distance of invertebrate environmental DNA in a natural river. *PLoS ONE*, **9**, e88786.

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**, 1–73.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Eiler A, Drakare S, Bertilsson S *et al.* (2013) Unveiling distribution patterns of freshwater phytoplankton by a next generation sequencing based approach. *PLoS ONE*, **8**, e53516.

Elbrecht V, Leese F (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.

Esling P, Lejzerowicz F, Pawlowski J (2015) Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, **43**, 2513–2524.

Falasco E, Bona F, Badino G, Hoffmann L, Ector L (2009) Diatom teratological forms and environmental alterations: a review. *Hydrobiologia*, **623**, 1–35.

Godhe A, Asplund ME, Härnström K *et al.* (2008) Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples

by real-time PCR. *Applied and Environmental Microbiology*, **74**, 7174–7182.
Gong J, Dong J, Liu X, Massana R (2013) Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist*, **164**, 369–379.
Gouy M, Guindon S, Gascuel O (2010) SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27**, 221–224.
Hill BH, Stevenson RJ, Pan Y *et al.* (2001) Comparison of correlations between environmental characteristics and stream diatom assemblages characterized at genus and species levels. *Journal of the North American Benthological Society*, **20**, 299–310.
Hofmann G, Werum M, Lange-Bertalot H (2011) k ; über 700 der häufigsten Arten und ihre Ökologie. Gantner.
Hürlimann J, Niederhauser P (2007) 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.
Jiang Y, Xu H, Hu X *et al.* (2011) An approach to analyzing spatial patterns of planktonic ciliate communities for monitoring water quality in Jiaozhou Bay, northern China. *Marine Pollution Bulletin*, **62**, 227–235.
Keck F, Bouchez A, Franc A, Rimet F (2016) Linking phylogenetic similarity and pollution sensitivity to develop ecological assessment methods: a test with river diatoms. *Journal of Applied Ecology*, **53**, 856–864.
Kelly MG, Adams C, Graves AC (2001) *The Trophic Diatom Index: a User's Manual*; Revised Edition. Environment Agency, Rotherham.
Kelly M, Bennett C, Coste M *et al.* (2008) A comparison of national approaches to setting ecological status boundaries in phytobenthos assessment for the European Water Framework Directive: results of an intercalibration exercise. *Hydrobiologia*, **621**, 169–182.
Kermarrec L, Franc A, Rimet F *et al.* (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, **13**, 607–619.
Kermarrec L, Franc A, Rimet F *et al.* (2014) A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, **33**, 349–363.
Krammer K, Lange-Bertalot H (1986-1992) *Bacillariophyceae. Süsswasserflora von Mitteleuropa*, Stuttgart Germany.
Lange-Bertalot H (2001) *Diatoms of the European Inland Waters and Comparable Habitats*. Ruggell, Lichtenstein.
Lange-Bertalot H, Metzeltin D (1996) Indicators of oligotrophy: 800 taxa representative of three ecologically distinct lake types : carbonate buffered, oligodystrophic, weakly buffered soft water. *Iconographia Diatomologica*, **2**, 390.
Lee S, Basu S, Tyler CW, Wei IW (2004) Ciliate populations as bio-indicators at Deer Island Treatment Plant. *Advances in Environmental Research*, **8**, 371–378.
MacGillivary ML, Kaczmarska I (2011) Survey of the efficacy of a short fragment of the rbcL gene as a supplemental DNA barcode for diatoms. *The Journal of Eukaryotic Microbiology*, **58**, 529–536.
Pawlowski J, Esling P, Lejzerowicz F, Cedhagen T, Wilding TA (2014) Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources*, **14**, 1129–1140.
Pawlowski J, Lejzerowicz F, Apotheloz-Perret-Gentil L, Visco J, Esling P (2016) Protist metabarcoding and environmental biomonitoring: time for change. *European Journal of Protistology*, **55**, 12–25.
Poikane S, van den Berg M, Hellsten S *et al.* (2011) Lake ecological assessment systems and intercalibration for the European Water Framework Directive: aims, achievements and further challenges. *Procedia Environmental Sciences*, **9**, 153–168.
Reichardt E (1999) Zur Revision der Gattung Gomphonema: Die Arten um G. affine/insigne, G. angustatum/micropus, G. acuminatum sowie gomphonemoide Diatomeen aus dem Oberoligozän in Böhmen. *Iconographia Diatomologica*, **8**, 250.
Rimet F, Trobajo R, Mann DG *et al.* (2014) When is sampling complete? the effects of geographical range and marker choice on perceived diversity in *Nitzschia palea* (Bacillariophyta). *Protist*, **165**, 245–259.
Rimet F, Chaumeil P, Keck F *et al.* (2016) R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database: The Journal of Biological Databases and Curation*, **2016**, baw016.
Rovira L, Trobajo R, Sato S, Ibáñez C, Mann DG (2015) Genetic and physiological diversity in the diatom *Nitzschia inconspicua*. *The Journal of Eukaryotic Microbiology*, **62**, 815–832.
Schönfeld J, Alve E, Geslin E *et al.* (2012) The FOBIMO (FOraminiferal BIo-MOnitoring) initiative—Towards a standardised protocol for soft-bottom benthic foraminiferal monitoring studies. *Marine Micropaleontology*, **94–95**, 1–13.
Stamatakis A (2014) RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
Stevenson RJ, Pan YD, van Dam H (2010) Assessing environmental conditions in rivers and streams with diatoms, In: *The Diatoms: Applications for the Environmental and Earth Sciences*, 2nd edn, pp. 57–85. Cambridge University Press, Cambridge, UK.
Swiss Federal Council (1998) *Waters Protection Ordinance*, https://www.admin.ch/opc/en/classified-compilation/19983281/index.html.
Van den Wyngaert S, Möst M, Freimann R, Ibelings BW, Spaak P (2015) Hidden diversity in the freshwater planktonic diatom *Asterionella formosa*. *Molecular Ecology*, **24**, 2955–2972.
Vidovic J, Dolenec M, Dolenec T, Karamarko V, Žvab Rožič P (2014) Benthic foraminifera assemblages as elemental pollution bioindicator in marine sediments around fish farm (Vrgada Island, Central Adriatic, Croatia). *Marine Pollution Bulletin*, **83**, 198–213.
Visco J, Apothéloz-Perret-Gentil L, Cordonier A *et al.* (2015) Environmental monitoring: inferring the diatom index from next-generation sequencing data. *Environmental Science & Technology*, **49**, 7597–7605.
Weber AA-T, Pawlowski J (2013) Can abundance of protists be inferred from sequence data: a case study of foraminifera (P López-García, Ed,). *PLoS ONE*, **8**, e56739.
Whitton BA, Rott E, Friedrich G (1991) Use of algae for monitoring rivers. *Journal of Applied Phycology*, **3**, 287–288.
Yu DW, Ji Y, Emerson BC *et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.
Zelinka M, Marvan P (1961) Zur Präzisierung der biologischen Klassifikation der Reinheit fliessender Gewässer. *Archiv fur Hydrobiologie*, **57**, 389–407.
Zimmermann J, Jahn R, Gemeinholzer B (2011) Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organisms Diversity & Evolution*, **11**, 173–192.
Zimmermann J, Abarca N, Enke N *et al.* (2014) Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. *PLoS ONE*, **9**, e108793.
Zimmermann J, Glöckner G, Jahn R *et al.* (2015) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, **15**, 526–542.

J.P., L.A.P.G. and P.E. conceived and designed the experiments. L.A.P.G. performed the experiments and analysed the data. A.C., F.S. and J.I. performed all the morphological work. J.P. and L.A.P.G. wrote the manuscript.

Data accessibility

Illumina raw data are deposited in Dryad (doi:10.5061/dryad.8m3kv).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Map of sampling sites.

Fig. S2 A. Schematic representation of the calculation of the D and G value for the molecular method.

Fig. S3 RAxML tree with sequences from the database and the OTUs from the HTS analysis.

Fig. S4 Venn diagram of morphospecies represented in the database (yellow), morphological analysis (blue) and found in HTS dataset by taxonomic assignment method (red).

Fig. S5 Scatter plot of the relative frequency for all the assigned species.

Fig. S6 Scatter plot of the relative frequency for the 4 most represented morphospecies in the HTS dataset.

Fig. S7 Graphical table representing the DI-MOLTAXFREE cross-validation results.

Fig. S8 Box plot of the DI-MOLTAXFREE Cross-Validation 25:75 test.

Fig. S9 Graphical representation of the DI-MOLTAXFREE Cross-Validation 25:75 test.

Fig. S10 Bar plots representing the proportion of D (blue) and G (green) values in function of their change during the cross-validation test.

Table S1 Illumina run code, station code, location, sampling date and geographic references for each site used in this study.

Table S2 List of database entries description with their NCBI or Rsyst accession number. Environmental sequences (ENV) are marked.

Table S3 List of primers and tags used in this study.

Table S4 Filtering process of the four Illumina runs used in this study.

Table S5 List of OTUs with their number of reads per sample.

Table S6 Number of OTUs from HTS analysis and species from morphological analysis for each site.

Table S7 List of species found during the morphological analysis with their relative abundance per site.

Table S8 DI-CH values given by morphology (DI-CH), taxonomic assignment (DI-MOLTAXASSIGN) and Leave-one-out cross-validation (DIMOLTAXFREE) for each site.

Table S9 Comparison of DG values given by morphology (D and G) and molecular (MOL-D and MOL-G) indices for each assigned OTUs.