Received: 27 August 2018 | Revised: 23 September 2018 | Accepted: 24 September 2018

DOI: 10.1002/ece3.4701

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

The impact of OTU sequence similarity threshold on diatom-based bioassessment: A case study of the rivers of Mayotte (France, Indian Ocean)

Kálmán Tapolczai[1,2] | Valentin Vasselon[2] | Agnès Bouchez[2] | Csilla Stenger-Kovács[3] | Judit Padisák[1,3] | Frédéric Rimet[2]

[1]MTA-PE Limnoecology Research Group, Veszprém, Hungary

[2]UMR CARRTEL, INRA, Thonon-les-Bains, France

[3]Department of Limnology, University of Pannonia, Veszprém, Hungary

Correspondence
Kálmán Tapolczai, MTA-PE Limnoecology Research Group, Veszprém, Hungary.
Email: tapolczai.kalman@gmail.com

Funding information
French National Agency for Water and Aquatic Environments (ONEMA - AFB).

Abstract

Extensive studies on the taxonomic resolution required for bioassessment purposes have determined that resolution above species level (genus, family) is sufficient for their use as indicators of relevant environmental pressures. The high-throughput sequencing (HTS) and meta-barcoding methods now used for bioassessment traditionally employ an arbitrary sequence similarity threshold (SST) around 95% or 97% to cluster sequences into operational taxonomic units, which is considered descriptive of species-level resolution. In this study, we analyzed the effect of the SST on the resulting diatom-based ecological quality index, which is based on OTU abundance distribution along a defined environmental gradient, ideally avoiding taxonomic assignments that could result in high rates of unclassified OTUs and biased final values. A total of 90 biofilm samples were collected in 2014 and 2015 from 51 stream sites on Mayotte Island in parallel with measures of relevant physical and chemical parameters. HTS sequencing was performed on the biofilms using the *rbcL* region as the genetic marker and diatom-specific primers. Hierarchical clustering was used to group sequences into OTUs using 20 experimental SST levels (80%–99%). An OTU-based quality index (Idx_{OTU}) was developed based on a weighted average equation using the abundance profiles of the OTUs. The developed Idx_{OTU} revealed significant correlations between the Idx_{OTU} values and the reference pressure gradient, which reached maximal performance using an SST of 90% (well above species level delimitation). We observed an interesting and important trade-off with the power to discriminate between sampling sites and index stability that will greatly inform future applications of the index. Taken together, the results from this study detail a thoroughly optimized and validated approach to generating robust, reproducible, and complete indexes that will greatly facilitate effective and efficient environmental monitoring.

KEYWORDS

Diatoms, high-throughput sequencing, OTU, pollution assessment, sequence similarity threshold, taxonomic resolution, Water Framework Directive

1 | INTRODUCTION

Benthic diatoms are widely used as ecological indicators for bodies of water due to their short generation time, large diversity, high and sensitivity to environmental changes (Mann & Vanormelingen, 2013). They serve as a proxy for the entire phytobenthos (Kelly, King, Jones, Barker, & Jamieson, 2008), which is one of the five biological quality elements (BQEs) required by the European Water Framework Directive (WFD) for the assessment of the ecological quality of bodies of water (European Commission, 2000).

Diatom-based quality indices are generally calculated using a weighted average equation (Zelinka & Marvan, 1961) based on the species' ecological optimum and tolerance, as defined by its abundance profile along a pollution gradient. Each index has its own reference database that contains the ecological values (optimum and tolerance) of a set of species. Diatom studies are largely dedicated to rigorously characterizing specimens down to the finest taxonomic level possible (species, subspecies), even though it is often challenging and not necessary for bioassessment purposes (Lavoie, Dillon, & Campeau, 2009; Rimet & Bouchez, 2012). Microscopy-based identification of diatoms is based on their morphological attributes and thus carries several drawbacks. The process is time-consuming and requires experienced analysts. Furthermore, misidentifications are common and cause discrepancies in the species inventories of different laboratories, which must be regularly rectified (Kahlert et al., 2009, 2012). Moreover, different indices may have different ecological values for the same species because their profiles were defined from different ecoregions with limited range of environmental variables (Besse-Lototskaya, Verdonschot, Coste, & Vijver, 2011).

DNA barcoding has enabled the rapid development of novel molecular techniques that have greatly improved the quality of species identification (Hebert, Cywinska, Ball, & deWaard, 2003). These approaches employ standard markers to identify taxa-specific sequences in the DNA of the organisms in question to serve as that organism's barcode. These DNA-based methods are efficient and reduce misidentifications due to phenotypic plasticity (Leliaert et al., 2014) or cryptic diversity (Kaczmarska, Mather, Luddington, Muise, & Ehrman, 2014). High-throughput sequencing (HTS) technology, in combination with the aforementioned meta-barcoding, allows for simultaneously identifying multiple taxa from multiple environmental samples (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). This makes the routine analysis of environmental samples faster, more cost-effective, and accurate than traditional microscopy-based methods and provides much information than ever before. This facilitates expanding the sampling network to include more sites monitored on a more frequent basis and has thereby revolutionized the field of biomonitoring (Baird & Hajibabaei, 2012; Keck, Vasselon, Tapolczai, Rimet, & Bouchez, 2017). The incorporation of molecular techniques in biomonitoring has caused remarkable progress over the past decade in terms of optimal genetic marker selection (Kermarrec et al., 2013), HTS platform (Loman et al., 2012; Shokralla, Spall, Gibson, & Hajibabaei, 2012), DNA extraction (Vasselon, Domaizon, Rimet, Kahlert, & Bouchez, 2017), and the bioinformatics required to analyze the HTS data.

Sequence data obtained from the HTS platform are subjected to a quality-filtering process and then typically clustered into operational taxonomic units (OTU). Three main approaches have been developed to effectively cluster sequences into OTUs (Westcott & Schloss, 2015). And which algorithm to apply depends on many factors, including the target taxa, the genetic markers, and the sequencing method (Flynn, Brown, Chain, MacIsaac, & Cristescu, 2015). The closed-reference clustering method compares sequences to a reference database and then clusters into OTUs based on similarity to the reference sequence. The most commonly used clustering approach is de novo clustering. Here, sequences are clustered into OTUs before taxonomic assignment. Hierarchical clustering is a form of de novo clustering that creates a distance matrix to compute sequence dissimilarity between all sequence pairs before generating the OTUs. While this method is widely used, it requires high computational capacity (Sun et al., 2012). Greedy heuristic clustering is a more computationally effective approach because it does not compare all of the sequence pairs but, instead, analyzes one input sequence at a time. If the distance between that sequence and an already existing OTU is smaller than the predefined threshold, the sequence is assigned to the existing OTU. If not, it serves as the seed sequence for a new OTU (Sun et al., 2012). Both the hierarchical and the greedy heuristic clustering methods use a defined yet arbitrary clustering threshold, called the sequence similarity threshold (SST), as a cutoff value to ensure that the sequences within an OTU are identical (Patin, Kunin, Lidström, & Ashby, 2013). The third approach is termed open-reference clustering and involves closed-reference clustering followed by de novo clustering. Thereby, this approach essentially combines the strengths of the two aforementioned methods (Westcott & Schloss, 2015).

While SST values can reach up to 99% (Apothéloz-Perret-Gentil et al., 2017), most range between 95% and 97% (Edgar, 2013; Elbrecht & Leese, 2015; Kelly et al., 2018; Patin et al., 2013), which is thought to effectively maximize genetic diversity while also minimizing the frequency of sequencing errors in the resulting HTS-based dataset (Birtel, Walser, Pichon, Bürgmann, & Matthews, 2015; Schloss & Handelsman, 2005). These thresholds are treated as quasispecies level delimitations regardless of the specific marker, clustering method, or model organism used, even though these parameters can greatly affect the final OTU composition (Flynn et al., 2015).

A taxonomic name is then assigned to each newly generated OTU by comparing a representative sequence, generally the most abundant (Patin et al., 2013), to reference barcodes available in public databases (Rimet et al., 2016). Most studies use a consensus confidence threshold (Schloss et al., 2009) to delineate the abundance of the representative sequences required within an OTU and those that fall below this threshold are labeled as "unclassified OTUs" (Rivera, Vasselon, Jacquet, et al., 2018b; Visco et al., 2015). To generate complete reference libraries, these unclassified OTUs must

FIGURE 1 One of the sampling site in Dapani River, Mayotte. Good quality sites are typically characterized by dense riparian vegetation, natural river bank, and low turbidity. Since the sampling was carried out during the dry season, the environmental conditions are stable and the water level is low

be resolved (Groendahl, Kahlert, & Fink, 2017; Vasselon, Rimet, Tapolczai, & Bouchez, 2017). This challenge represents a considerable and pressing issue because a portion of the taxonomic diversity of the site remains unknown. As such, the quality index calculation will only be based on the ecological values from a portion of the species or genera while others go unidentified, among which may be dominant, relevant species (Rivera, Vasselon, Jacquet, et al., 2018b). An alternative approach was proposed (Apothéloz-Perret-Gentil et al., 2017) that avoids the taxonomic assignment of OTUs by using the ecological values of the OTUs directly.

The aim of this study was to determine the impact of taxonomic resolution on a quality index using molecular data. Toward this end, we carried out DNA-meta-barcoding on environmental biofilm samples collected from streams on the main island of the Department of Mayotte, a French archipelago in the Indian Ocean. We then investigate the impact of the SST on the OTU quality index (Idx_{OTU}) using the abundance profiles (ecological profiles) of the different OTUs that result from a gradient of different SSTs (80%–99%).

We hypothesize that the SST serves as a proxy for taxonomic resolution whereby OTUs of high SSTs represent fine taxonomic characterization (e.g., species, populations) and OTUs of low SSTs represent coarser taxonomic classification (e.g., genera, families). Our approach is similar to studies that analyzed the effect of taxonomic resolution on classical diatom quality indices (Lavoie et al., 2009; Rimet & Bouchez, 2012).

Additionally, we hypothesize that at low SSTs, the ecological profile of an OTU is the result of merging good indicator sequences and thus results in a low-performance quality index according to its capacity to separate "high-" and "poor-"quality samples from each other. On the other hand, at high SSTs, the ecological profiles of the more rare OTUs are based on fewer data points. Thus, they are more sensitive to the outliers that do not fit in the model provided by the ecological profile. We, therefore, hypothesize that after an "optimal" point, increasing the SST results in a less stable index that does not confer additional benefit in terms of quality evaluation. We suggest that an optimal SST, analogous to the optimal taxonomic resolution in previous studies on microscopy-based approaches (Rimet & Bouchez, 2012), can be identified that maximizes the index's performance and stability.

The effect of the SST on Idx_{OTU} was analyzed from several aspects: (i) the number of OTUs defined at each SST, (ii) the proportion of OTUs identified at the species and genus level after taxonomic assignment (this aspect was studied but not included in the index development), (iii) the performance of the quality index using three different indicators: (iii-a) the correlation between quality values and the reference environmental gradient, (iii-b) the index's capability to discriminate between sites with different quality values, and (iii-c) the variability in the index values conferred by the process used for index development (stability).

2 | MATERIALS AND METHODS

2.1 | Study site and sampling network

The French overseas Department of Mayotte is an island in the Comoros archipelago located in the Indian Ocean, northwest of Madagascar (12°50′35″S 45°08′18″E; Appendix S1). Following the change to its legal status in 2011, the implementation of the Water Framework Directive (WFD) became obligatory for its bodies of water (Figure 1). Toward this end, a surveillance monitoring network (RCS) was set up in 2008. This network was complemented with a "reference" (REF) network in 2013 and a "polluted" (POLL) network in 2014 to enlarge the environmental gradient. The classification of sites into these networks was predefined and based, in general, on visible conditions of the area (Tapolczai, Bouchez, Stenger-Kovács, Padisák, & Rimet, 2017). For the purpose of this study, a total of 90 samplings were collected from the three monitoring networks: 30, 23, and 37 samples from the RCS, POLL, and REF networks, respectively. These were collected from 51 river sites in 2014 and 2015, in parallel with the physical and chemical data associated with each site (Appendix S2).

2.2 | Phytobenthos sampling, physical, and chemical parameters

The phytobenthos sampling procedure followed both French and European standards (Afnor, 2014, 2016) and was carried out during the dry season (July–August) when flow conditions are more stable compared those the rainy season, which are affected by heavy flooding. The samples were collected using clean toothbrushes to remove the biofilm from the surface of at least five stones from lotic regions. These were then preserved by adding sufficient 99% ethanol to ensure a final concentration of over 70%. The sampling and analysis of the physical and chemical parameters were carried out

during the same time period according to APHA standards (APHA, 2012).

2.3 | HTS procedure

Total DNA was extracted from 2 ml of each phytobenthos samples using the GenEluteTM-LPA method. A detailed protocol can be found in previous publications (Chonova et al., 2016; Kermarrec et al., 2013). This method is preferred for diatom meta-barcoding (Vasselon, Domaizon, et al., 2017) because it uses multiple lysis mechanisms (mechanical, enzymatic, and heat-based) that when combined greatly increase the efficiency of diatom cell lysis and DNA yield.

A short 312-bp segment of the *rbcL* gene was used as the DNA marker and amplified by PCR using an equimolar mix of a modified version of the Diat_rbcL_708F forward and the R3 reverse primers (Rimet, Abarca, et al., 2018a; Vasselon, Rimet, et al., 2017). Each DNA sample was amplified in triplicate using 1 µl of extracted DNA in a final reaction volume of 25 µl. Detailed information on the PCR mixture and amplification conditions is summarized in Appendix S3.

The three PCR replicates of each DNA sample were pooled and purified using Agencourt AMPure beads (Beckman Coulter, Brea, CA, USA). The quality and quantity of the purified amplicons were checked using the 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA). Following the library preparation method described by Vasselon, Domaizon, et al. (2017), individual A-X tag adapters (Ion ExpressTM Barcode Adapters, Life Technologies, Carlsbad, CA, USA) were ligated to the amplicons using the NEBNext® Fast DNA Library Prep Set for Ion TorrentTM (BioLabs, Ipswich, MA, USA). The sample libraries were pooled into two mixes corresponding to the Mayotte 2014 and 2015 sampling campaigns that contained 49 and 41 samples, respectively. Each mix was adjusted to a final concentration of 100 pm and sequenced independently on two Ion 318TMChip Kit v2 (Life Technologies, Carlsbad, USA) using the PGM Ion Torrent machine.

The sequencing was performed by the "Plateforme Génome Transcriptome" (PGTB, Bordeaux, France) who provided one fastq file per sample for the 90 libraries with demultiplexed DNA reads. A quality-filtering step excluded DNA reads under 250 bp with a Phred quality score below 23 over a moving window of 25 bp, more than one mismatch in the primer sequence, a homopolymer over 8 bp, or an ambiguous base. The 90 trimmed files were merged in order to manipulate all of the samples concurrently using the bioinformatics processes described in Vasselon, Rimet, et al. (2017) using the Mothur software (Schloss et al., 2009). In addition to bioinformatics, the DNA reads were dereplicated to obtain individual sequence units (ISUs). The abundance of ISUs, corresponding to the number of sequence replicates per ISU, was used to remove ISUs with only one sequence. Retained ISUs were then clustered into OTUs using different SSTs ranging from 80% to 99%. Finally, 20 OTU lists, corresponding to each threshold and including the number of DNA reads within the 90 samples, were produced. Based on the taxonomy assigned to each ISU with the classify.seq command (RDP classifier with bootstrap cutoff = 85%, Wang, Garrity, Tiedje, & Cole, 2007) and the R-syst::diatom library (Rimet et al., 2016, 13–02–2015: R-Syst::diatom v3, https://www.rsyst.inra.fr/en), a consensus taxonomy was provided to each OUT using the classify.otu command with a confidence threshold of 80%. The Supplementary Data contains the following: the Fastq files with the demultiplexed DNA reads (Appendix S4); information on the sequencing depth before and after trimming (Appendix S5); the final OTU summary, including the proportion of DNA reads, representative DNA sequences for each OTU, and their taxonomic assignments (Appendix S6); descriptions of the sampling sites (Appendix S7); and the script run in Mothur from trimming to obtain the used OTU lists (Appendix S8).

2.4 | The development and testing of Idx_{OTU}

Sequence reads were transformed into relative abundances in order to normalize the OTU database. Although this is not the ideal approach toward achieving comparable quantification between samples, it is one of the most frequently used, second to rarefying (McMurdie & Holmes, 2014). Additionally, rare OTUs were removed from the 20 OTU lists and only those present in more than the 5% of the samples were kept, resulting in a total of five samples from the original 90 (Figure 2). This arbitrary limit, well established in previous studies (Bere, Mangadze, & Mwedzi, 2014; Stenger-Kovács, Buczkó, Hajnal, & Padisák, 2007), was necessary to keep a minimum number of samples based on which robust ecological profiles of the OTUs are ensured.

Principal component analysis (PCA) was executed using the "prcomp" function in R (R Development Core Team, 2008; Venables & Ripley, 2002) to study the structure of the samples and their relationship to the environmental (physical and chemical) variables (Figures 2, 3). We used the variables shown to be related to anthropogenic pressure in previous study: turbidity [NFU], total suspended solids (TSS [mg/L]), dissolved organic carbon (DOC [mg/L]), total organic carbon (TOC [mg/L]), total nitrogen (TN [mg/L]), total phosphorus (TP [mg/L]), nitrite (NO_2^- [mg/L]), nitrate (NO_3^- [mg/L], phosphate (PO_4^{3-} [mg/L]), and ammonium ($NH_4^{\pm}$ [mg/L]) (Tapolczai et al., 2017). Logarithmic transformation was applied to the environmental variables in order to ensure the normal distribution required for PCA. The first axis of the PCA (PC1) represents the reference pressure gradient; that is, the position of the samples along this gradient represents the "reference" quality to which the Idx_{OTU} values were compared. The values and summary statistics describing the environmental variables are presented in Appendix S2.

All 20 datasets were randomly divided according to the 20 SST levels into a training dataset containing the 75% of the samples including their position along PC1 and their associated OTU relative abundances and into a test dataset containing the remaining 25% of the samples (Figure 2). Therefore, the index could be tested on an independent dataset (test) that was not included in index development

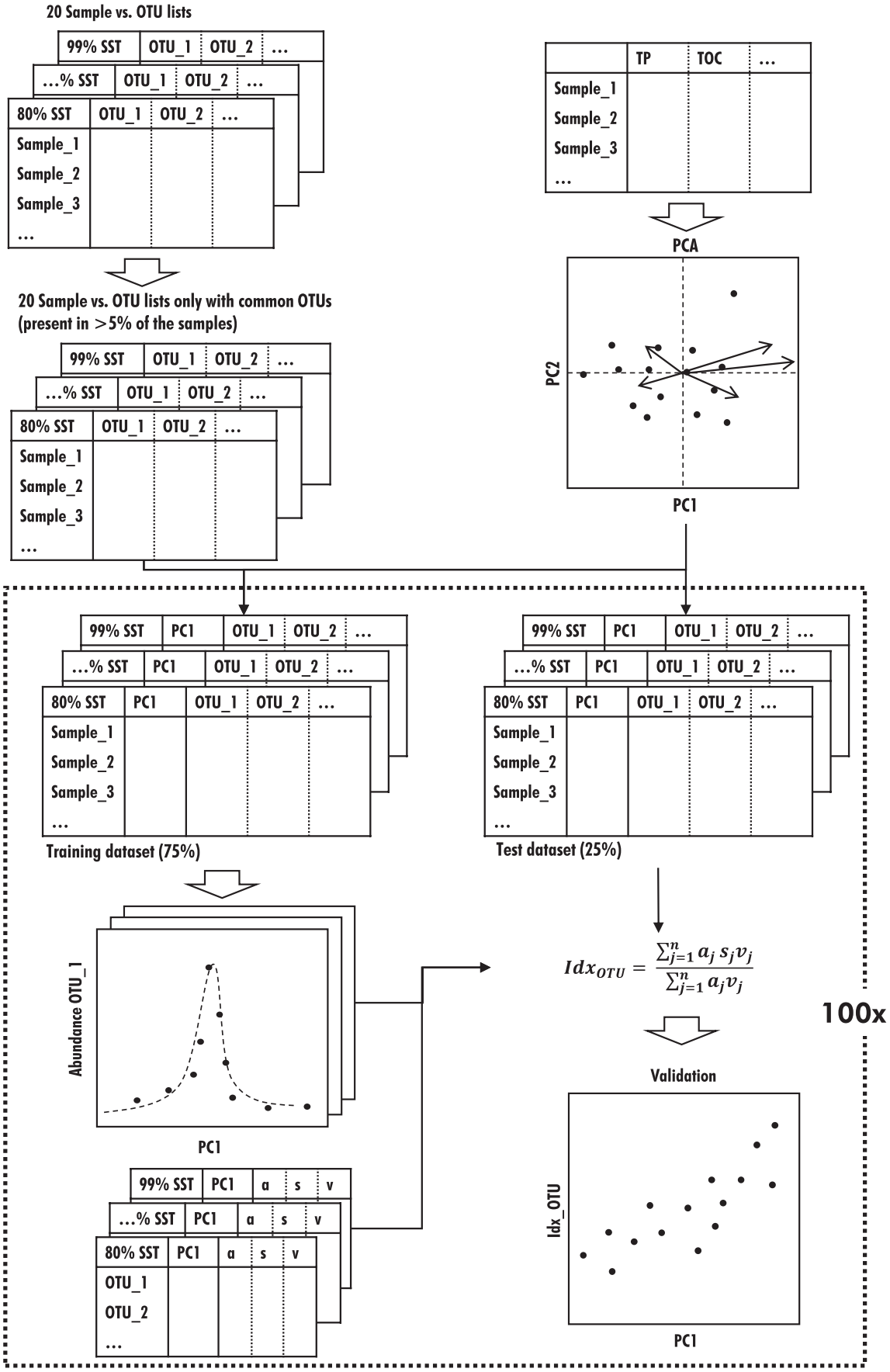

FIGURE 2 Schematic representation of the index development process and statistical analyses

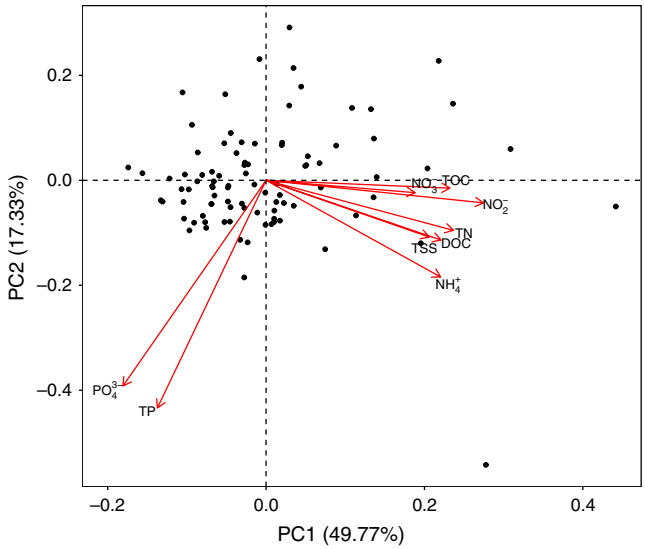

FIGURE 3 Two-dimensional graphical representation of principal component analysis results. The environmental variables in this analysis were ammonium (NH_4^+), dissolved organic carbon (DOC), nitrate (NO_3^-), nitrite (NO_2^-), phosphate (PO_4^{3-}), total nitrogen (TN), total organic carbon (TOC), total phosphorus (TP), and total suspended solids (TSS). Sample locations on the primary axis (PC1) represent the reference pressure gradient, from "high" to "poor" quality

(training). Although the selection of the two datasets was random, the proportions of the samples belonging to the three sampling networks (see Section 2.1) were maintained (0.41, 0.33, and 0.26 for REF, RCS, and POLL, respectively), to ensure a reasonable range of the pressure gradients. At each SST, a random selection of datasets was executed 100 times to measure the average and standard deviation of the Idx_{OTU} values at each sample instead of a single measure that could bias the results. The 100 iterations also allowed for all of the samples to be included in the training and test datasets too. This resulted in 100 training and test datasets at each SST. The whole process resulted in 100 indices tested for each of the 20 SSTs datasets (2,000 indices in total). It is important to note that quality values in the results only contain the Idx_{OTU} values calculated on the test dataset.

The ecological profiles of the OTUs in the training datasets were defined by modeling the relative abundance of each OTU in the samples along PC1 (Figure 2). Weighted averages and standard deviations of the profiles were calculated to estimate the ecological optimum (s) and the tolerance (v) values of the OTUs. The Zelinka-Marvan equation (Zelinka & Marvan, 1961) was adapted to our data to define Idx_{OTU}:

$$Idx_{OTU} = \frac{\sum_{j=1}^{n} a_j s_j v_j}{\sum_{j=1}^{n} a_j v_j},$$

where a_j = relative abundance of OTU j, s_j = sensitivity value or optimum of OTU j, and v_j = indicator value or tolerance of OTU j in the sample. Sensitivity and indicator values for each OTU were calculated from the abundance values plotted as functions of the samples' PC1 values. The two ecological values of each OTU comprised a database that was used together with the abundance of the OTUs in the samples for which the index was calculated. Only the data from the training dataset were used to define these profiles. The Idx_{OTU} was calculated for each site in the test dataset and then correlated with its position on PC1 (Figure 2).

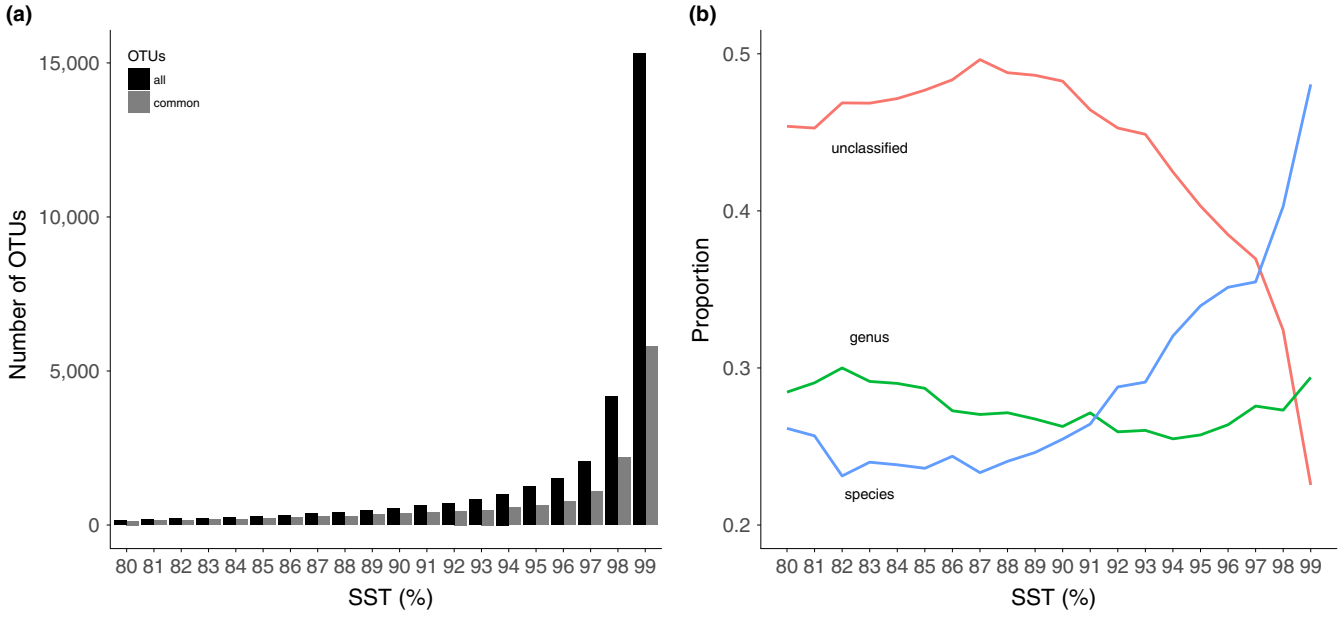

FIGURE 4 Number of OTUs (a) across the SST gradient before (black) and after (gray) removing the rare ones (present in less than 5% of the samples). Taxonomic affiliation of the OTUs with each experimental SST (b): proportion of OTUs identified at the species and genus level and of unclassified OTUs

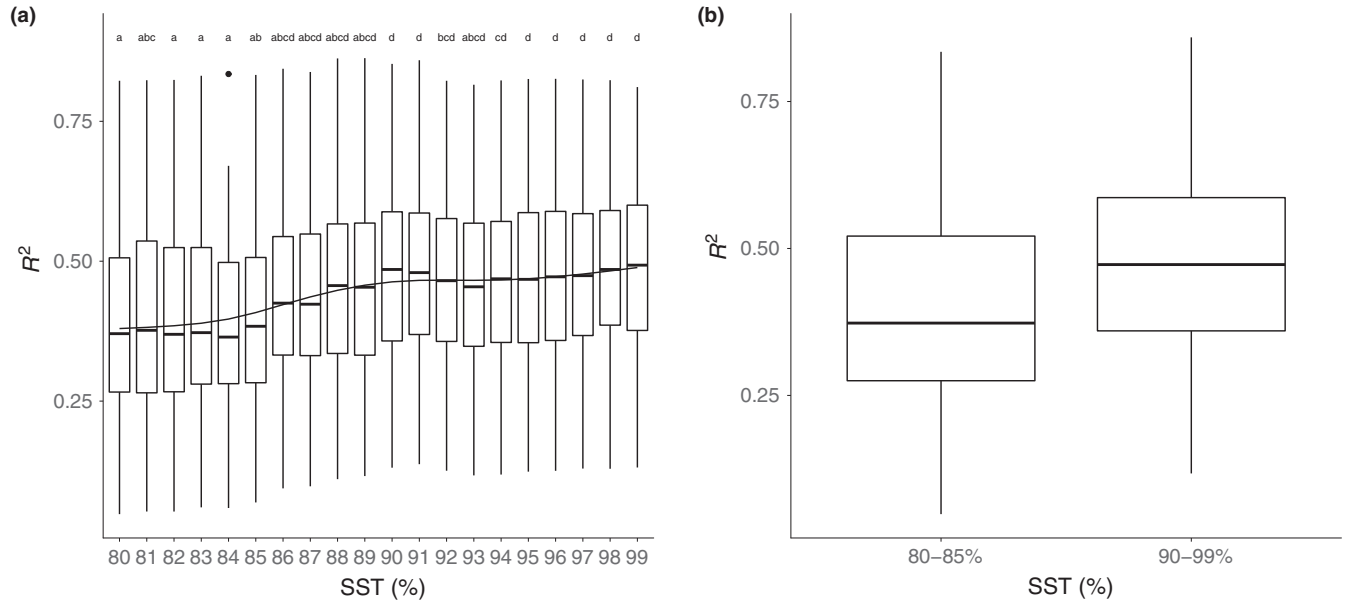

FIGURE 5 Linear regression coefficients (R^2) obtained using linear models fitted to the relationship between Idx_{OTU} and PC1 values from the test dataset are presented for each SST (a). The difference in the R^2 values between the SSTs of 80%–85% (mean = 0.39) and 90%–99% (mean = 0.47) was statistically significant (Student's *t* test, $p < 0.01$).

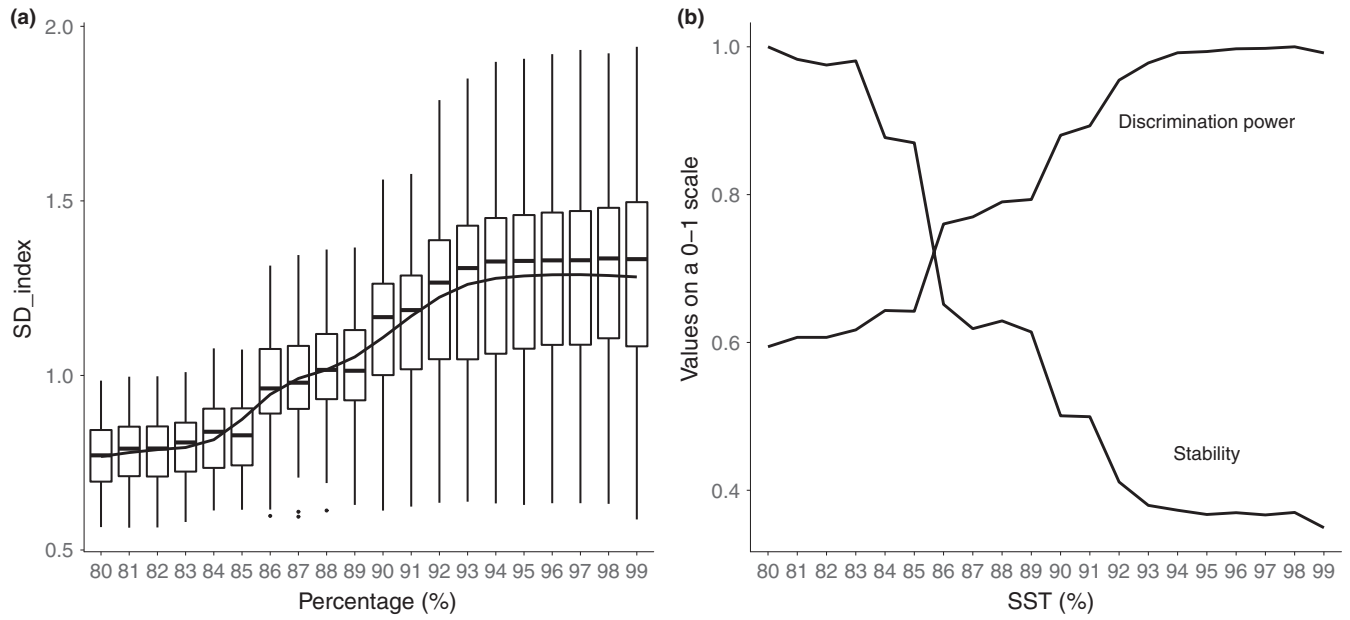

FIGURE 6 Standard deviations from the Idx_{OTU} values (SD_index) of samples within each of the 100 randomization steps were calculated and plotted against the associated SST (a). The mean (reflective of the discrimination power of the index) and the standard deviation (reflective of the stability of the index) of the SD_index as a function of the SST, normalized to a scale ranging from 0 to 1 (b)

2.5 | Idx_{OTU} performance

Three different parameters were examined to assess Idx_{OTU}'s performance.

1. Fitting a linear model using the "lm" function in R (Chambers, 1992; R Development Core Team, 2008) to ascertain the relationship between the calculated Idx_{OTU} values and their reference quality conditions (PC1). At each SST level, 100 linear models were fitted due to the 100 randomizations used when selecting the training and test datasets. The regression coefficients (R^2) of the linear models were used to reflect the performance of the index. These R^2 values were then plotted as a function of the SST, as described in Section 1,2 and in Figure 5.
2. Another aspect of the index's performance was studied through the variability of the Idx_{OTU} values among the samples within each

randomization step and SST. For this purpose, the standard deviations (*SD*) of the calculated Idx_{OTU} values for the samples within each randomization step were compared and then plotted against the SST gradient (Figure 6). Here, the *SD* was considered as a proxy for the discrimination power of the index, that is, the ability to distinguish between samples with different index values from each other (see Section 3.2.3 and Figure 6).

3. The stability of the index was tested by its ability to reproduce the same results over the course of the 100 randomization steps within each SST. Stability values were calculated for (*a*) the index, using the *SD* of the discrimination powers (see Section 3.2.4 and Figure 5) and for (*b*) the samples themselves, using the *SD* of the Idx_{OTU} values per samples (see Section 3.2.5 and Figure 7).

3 | RESULTS

3.1 | Taxonomic resolution and number of OTUs

Of the 20 OTU lists, the number of OTUs increased exponentially with the SST (Figure 4a). The number of OTUs ranged from 159 at 80% SST to 15,296 at 99% SST. Common OTUs, those that were present in over 5% of the samples, showed similar trends; however, the ratio of the removed rare OTUs increased too: at 99%, over 60% of the OTUs were removed, whereas at 80%, the percentage dropped to only 18%. Assigning taxonomy to the OTUs revealed that the taxonomic resolution changed dramatically with the SST (Figure 4b). From 80% to 93%, the percentage of unclassified OTUs varied between 45% and 50% and then steeply decreased with SSTs over 93% (Figure 4b). The proportion of OTUs identified at the species level exhibited the opposite trend. For SSTs up to 90%, the proportion was around 25%, followed by a sharp increase that reached approximately 50% at 99% SST (Figure 4b). The proportion of OTUs identified at the genus level did not display such a clear pattern, with variations from 25% to 30% across the SST gradient (Figure 4b).

3.2 | Index performance

3.2.1 | Relationship between Idx_{OTU} and the reference gradient

After developing Idx_{OTU} on the training dataset for each SST and randomization step, quality values were calculated for the corresponding test datasets. Then, the relationship between the calculated index values and PC1 was studied using linear models and their R^2 (Figure 5a). The range of the R^2 values is always high, due to the outlier datasets generated during the randomization process. The index's performance showed an increase between 85% and 91% SST. This increase was evaluated by comparing the difference in R^2 values between the 80%–85% and 91%–99% SSTs (Figure 5b). It increased significantly from a mean value of 0.39 to 0.47 (Student's *t* test, $p < 0.01$). The correlations between the index values were calculated on the samples from the test dataset, and the PC1 values were statistically significant in ~95% of the cases.

3.2.2 | Discrimination power of Idx_{OTU}

The SDs of the Idx_{OTU} values (SD_index) among the samples from each of the randomization and SST steps were calculated and then used to reflective discrimination power (Figure 6). An increase in the SD_index was observed along the SST gradient with a steep transition at 86%–93% at which point it reached a plateau without any further increases (Figure 6a).

3.2.3 | Stability of Idx_{OTU}

Concomitantly with the increase in discrimination power, we observed an increase in the interquartile ranges of the boxplots (Figure 6a). The *SD* of the SD_index was used to estimate the stability of the discrimination power against the 100 random selection processes applied to generate the training and test datasets. Higher *SD* values correspond to higher levels of variability in the discrimination power of the 100 randomization steps at one given SST. The observed increase in the discrimination power and associated decrease in stability along the SST gradient is presented in Figure 6b. Values of both parameters were standardized to a scale that ranges from 0 to 1.

3.2.4 | Stability of the samples' Idx_{OTU} values

Figure 7a depicts the variation (also measured in *SD*) in Idx_{OTU} values due to the randomization steps for each sample at each SST. The samples are ordered by mean Idx_{OTU} value on the y-axis from poor to high quality. This graphical representation illuminates that the quality values of the samples at the two ends of the quality gradient varied greatly. Mainly, the poor-quality samples had more variation in their Idx_{OTU} values and this variation increased with the SST. The variability in the Idx_{OTU} values of samples in the middle of the pressure gradient was lower, with only a few exceptions (Figure 7a).

3.2.5 | OTU richness and index stability

We linked the instability of the Idx_{OTU} values with OTU richness (number of OTUs per samples). The index value variability of each site (*SD*) and OTU richness (number of OTUs normalized across the experimental SSTs) correlated negatively, with higher *SD* values corresponding to low richness and lower *SD* being associated with higher richness ($r = -0.54$, $p < 0.05$; Figure 7b).

4 | DISCUSSION

4.1 | Sequence similarity threshold and taxonomic resolution

In the present study, the SST used for clustering sequences into OTUs was regarded as a proxy for taxonomic resolution but because this assumption can be up for debate, it requires further explanation. As discussed comprehensively by Mann (1999), the lack of a

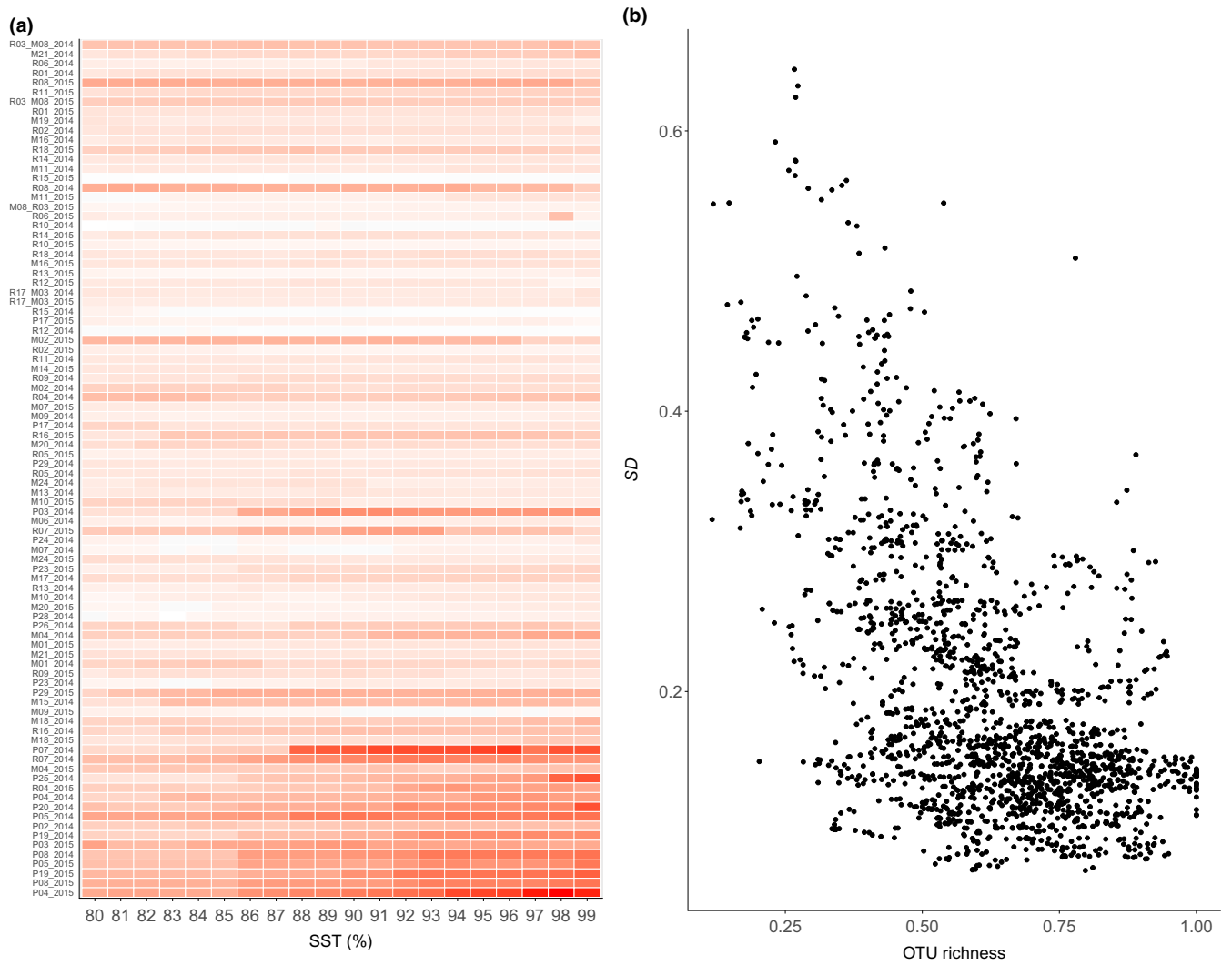

FIGURE 7 Variability (standard deviation) in the Idx_{OTU} values for each sample as a function of the associated SST. (a) Dark red cells represent higher standard deviation while the lighter colors indicate lower standard deviations. (b) Plot of standard deviations against OTU richness illustrates their statistically significant negative correlation (Pearson's correlation test, $p < 0.01$, $r = -0.54$).

solid conceptual basis for diatom taxonomy has resulted in a rapidly changing, unstable classification system of diatoms. Original approaches were based on the morphological characteristics of the specimens but the inclusion of DNA barcoding techniques (Hebert et al., 2003) helped create a taxonomy based on morphology and supplemented by molecular characters. Another important aspect to consider is that both the intragenomic variation and the intraspecies variation of the barcoding gene differ among taxa (Hamsher, Evans, Mann, Poulíčková, & Saunders, 2011). Although we used a hierarchical clustering method with predefined global thresholds in the current study, there are other clustering methods available that we have not tested. The importance of clustering methods presented in the introduction should not be overlooked and warrants further analysis in further studies.

The choice of appropriate taxonomic resolution for bioassessment purposes is a common and active topic of debate in biomonitoring most every biota studied. The identification to the lowest taxonomic level is important for complex ecological questions, fundamental studies, and for simply expanding the common knowledge base. One of the arguments in favor of precise taxonomic resolution (i.e., species-level) is built on the fact that species represent the basic units of an ecosystem and a clear and thorough understanding of their ecological niches directly impacts the amount, quality, and value of the information they provide (Salmaso, Naselli-Flores, & Padisák, 2015). However, in practice, classifying specimens to species (or finer) taxonomic level does not necessarily further inform or improve bioassessment. For example, aquatic macroinvertebrates are usually identified down to species or genus level, depending on the taxa and the life stages of the organisms. However, several studies have been unable to find significant differences in the bioindication efficiency of the same community at different taxonomic levels (even family) for the same type of pollution (Bailey, Norris, & Reynoldson, 2001; Bowman & Bailey, 1997). Furthermore, the applicability depends on the metrics being used. For macroinvertebrates, more complex metrics exist than for diatoms, including functionality, life-forms, and habitat preferences. A comprehensive study by Schmidt-Kloiber and

Nijboer (2004) showed that while some metrics were not sensitive to changes in the taxonomic level (e.g., richness, diversity measures), others (e.g., functional metrics) did impact final quality values.

The question is of greater relevance for the study of protists, where the microscopic identification at the species level is technically challenging and labor intensive (Lavoie et al., 2009; Rimet & Bouchez, 2012). Published studies on the effect of taxonomic resolution are contradictory, and their results seem to depend largely on the index characteristics. For instance, Lavoie et al. (2009) studied the effect of reducing the taxonomic resolution to the genus level for the Eastern Canadian Diatom Index. They found that while it could still successfully separate impacted sites from reference ones, its ability to detect fine changes in the environment had diminished. Other studies, however, have found that the change from species level to genus level does not impact bioassessment efficiency. This has been tested using indicators of river regulation (Growns, 1999) and organic and nutrient pollution (Rimet & Bouchez, 2012).

DNA barcoding has enabled the detection of intraspecific variations that were not detectable using microscopy-based analysis (Keck et al., 2017; Vasselon, Rimet, et al., 2017). Generally, a sequence similarity of 95% has been used for species-level delimitation when meta-barcoding diatoms. In this study, 1,239 OTUs at 95% similarity were identified, which is clearly dwarves the 382 species identified by microscopy (Tapolczai et al., 2017). Such striking differences between the results obtained through microscopy and HTS are commonly reported in the study of diatoms (Rivera, Vasselon, Jacquet, et al., 2018b). These differences are largely due to the cryptic diversity common to diatoms; indeed, it has been shown in several species that the genetic diversity is substantially richer than the morphological diversity (Evans, Wortley, Simpson, Chepurnov, & Mann, 2008; Mann et al., 2004; Souffreau et al., 2013). This relatively newfound ability to recognize these cryptic species is important because their ecological niches may differ even when they live in sympatry (Kelly, Trobajo, Rovira, & Mann, 2015; Rovira, Trobajo, Sato, Ibáñez, & Mann, 2015).

While the effect of taxonomic resolution on bioassessment has been extensively studied using microscopy-based identification, "OTU studies" have relied on arbitrary clustering thresholds until now. Our study revealed maximum index performance at a 91%–92% SST, lower than traditionally thought necessary. However, it must be noted that the validity and applicability of this threshold are potentially limited to the conditions included in the present analysis (e.g., pollution gradient, community structure) and further experimental consideration and validation are required prior to being exported for widespread use.

4.2 | Performances of the OTU-based indices depend on the SST choice

In the present study, we developed a diatom index, based on the same principles as classical ones (e.g., PSI, Coste, 1982; BDI, Coste, Boutry, Tison-Rosebery, & Delmas, 2009; TDI Kelly, 1998). However, in this case, we directly applied the ecological profile of the OTUs, without taxonomic assignment, avoiding the problem associated with incomplete DNA reference databases, which can easily inject bias (Apothéloz-Perret-Gentil et al., 2017; Groendahl et al., 2017; Rivera, Vasselon, Jacquet, et al., 2018b; Zimmermann, Glöckner, Jahn, Enke, & Gemeinholzer, 2015). Even though the number of species included in the DNA reference libraries is constantly increasing (Rivera, Vasselon, Jacquet, et al., 2018b), the proportion of OTUs that can be assigned to the species level remains far from satisfying. Published reports have described a wide range of classification coverage, including 35.7% (Vasselon, Rimet, et al., 2017), 35% (Apothéloz-Perret-Gentil et al., 2017), 23% (Rivera, Vasselon, Jacquet, et al., 2018b), and as low as 10% for marine samples (Rivera, Vasselon, Ballorain, et al., 2018a). Our approach is similar to those of Apothéloz-Perret-Gentil et al. (2017) but the method for defining the OTUs' indicator and sensitivity values used in the Zelinka–Marvan equation (1961) is different. In Apothéloz-Perret-Gentil et al.'s study, sites were preclassified using the original Swiss morphology-based index and served as the reference from which the OTU indicator and sensitivity values were defined. In the present study, we directly incorporated the environmental pressure gradients of both physical and chemical parameters using multivariate analysis. Thus, the method described here is completely independent of morphology-based taxonomy. The disadvantage, however, is that rare OTUs have unreliable ecological profiles and must be removed to safeguard the accuracy and of a system that is most effective when only based on robust OTUs. Our results indicate that the index values calculated for the test dataset correlated significantly with the pressure gradient regardless of the SST; however, an important transitional zone in the SST gradient was observed from 86% to 91%, described by an increasing R^2 value.

One technical drawback of the index described in this investigation is that when new samplings are carried out in Mayotte, the OTUs generated from these new data may differ from those obtained from the datasets used our index development. Indeed, the sequence composition of the sampling sites can fluctuate over time. This means that the ecological profile, together with the representative sequence of an OTU, must be fixed and assigned to the correct OTU, generated by another sequencing run. However, this requires the calibration, standardization, and extensive validation of the OTU clustering method given its potential to greatly effect final OTU composition.

Interesting, our data uncovered an important trade-off between the index's discrimination power and its stability. The stability of Idx_{OTU} during the randomization process decreased with increasing SSTs. This is due to the exponential increase in OTUs, many of which become less frequent and a higher number warrant removal. Regardless, these OTUs consist of fewer sequences, and thus, their ecological profile cannot be established with any robustness. Thus, the biasing effect of one outlier abundance data point becomes higher and this makes the dataset very sensitive to the random selection process for the training and test datasets. In contrast, coarse

taxonomic resolution generates fewer OTUs with wider but more reliable ecological profiles; this leads to a more stable Idx_{OTU} with weaker discrimination power. This instability was particularly important when samples presented with low OTU richness. Low OTU richness is observed at highly polluted sites, where only a few resistant species could survive (Blanco et al., 2012; Stevenson, Pan, & Van Dam, 2010) and in reference sites where nutrient limitation has selected for a limited number of taxa (Blanco et al., 2012; Stevenson, Hill, Herlihy, Yuan, & Norton, 2008). A technical drawback of using a high similarity threshold is the elevated risk of misappropriating biological relevance to artifacts and sequence errors, which results in biased and inaccurate results (Patin et al., 2013).

An ecological disadvantage of the de novo hierarchical clustering used in this study is that it hinges on a single, global SST regardless of the species being considered. Given the differential levels of intra and interspecies genetic variation, a clustering method that tailors the SST to the specific characteristics of each taxon would be a valuable tool. Without this, there is always a risk of undergrouping heterogeneous sequences and thereby creating an ecological profile that is not indicative or representative of some taxa, while simultaneously running the risk of overgrouping and thereby separating groups of sequences with similar ecological preferences for other taxa. Further studies should implement approaches similar to that described by Preheim, Perrotta, Martin-Platero, Gupta, and Alm (2013) that employs the ecological preferences of bacterial sequences (termed distribution-based clustering) to refine the OTUs.

4.3 | The ecologically naïve paradigm of diatom indices

The taxonomy-free approach delineated in this study proposes a solution to overcome the considerable technical challenge posed by incomplete reference databases. However, it is important to highlight that the fundamental basis of Idx_{OTU} and the classical taxonomy-based diatom indices are the same: using uncritically the relative abundance of a list of species (or OTUs) and their ecological values to develop an index calculated from a weighted average equation that correlates with the physical and chemical parameters of an environment. These fundamental aspects have been analyzed in several previous studies: Instead of using this ecologically naïve approach, other groups have proposed reconsidering the functional aspects underlying the ecological indication, for example, using relative biovolume instead of relative abundance (Tapolczai et al., 2017), trait-based functional groups (B-Béres et al., 2016), or diversity metrics (Blanco et al., 2012). Nevertheless, the indices currently employed in the EU WFD are criticized for their lack of ecological basis (Schneider, Hilt, Vermaat, & Kelly, 2016) and all of the indices based on molecular techniques use the same naïve approaches (Kelly et al., 2018; Leese et al., 2016), which are qualitatively not ideal. In this study, we are clearly not advocating the uncritical widespread use of Idx_{OTU} but instead used it as a test object to assess the effect of the SST was tested.

Sequencing methods can potentially address some of these drawbacks. It has been shown in a previous study that DNA read abundances, using the *rbcL* marker correlate reliably with species' relative biovolume (Vasselon et al., 2018), thus enabling the generation of more ecologically relevant taxa quantification data (Tapolczai, 2017). Moreover, molecular methods facilitate the analysis of other benthic taxa beyond diatoms (e.g., Cyanobacteria, Chlorophyta, Rhodophyta) without the need of experts that specialize in each of these groups. The application of this rapidly advancing technology has the potential to provide a much more holistic, representative view of phytobenthic composition (Groendahl et al., 2017).

ACKNOWLEDGMENTS

This study was funded by ONEMA (French National Office for Water and Aquatic Ecosystems) under the 2013–2018 "Développement d'outils de bioindication « phytobenthos » et «macroinvértébrés benthiques » pour les eaux de surface continentales de Mayotte" program.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

FR, AB, and KT devised the concept and designed the study. FR, AB, KT, and VV carried out the sampling. VV prepared the samples for sequencing and the bioinformatic analyses for OTU clustering. KT carried out the statistical analyses and tests, index development, and wrote the manuscript together with the supporting material and figures. FR, AB, CSK, and JP provided constructive feedback and suggestions toward improving the manuscript.

DATA ACCESSIBILITY

All data are uploaded as supporting material and also stored in Zenodo file repository (https://doi.org/10.5281/zenodo.1443416).

ORCID

Kálmán Tapolczai https://orcid.org/0000-0003-1453-767X
Valentin Vasselon https://orcid.org/0000-0001-5038-7918
Agnès Bouchez https://orcid.org/0000-0001-8802-6966
Csilla Stenger-Kovács https://orcid.org/0000-0001-6175-4904
Judit Padisák https://orcid.org/0000-0001-8285-2896
Frédéric Rimet https://orcid.org/0000-0002-5514-869X

REFERENCES

Afnor. (2014). NF EN 13946.Qualité de l'eau – Guide pour l'échantillonnage en routine et le prétraitement des diatomées benthiques. De Rivières Et De Plans.

Afnor. (2016). NF T90 354 - Qualité de l'eau - Échantillonnage, traitement et analyse de diatomées benthiques en cours d'eau et canaux. pp. 1–79.

APHA. (2012). *Standard methods for the examination of water and wastewater*, 22nd ed. Washington, DC: APHA American Public Health Association.

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources*, https://doi.org/10.1111/1755-0998.12668.

Bailey, R. C., Norris, R. H., & Reynoldson, T. B. (2001). Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. *Journal of the North American Benthological Society.*, *20*(2), 280–286. https://doi.org/10.2307/1468322

Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology.*, *21*(8), 2039–2044. https://doi.org/10.1111/j.1365-294X.2012.05519.x

B-Béres, V., Lukács, Á., Török, P., Kókai, Z., Novák, Z., T-Krasznai, E., … Bácsi, I. (2016). Combined eco-morphological functional groups are reliable indicators of colonisation processes of benthic diatom assemblages in a lowland stream. *Ecological Indicators.*, *64*, 31–38. https://doi.org/10.1016/j.ecolind.2015.12.031

Bere, T., Mangadze, T., & Mwedzi, T. (2014). The application and testing of diatom-based indices of stream water quality in Chinhoyi Town, Zimbabwe. *Water SA.*, *40*(3), 503. https://doi.org/10.4314/wsa.v40i3.14

Besse-Lototskaya, A., Verdonschot, P. F. M., Coste, M., & Van de Vijver, B. (2011). Evaluation of European diatom trophic indices. *Ecological Indicators.*, *11*(2), 456–467. https://doi.org/10.1016/j.ecolind.2010.06.017

Birtel, J., Walser, J.-C., Pichon, S., Bürgmann, H., & Matthews, B. (2015). Estimating bacterial diversity for ecological studies: Methods, metrics, and assumptions. *PLoS ONE*, *10*(4), e0125356. https://doi.org/10.1371/journal.pone.0125356

Blanco, S., Cejudo-Figueiras, C., Tudesque, L., Bécares, E., Hoffmann, L., & Ector, L. (2012). Are diatom diversity indices reliable monitoring metrics? *Hydrobiologia*, *695*(1), 199–206. https://doi.org/10.1007/s10750-012-1113-1

Bowman, M. F., & Bailey, R. C. (1997). Does taxonomic resolution affect the multivariate description of the structure of freshwater benthic macroinvertebrate communities? *Canadian Journal of Fisheries and Aquatic Science*, *54*(8), 1802–1807. https://doi.org/10.1139/f97-085

Chambers, J. M. *Linear models. Statistical Models in S*. 4th ed. Wadsworth & Brooks/Cole, 1992.

Chonova, T., Keck, F., Labanowski, J., Montuelle, B., Rimet, F., & Bouchez, A. (2016). Separate treatment of hospital and urban wastewaters: A real scale comparison of effluents and their effect on microbial communities. *Science of the Total Environment*, *542*(Part A), A:965–75.

Coste, M. (1982) *Étude des méthodes biologiques d'appréciation quantitative de la qualité des eaux*. Rapport Cemagref QE Lyon-AF Bassin Rhône Méditerranée Corse.

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*(4), 621–650. https://doi.org/10.1016/j.ecolind.2008.06.003

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods.*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

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. Hajibabaei M, editor. *PLoS ONE*. *10*(7):e0130324.

European Commission. (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23rd October 2000 establishing a framework for Community action in the field of water policy. *Official Journal of the European Communities.*, *327*, 1–72.

Evans, K. M., Wortley, A. H., Simpson, G. E., Chepurnov, V. A., & Mann, D. G. (2008). A molecular systematic approach to explore diversity within the *Sellaphora pupula* species complex (Bacillariophyta) 1. *Journal of Phycology.*, *44*(1), 215–231.

Flynn, J. M., Brown, E. A., Chain, F. J. J., MacIsaac, H. J., & Cristescu, M. E. (2015). Toward accurate molecular identification of species in complex environmental samples: Testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, *5*(11), 2252–2266. https://doi.org/10.1002/ece3.1497

Groendahl, S., Kahlert, M., & Fink, P. (2017). The best of both worlds: A combined approach for analyzing microalgal diversity via metabarcoding and morphology-based methods. Doi H, editor. *PLoS ONE*. *12*(2):e0172808.

Growns, I. (1999). Is genus or species identification of periphytic diatoms required to determine the impacts of river regulation? *Journal of Applied Phycology.*, *11*(3), 273–283.

Hamsher, S. E., Evans, K. M., Mann, D. G., Poulíčková, A., & Saunders, G. W. (2011). Barcoding diatoms: Exploring alternatives to COI-5P. *Protist.*, *162*(3), 405–422. https://doi.org/10.1016/j.protis.2010.09.005

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(1512), 313–321.

Kaczmarska, I., Mather, L., Luddington, I. A., Muise, F., & Ehrman, J. M. (2014). Cryptic diversity in a cosmopolitan diatom known as *Asterionellopsis glacialis* (Fragilariaceae): Implications for ecology, biogeography, and taxonomy. *American Journal of Botany.*, *101*(2), 267–286.

Kahlert, M., Albert, R.-L., Anttila, E.-L., Bengtsson, R., Bigler, C., Eskola, T., et al. (2009). Harmonization is more important than experience—results of the first Nordic-Baltic diatom intercalibration exercise 2007 (stream monitoring). *Journal of Applied Phycology*, *21*(4), 471–482. https://doi.org/10.1007/s10811-008-9394-5

Kahlert, M., Kelly, M., Albert, R.-L., Almeida, S. F. P., Bešta, T., Blanco, S., et al. (2012). Identification versus counting protocols as sources of uncertainty in diatom-based ecological status assessments. *Hydrobiologia*, *695*(1), 109–124. https://doi.org/10.1007/s10750-012-1115-z

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., & Bouchez, A. (2017). Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, *15*(5), 266–274. https://doi.org/10.1002/fee.1490

Kelly, M. G. (1998). Use of the trophic diatom index to monitor eutrophication in rivers. *Water Research.*, *32*(1), 236–242. https://doi.org/10.1016/S0043-1354(97)00157-7

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann, D., Pass, D., … Glover, R. (2018). *A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers* [Internet]. Bristol, UK: Environment Agency; p. 157. Report No.: SC140024/R. Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/684493/A_DNA_based_metabarcoding_approach_to_assess_diatom_communities_in_rivers_-_report.pdf

Kelly, M. G., King, L., Jones, R. I., Barker, P. A., & Jamieson, B. J. (2008). Validation of diatoms as proxies for phytobenthos when assessing ecological status in lakes. *Hydrobiologia*, *610*(1), 125–129. https://doi.org/10.1007/s10750-008-9427-8

Kelly, M. G., Trobajo, R., Rovira, L., & Mann, D. G. (2015). Characterizing the niches of two very similar Nitzschia species and implications for ecological assessment. *Diatom Research*, *30*(1), 27–33.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J. F., & Bouchez, A. (2013). Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: A test for freshwater diatoms. *Molecular Ecology Resources*, *13*(4), 607–619. https://doi.org/10.1111/1755-0998.12105

Lavoie, I., Dillon, P. J., & Campeau, S. (2009). The effect of excluding diatom taxa and reducing taxonomic resolution on multivariate analyses and stream bioassessment. *Ecological Indicators.*, *9*(2), 213–225. https://doi.org/10.1016/j.ecolind.2008.04.003

Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., et al. (2016). DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes.*, *2*, e11321.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., et al. (2014). DNA-based species delimitation in algae. *European Journal of Phycology.*, *49*(2), 179–196. https://doi.org/10.1080/09670262.2014.904524

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., … Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology.*, *30*(5), 434–439. https://doi.org/10.1038/nbt.2198

Mann, D. G. (1999). The species concept in diatoms. *Phycologia*, *38*(6), 437–495. https://doi.org/10.2216/i0031-8884-38-6-437.1

Mann, D. G., McDonald, S. M., Bayer, M. M., Droop, S. J. M., Chepurnov, V. A., Loke, R. E., et al. (2004). The Sellaphora pupula species complex (Bacillariophyceae): Morphometric analysis, ultrastructure and mating data provide evidence for five new species. *Phycologia*, *43*(4), 459–482. https://doi.org/10.2216/i0031-8884-43-4-459.1

Mann, D. G., & Vanormelingen, P. (2013). An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology*, *60*(4), 414–420. https://doi.org/10.1111/jeu.12047

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLOS Computational Biology.*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

Patin, N. V., Kunin, V., Lidström, U., & Ashby, M. N. (2013). Effects of OTU clustering and PCR artifacts on microbial diversity estimates. *Microbial Ecology.*, *65*(3), 709–719. https://doi.org/10.1007/s00248-012-0145-4

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., & Alm, E. J. (2013). Distribution-based clustering: Using ecology to refine the operational taxonomic unit. *Applied and Environment Microbiology*, *79*(21), 6593–6603. https://doi.org/10.1128/AEM.00342-13

R Development Core Team. (2008). *A language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org

Rimet, F., Abarca, N., Bouchez, A., Kusber, W.-H., Jahn, R., Kahlert, M., … Trobajo, R. (2018a). The potential of high throughput sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. *Fottea*, *18*(1), 37–54.

Rimet, F., & Bouchez, A. (2012). Biomonitoring river diatoms: Implications of taxonomic resolution. *Ecological Indicators.*, *15*(1), 92–99. https://doi.org/10.1016/j.ecolind.2011.09.014

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., … Bouchez, A.. (2016). R-Syst:Diatom: An open-access and curated barcode database for diatoms and freshwater monitoring. *Database (Oxford)*. https://doi.org/10.1093/database/baw016

Rimet, F., Vasselon, V., A.-Keszte, B., & Bouchez, A. (2018b). Do we similarly assess diversity with microscopy and high-throughput sequencing? Case of microalgae in lakes. *Organisms Diversity & Evolution*, *18*(1), 51–62.

Rivera, S. F., Vasselon, V., Ballorain, K., Carpentier, A., Wetzel, C. E., Ector, L., et al. (2018a). DNA metabarcoding and microscopic analyses of sea turtles biofilms: Complementary to understand turtle behavior. *PLoS ONE*, *13*(4), e0195770.

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., & Rimet, F. (2018b). Metabarcoding of lake benthic diatoms: From structure assemblages to ecological assessment. *Hydrobiologia*, *807*(1), 37–51.

Rovira, L., Trobajo, R., Sato, S., Ibáñez, C., & Mann, D. G. (2015). Genetic and Physiological Diversity in the Diatom Nitzschia inconspicua. *Journal of Eukaryotic Microbiology*, *62*(6), 815–832.

Salmaso, N., Naselli-Flores, L., & Padisák, J. (2015). Functional classifications and their application in phytoplankton ecology. *Freshwater Biology*, *60*(4), 603–619. https://doi.org/10.1111/fwb.12520

Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology.*, *71*(3), 1501–1506. https://doi.org/10.1128/AEM.71.3.1501-1506.2005

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environment Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schmidt-Kloiber, A., & Nijboer, R. C. (2004). The effect of taxonomic resolution on the assessment of ecological water quality classes. *Hydrobiologia*, *516*(1–3), 269–283. https://doi.org/10.1023/B:HYDR.0000025270.10807.10

Schneider, S. C., Hilt, S., Vermaat, J. E., & Kelly, M. (2016). The "Forgotten" ecology behind ecological status evaluation: Re-assessing the roles of aquatic plants and benthic algae in ecosystem functioning. In *Progress in Botany* (vol. *78*, pp. 285-304). Cham, Switzerland: Springer.

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology.*, *21*(8), 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Souffreau, C., Vanormelingen, P., Van de Vijver, B., Isheva, T., Verleyen, E., Sabbe, K., et al. (2013). Molecular evidence for distinct Antarctic lineages in the cosmopolitan terrestrial diatoms Pinnularia borealis and Hantzschia amphioxys. *Protist.*, *164*(1), 101–115. https://doi.org/10.1016/j.protis.2012.04.001

Stenger-Kovács, C., Buczkó, K., Hajnal, É., & Padisák, J. (2007). Epiphytic, littoral diatoms as bioindicators of shallow lake trophic status: Trophic Diatom Index for Lakes (TDIL) developed in Hungary. *Hydrobiologia*, *589*(1), 141–154. https://doi.org/10.1007/s10750-007-0729-z

Stevenson, R. J., Hill, B. H., Herlihy, A. T., Yuan, L. L., & Norton, S. B. (2008). Algae–P relationships, thresholds, and frequency distributions guide nutrient criterion development. *Journal of the North American Benthological Society.*, *27*(3), 783–799. https://doi.org/10.1899/07-077.1

Stevenson, R. J., Pan, Y., & Van Dam, H. (2010). Assessing environmental conditions in rivers and streams with diatoms. In: *The diatoms: Applications for the environmental and Earth Sciences* [Internet], 2nd ed. Cambridge, UK: Cambridge University Press.

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., & Mai, V. (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in Bioinformatics*, *13*(1), 107–121. https://doi.org/10.1093/bib/bbr009

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology.*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tapolczai, K. (2017). *Time for change: Towards the implementation of new approaches in diatom-based ecological quality assessment for rivers* (Doctoral dissertation). Retrieved from. http://www.theses.fr/s120055

Tapolczai, K., Bouchez, A., Stenger-Kovács, C., Padisák, J., & Rimet, F. (2017). Taxonomy- or trait-based ecological assessment for tropical rivers? Case study on benthic diatoms in Mayotte island (France, Indian Ocean). *Science of the Total Environment.*, *607–608*, 1293–1303. https://doi.org/10.1016/j.scitotenv.2017.07.093

Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., … Domaizon, I. (2018). Avoiding quantification bias in metabarcoding: Application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods in Ecology and Evolution*, *9*(4), 1060–1069. https://doi.org/10.1111/2041-210X.12960

Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., & Bouchez, A. (2017). Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? *Freshwater Science.*, *36*(1), 162–177.

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

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. 4th ed. [internet]. New York, NY: Springer. https://www.stats.ox.ac.uk/pub/MASS4

Visco, J. A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., & Pawlowski, J. (2015). Environmental monitoring: Inferring the diatom index from next-generation sequencing data. *Environmental Science & Technology.*, *49*(13), 7597–7605. https://doi.org/10.1021/es506158m

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Westcott, S. L., & Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ.*, *8*(3), e1487. https://doi.org/10.7717/peerj.1487

Zelinka, M., & Marvan, P. (1961). Zur präzisierung der biologischen klassifikation der reinheit flie\s sender gewässer. *Archiv Fur Hydrobiologie*, *57*(3), 389–407.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., & Gemeinholzer, B. (2015). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources.*, *15*(3), 526–542. https://doi.org/10.1111/1755-0998.12336

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Tapolczai K, Vasselon V, Bouchez A, Stenger-Kovács C, Padisák J, Rimet F. The impact of OTU sequence similarity threshold on diatom-based bioassessment: A case study of the rivers of Mayotte (France, Indian Ocean). *Ecol Evol*. 2019;9:166–179. https://doi.org/10.1002/ece3.4701