


# Monitoring the ecological status of rivers with diatom eDNA metabarcoding: A comparison of taxonomic markers and analytical approaches for the inference of a molecular diatom index

Laure Apothéloz-Perret-Gentil<sup>1,2</sup>  | Agnès Bouchez<sup>3</sup> | Tristan Cordier<sup>1,2</sup> | Arielle Cordonier<sup>4</sup> | Julie Guéguen<sup>3</sup> | Frederic Rimet<sup>3</sup> | Valentin Vasselon<sup>5,6</sup> | Jan Pawlowski<sup>1,2,7</sup>

<sup>1</sup>Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland

<sup>2</sup>ID-Gene ecodiagnosics, Geneva, Switzerland

<sup>3</sup>UMR CARTELE, INRAE, Université Savoie Mont-Blanc, Thonon, France

<sup>4</sup>Department of Territorial Management, Water Ecology Service, Geneva, Switzerland

<sup>5</sup>Pôle R&D "ECLA", Thonon-les-Bains, France

<sup>6</sup>OFB, Site INRA UMR CARTELE, Thonon-les-Bains, France

<sup>7</sup>Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland

## Correspondence

Laure Apothéloz-Perret-Gentil, Department of Genetics and Evolution, University of Geneva, 1211 Geneva, Switzerland.  
Email: laure.perret-gentil@unige.ch

## Funding information

European Cross-Border Cooperation Program; European Regional Development Fund; Swiss Federal; Swiss cantons; Swiss National Science Foundation, Grant/Award Number: 31003A\_179125; European Union

## Abstract

Recently, several studies demonstrated the usefulness of diatom eDNA metabarcoding as an alternative to assess the ecological quality of rivers and streams. However, the choice of the taxonomic marker as well as the methodology for data analysis differ between these studies, hampering the comparison of their results and effectiveness. The aim of this study was to compare two taxonomic markers commonly used in diatom metabarcoding and three distinct analytical approaches to infer a molecular diatom index. We used the values of classical morphological diatom index as a benchmark for this comparison. We amplified and sequenced both a fragment of the *rbcl* gene and the V4 region of the 18S rRNA gene for 112 epilithic samples from Swiss and French rivers. We inferred index values using three analytical approaches: by computing it directly from taxonomically assigned sequences, by calibrating de novo the ecovalues of all metabarcodes, and by using a supervised machine learning algorithm to train predictive models. In general, the values of index obtained using the two "taxonomy-free" approaches, encompassing molecular assignment and machine learning, were closer correlated to the values of the morphological index than the values based on taxonomically assigned sequences. The correlations of the three analytical approaches were higher in the case of *rbcl* compared to the 18S marker, highlighting the importance of the reference database which is more complete for the *rbcl* marker. Our study confirms the effectiveness of diatom metabarcoding as an operational tool for rivers ecological quality assessment and shows that the analytical approaches by-passing the taxonomic assignments are particularly efficient when reference databases are incomplete.

## KEYWORDS

aquatic bioindication, biomonitoring, diatoms, environmental DNA, metabarcoding, taxonomy-free

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *Molecular Ecology* published by John Wiley & Sons Ltd

## 1 | INTRODUCTION

Increasing anthropogenic impacts on the environment prompts many countries to implement special measures to assess these impacts and mitigate their effects. Current legislation in EU (WFD, European Commission, 2000) and Switzerland (Swiss Federal Council, 1998) recommends using several different biological groups to assess the ecological status of rivers and streams. Diatoms are one of these groups of organisms that are used for biomonitoring because they respond quickly to environmental changes and are highly sensitive to physicochemical stressors (Rimet & Bouchez, 2012). Therefore, several biotic indices based on taxonomic and ecological knowledge of diatom communities have been developed to assess environmental impacts (e.g., TDI in UK, IBD and IPS in France, the latter is also used in several European countries). Most of these indices followed the weighted average equation of Zelinka and Marvan (1961) that is based on the relative frequency of species weighted by their autecological value.

In Switzerland, the Swiss Diatom Index (DI-CH) was first developed using more than 3,500 river samples. Out of the 780 morphological taxa found in those sites, autecological values and weighting factors for the calculation of the index were kept for 300 taxa. The calculation also includes the relative frequency of each taxon in the sample (Hürlimann & Niederhauser, 2007). This index was calibrated to fit the recommendations given by the federal ordinance in terms of chemical pollution and organic enrichment of running water (Swiss Federal Council, 1998). Traditionally, the majority of diatom indices are calculated based on morphotaxonomic identification of diatoms in biofilm samples following identification keys. This process is particularly time consuming as samples have to be analysed one after the other under a microscope by an expert in diatoms' identification (Keck et al., 2017).

Environmental (e)DNA metabarcoding (as defined by Taberlet et al., 2012) applied to high-throughput sequencing of diatom taxa present in biofilm samples could overcome the limitations of the traditional microscopic approach. A molecular diatom index (MDI) inferred from eDNA metabarcoding data could provide several advantages for routine bioassessment. First, the ecological status could be inferred far more rapidly, and at lower costs. Moreover, the results could be much easier to control and compare because they would be based on comparable sequencing data. Indeed, the classical morphotaxonomic approach also presents its own biases and limitations, such as low taxonomic resolution and gaps in knowledge of morphospecies ecology (Pawlowski et al., 2018). Finally, many samples could be processed in parallel with molecular methods, which constitute a key advantage for large-scale and more continuous surveys.

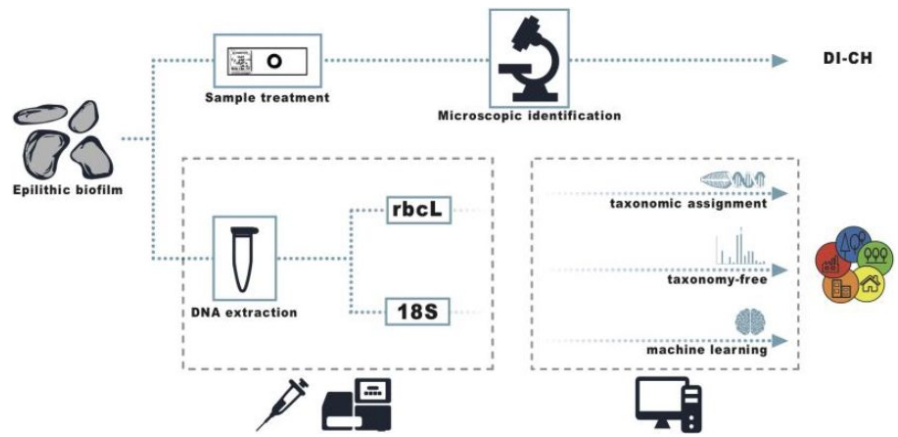
The eDNA metabarcoding has been successfully applied to monitoring past and present biodiversity in various types of environment (reviewed in; Bohmann et al., 2014; Deiner et al., 2017; Pedersen et al., 2015; Taberlet et al., 2012, 2018; Valentini et al., 2016). Its practical applications range from the detection of invasive and

endangered species (Egan et al., 2015; Thomsen et al., 2012; Zaiko et al., 2015) to biodiversity surveys for assessing ecosystem conditions (Bista et al., 2017; Chariton et al., 2015; Djurhuus et al., 2020) and industrial impacts (Lanzén et al., 2016; Laroche et al., 2018; Pawlowski et al., 2014). Special attention was paid to inferring biotic indices from eDNA metabarcoding data (Pawlowski et al., 2018). Some studies compared the traditional morphotaxonomy-based indices to those inferred from metabarcoding data (Lejzerowicz et al., 2015). Others proposed new indices based on metabarcoding data (Aylagas et al., 2014, 2017; Keeley et al., 2018). New analytic approaches have been developed using machine learning to predict biotic indices from metabarcoding data (Cordier et al., 2017, 2018, 2019; Frühe et al., 2020).

The usefulness of diatom metabarcoding to infer a species list and/or assess the water quality has been explored in several studies of European rivers and streams (Bailet et al., 2019; Kelly et al., 2018; Kermarrec et al., 2013, 2014; Pérez-Burillo et al., 2020; Rivera et al., 2020; Vasselon et al., 2017; Visco et al., 2015; Zimmermann et al., 2015). Some studies have been conducted to address methodological issues inherent to metabarcoding methods, focusing either on the generation of the data (DNA extraction: Vasselon et al., 2017; OTU clustering threshold: Tapolczai, Vasselon, et al., 2019) or on the interpretation of results based the current ecological knowledge of diatoms (Keck et al., 2018; Mortágua et al., 2019; Vasselon et al., 2018). Extensive efforts have also been made to build a comprehensive reference sequence database for freshwater diatoms (Rimet et al., 2019). More recently, analytical approaches aiming to expand the range of possible bioindicators beyond the sole fraction of assigned sequences, i.e. the "taxonomy-free" approaches, have been proposed to optimize the use of diatom metabarcoding data for biomonitoring (Apothéloz-Perret-Gentil et al., 2017; Tapolczai, Keck, et al., 2019). Finally, a recent study proposed using machine learning algorithms to predict diatom communities and infer water quality based on the divergence of the genetic community compared to those on newly defined reference sites (Feio et al., 2020).

In this study, we compared the efficiency of two taxonomic markers (*rbcL* and 18S V4) and three analytical approaches for the inference of the MDI, in reference to the morphology-based Swiss Diatom Index (DI-CH). We used 112 biofilm river samples collected within the SYNAQUA project (Lefrançois et al., 2018). We employed (a) a taxonomy-based approach to compute the MDI values in a same manner as for morphological approach (Tax-Assign); (b) a taxonomy-free, indicator value approach, that calibrates *de novo* the ecological optimum and tolerance of all metabarcodes (Mol-Assign, see Apothéloz-Perret-Gentil et al., 2017); and (c) a supervised machine learning approach, that scrutinizes full community profiles across samples of known disturbance level to train a predictive model (ML) (Figure 1). We analysed the correlation between the MDI values inferred from molecular data and the DI-CH values obtained from morphological data for the same samples. We ranked the different combinations of marker-analytical approaches depending on their congruence with the morphological index and discuss the pros and cons of each approach.

**FIGURE 1** Schematic representation of the workflow used in this study



## 2 | MATERIALS AND METHODS

### 2.1 | Sampling

As part of the INTERREG SYNAQUA project (Lefrançois et al., 2018), 112 samples of biofilms from rivers were collected in 2017. The sampling campaigns were carried by the authorized authorities in both France (80 samples) and Switzerland (32 samples), these sites are part of the national surveillance networks (Table S1). Epilithic biofilm samples were obtained by scraping the surface of 3–5 stones per site following the official recommendations for each country (CH: Hürlimann & Niederhauser, 2007; FR: AFNOR, 2014a). Half of each sample was used for morphological analysis and the other half for molecular analysis. All biofilm samples were preserved by adding 99% molecular grade ethanol for a final ethanol concentration > 70%, except the samples collected for Swiss morphological identification, for which a subsample was preserved in 4% formalin according to the Swiss legislation.

### 2.2 | DNA extraction

For French samples, 2 ml of biofilm was centrifuged at 13,000 rpm for 30 min at 4°C, the supernatant containing ethanol was removed and the pellet used as starter for DNA extraction. Samples were then extracted using the DNA extraction kit Macherey-Nagel NucleoSpin Soil kit following manufacturer recommendation starting with SL1 solution and with a final elution of 30 µl with the solution provided in the kit, as described in Vasselon, et al. (2017). For Swiss samples, DNA extraction was performed using 1 ml of the preserved sample. After centrifugation at 3'500 g during 5 min, the supernatant containing ethanol was removed and the pellet used as starter for DNA extraction. Each sample was extracted using the DNeasy PowerBiofilm kit (Qiagen) according to the manufacturer's instructions, as described in Apothéloz-Perret-Gentil et al. (2017) in a final elution of DNA in 100 µl with the solution provided in the kit. All DNA extracts were then stored at -20°C until PCR amplifications.

### 2.3 | PCR amplification and high-throughput sequencing (HTS)

The DNA extracts were amplified by PCR targeting two taxonomic markers: the barcoding fragment of chloroplastic *rbcL* gene and the V4 region of the nuclear 18S gene. The primers corresponding to both markers were optimized for diatoms. Primers and PCR conditions for *rbcL* and 18S V4 were described in Vasselon, et al. (2017) and Visco et al. (2015), respectively. 35 cycles were performed for the 18S marker using the FastStart Taq DNA Polymerase (Roche Applied Science) in a final volume of 25 µl and 33 cycles using the TaKaRa LA Taq DNA Polymerase in a final volume of 25 µl for *rbcL* marker. For each sample, three PCR reactions were performed. A negative control was performed for each sample.

For 18S amplifications, tagged primers bearing eight nucleotides attached at each primer's 5'-extremity were used to enable multiplexing of all PCR products in a unique sequencing library (Esling et al., 2015). All PCR replicates were then pooled and quantified with capillary electrophoresis using QIAxcel instrument (Qiagen). Equimolar concentrations of PCR products were pooled into three libraries and purified using High Pure PCR Product Purification kit (Roche Applied Science). The libraries preparation were performed using Illumina TruSeq DNA PCR-Free Library Preparation Kit. 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 standard kit v3.

For *rbcL* amplifications, to enable the sequencing of all samples in a single Illumina run, two successive PCR were performed to prepare HTS libraries. As described in Keck et al. (2018), half of the Illumina adapters were included to the 5' end of *rbcL* primers for the first amplification. The three pooled PCR were then sent to the "GenoToul Genomics and Transcriptomics" facility (GeT-PlaGe) where amplicons were purified and used as templates for the second PCR which used Illumina-tailed primers targeting the half of Illumina adapters used in the first PCR. Finally, all generated amplicons were dual indexed and pooled into a single tube. Final pool was sequenced on an Illumina MiSeq platform using the V2 paired-end sequencing kit (250 bp × 2).

Raw FASTQ reads were demultiplexed to retrieve the R1 and R2 fastq files for each sample using the demultiplexer module implemented in SLIM (Dufresne et al., 2019). Quality filtering, removal of chimera and generation of the amplicon sequence variant (ASV) table were done using dada2 R package v.1.10.1 (Callahan et al., 2016). For taxonomic assignment, DECIPHER R package v.2.10.2 (Wright, 2016) was used with the v7 version of the Diat.barcode database (Rimet et al., 2019) and a confidence threshold of 60 (i.e., minimal proportion of bootstrap replicates that yielded a given taxonomic label, see Murali et al., 2018). To investigate the proportion of nondiatom sequences amplified with the two set of primers, we performed BLAST analyses against the GenBank database with 80% of identity on representative ASV and calculate the proportion of diatom, algae and other sequences for each site.

## 2.4 | Morphological analysis

In the case of Swiss samples, the preparation of diatoms slides for microscopic observation was performed as recommended by the protocol of Swiss Federal Office for the Environment (Hürlimann & Niederhauser, 2007). The samples collected in France were processed according to the European Standard (AFNOR, 2014b) and the French Standard NFT 0-354 (AFNOR, 2016). The Swiss and French protocols are compatible with each other. At least 400 valves per sample were counted and identified mainly with the bibliographic support of The Flora of Diatoms (Krammer & Lange-Bertalot, 1986), 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). The DI-CH values were calculated for both Swiss and French sites with the generated species lists according to the Swiss guidance (Hürlimann & Niederhauser, 2007). The DI-CH index uses two types of ecovalues that correspond to the optimum condition (*D*-value, ranges from 1 to 8 with a step of 0.5, from 1 for good condition to eight for bad condition) and a weighting factor (*G*-value: 0.5, 1, 2, 4 or 8) which correspond to the environmental tolerance of each species. Species with a high *G*-value will be very representative of an ecological status. The calculation of the index follows the weighted average equation of Zelinka and Marvan (1961) using the relative frequency of each species in the sample. This index 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: poor; 6.5–8: bad) (see Hürlimann & Niederhauser, 2007 for more details).

## 2.5 | Molecular index calculation

Three analytical approaches to infer the MDI values were performed on the molecular data sets obtained from the two taxonomic markers. For all the approaches, the counts of reads of each ASV table were normalized using the cumulative-sum scaling method (CSS) implemented in the metagenomeSeq R package v.1.22.0 (Paulson

et al., 2013). First, for the Tax-Assign approach, CSS normalized abundance was used with the ecovalues related to morphospecies registered in Hürlimann and Niederhauser (2007) to calculate the MDI, only species level assignment were used since the DI-CH gives ecovalues only to species and not genus level. The taxonomic assignment was obtained using curated Diat.barcode database (Rimet et al., 2019) for rbcL and Genbank database for 18S. For the Mol-Assign approach, we used the occurrence distribution and the CSS normalized abundance across all samples to calculate optimum and tolerance values for each ASV as described in Apothéloz-Perret-Gentil et al. (2017). Samples were classified by their morphological DI-CH value rounded to 0.5 (according to the range of the *D*-value) and the relative frequency of each ASV was plotted for each class. The class with the highest 95e percentile value was used as optimum value (corresponding to morphological *D*-value). For the tolerance value, samples were classified by their morphological DI-CH value rounded to the unit and the relative frequency of each ASV was plotted for each class. The tolerance value was determined given the distribution of the ASV across classes. Then the distribution of 80% of the total abundance of the ASV was used to determine the weighting factor (i.e., 80% in the extreme classes (1:3, corresponding to very good or 7–8, corresponding to bad status), the strongest weighting factor was given (value 8). On the opposite, if the 80% were distributed across more than three classes, the lowest weighting factor was given (value 0.5), see graphical example in supporting information from Apothéloz-Perret-Gentil et al. (2017). Finally, for the ML approach, we trained predictive models using the Random Forest algorithm (Breiman, 2001) implemented in the ranger R package v.0.11.2 (Wright & Ziegler, 2017) as described in Cordier et al. (2017). For both of the tested taxonomy-free approach (Mol-Assign and ML), a leave-one-sample-out cross validation procedure was performed to assess their accuracy.

## 3 | RESULTS

### 3.1 | Morphological analysis

The number of taxa identified to species level at least ranged from 8 to 56 in all samples, with a median value at 24 (Table S2). This list, in addition with the relative abundance of each species per site, was used to calculate the DI-CH. The index values ranged from 1.18 to 6.75 (eight being the worst possible value). A total of 60, 31, 13, four and three samples belonged to the very good, good, average, poor and bad quality status classes, respectively, following the classification given by the DI-CH (Table S3).

### 3.2 | Sequence data

The molecular data were obtained from two independent Illumina MiSeq runs, one for each taxonomic marker. Details of the filtration statistics during sequence data processing are given in Table S4.

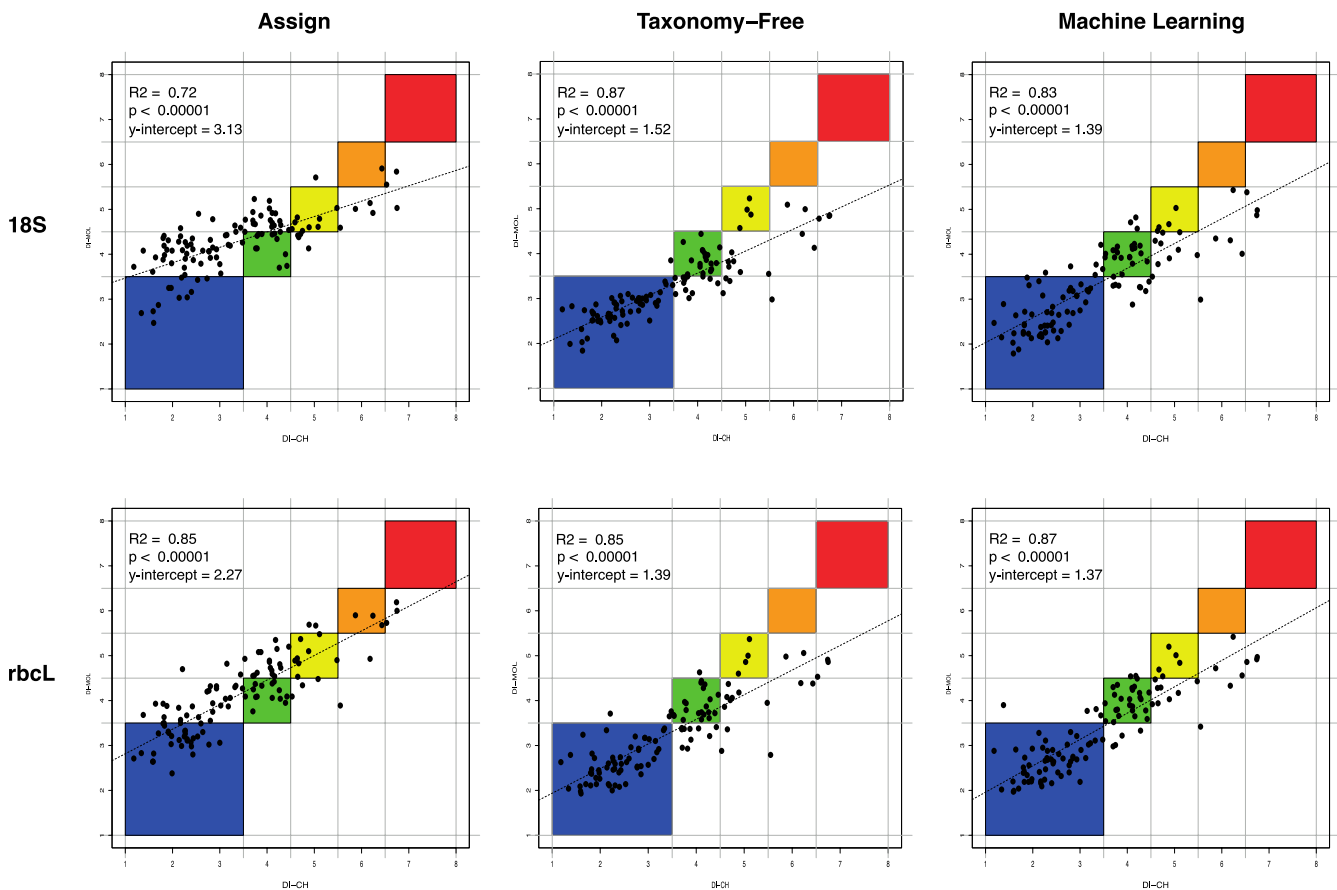
For the *rbcl* marker, a total of 3,245,094 high-quality reads distributed across 112 samples were obtained, ranging from 10,127 to 54,559 reads per sample with a median value around 30,000 sequences. The proportion of diatoms, other algae and non-algal eukaryotes sequences were investigated using a BLAST analysis against the GenBank database (Figure S1, Table S5). The median values are 82% of diatoms, 1% of other algae and 15% of other eukaryotes. For more accurate identification of diatoms, the assignment was also performed on the curated Diat.barcode database (Rimet et al., 2019). The percentage of assigned sequences to species level (only diatom species) ranged from 22% to 99% per sample with a median value at 90% and the number of assigned taxa ranged from 12 to 64 per sample with a median value at 30 (Table S6).

For the 18S marker, a total of 6,568,535 high-quality reads distributed across 111 samples were obtained. The CH12 sample failed to amplify for the 18S marker and therefore no sequences were retrieved from it. The total number of sequences per sample ranged from 12,038 to 255,004 with a median value around 43,000 sequences. The proportion and taxonomic composition of diatoms, other algae and nonalgal eukaryotes were also investigated using the GenBank database (Figure S1, Table S5). The median values are

94% of diatoms, 5% of other algae and 0.4% of other eukaryotes. Assignment with the Diat.barcode database showed that the percentage of assigned sequences to species level (diatom and other algae) ranged from less than 1% to 99% per sample with a median value at 72% and the number of assigned taxa ranged from 2 to 48 per sample with a median value at 24 (Table S7). Raw ASV table with representative sequences and taxonomic assignment with the confidence threshold are given in Table S8 and S9 for *rbcl* and 18S markers respectively.

### 3.3 | Indices comparison

All inferred index values (Tax-Assign, Mol-Assign and ML) based on each taxonomic marker (*rbcl* and 18S) were compared to the morphological DI-CH values, considered as ground-truth (Figure 2). The R square value ranged from 0.72 for Tax-Assign-18S to 0.87 for Mol-Assign-18S and ML-*rbcl*. The y-intercept values for all taxonomy free methods were around 1.38, except a slightly higher value for Mol-Assign-18S (1.52). However, this value diverges more greatly from the ideal 1:1 slope for the Tax-Assign method with 2.27 and 3.13 values for *rbcl* and 18S markers respectively.



**FIGURE 2** Scatter plots showing 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 a dashed line, and the R<sup>2</sup>, p-value and y-intercept value are indicated for each plot

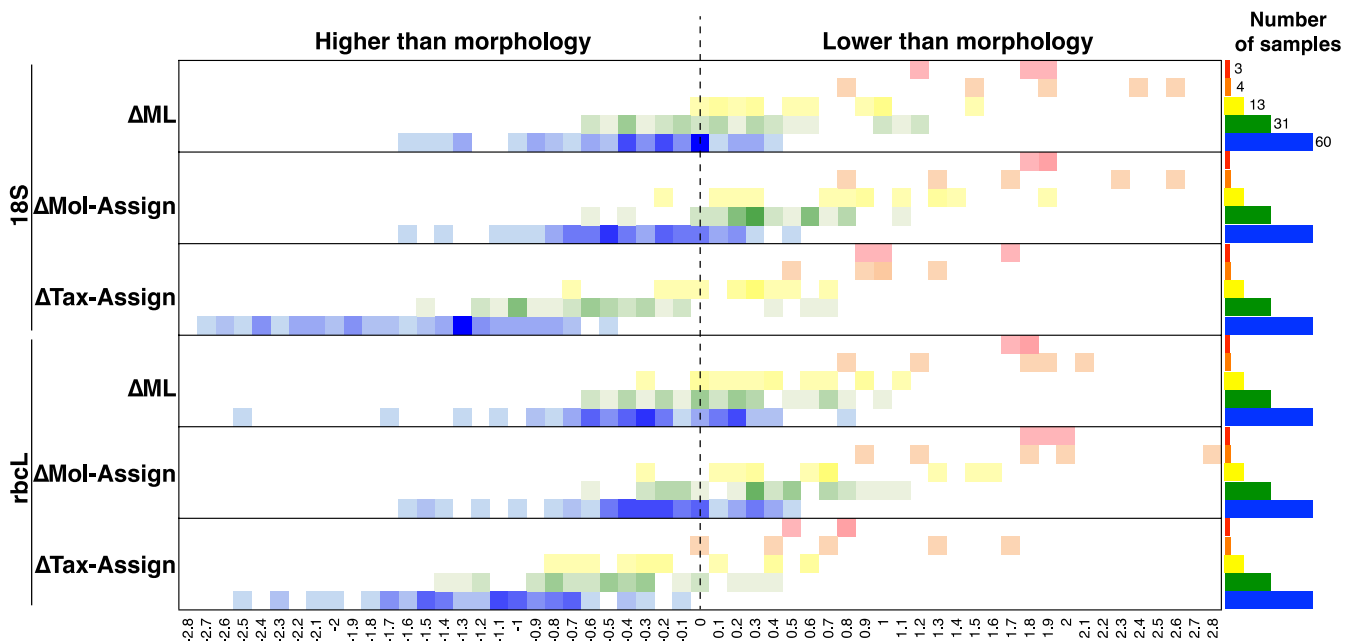
The divergence of each MDI value from the reference morphological DI-CH was plotted separately for each quality classes in the Figure 3. The heat maps showed a clear difference between the Tax-Assign approach and both taxonomy-free approaches. For the Tax-Assign approach, very good quality sites are always overestimated, meaning that the molecular index values are higher than the morphological ones (leading to worse ecological status with the Tax-Assign than with the morphology). The bad sites deviate less from the morphological value in the Tax-Assign method compared to taxonomy-free approaches, especially in the case of the *rbcl* marker. Both taxonomy-free approaches gave very similar results for both markers, providing slightly higher values for very good sites and lower values for poor and bad sites, compared to the morphological index.

When comparing the three analytical approaches to the reference morphological assessments, we observe relatively small differences for very good, good and average quality classes. In median, Tax-Assign method differs from morphology by less than or equal to 1 point in all classes except the very good class. For both taxonomic markers, it even differs by less than 0.5 point in the average class. The Mol-Assign approach differs from morphology by less than 1 point in very good, good and average classes and even less than 0.5 point in very good and good classes. The ML approach showed the stronger correlation to morphology with a difference by less than or equal to 0.5 point for all three first classes. However, the difference is much higher in the case of poor and bad classes, especially for both

taxonomy-free approaches that differ up to 1.7 to 1.9 point, which may lead to a difference of two quality classes in the bioassessment.

## 4 | DISCUSSION

Our study shows that, among the three tested analytical molecular approaches, the two which are taxonomy-free provide correlations closer to reference morphotaxonomy assessments than the taxonomy-based approach, for both tested markers. Although the Tax-Assign approach provided a good correlation with reference using the *rbcl* marker, the slope was more distant from the ideal 1:1 than with taxonomy-free approaches, even when the correction factor based on cell biovolume proposed by Vasselon et al. (2018) was applied (Figure S2). However, the congruence between molecular indices and morphological reference index differs depending on quality classes. We observed that taxonomy-free approaches tend to improve the ecological status of poor quality sites, meaning that the inferred value were often lower than the morphological ones, as opposed to the taxonomy-based approach that shows higher correlation to DI-CH for those sites (especially with *rbcl* marker). Conversely, taxonomy-free approaches are more congruent with morphological index in very good and good quality sites where the taxonomy-based approach tends to overestimate the DI-CH values, meaning that the inferred value were often higher than the morphological ones (leading to a worse ecological status). This observation



**FIGURE 3** Heat maps showing the difference in the value of the morphological DI-CH and the one predicted by the molecular methods. Each line corresponds to the heat map of one of the three methods (ML, Mol-Assign and Tax-Assign) and one of the two markers (18S and *rbcl*). Negative value means that the molecular index gave a value higher than the morphology (ecological status worse than morphology) while positive value means that the molecular index gave a value lower than the morphology (ecological status better than morphology). Next to the heat maps, the bar plots indicate the number of sites in each quality status class given by the morphological assessment, following the DI-CH guidance (blue, very good; green, good; yellow, average; orange, poor; red, bad)



may stem from the unbalanced sampling of possible ecological quality classes within our data set. We had only few sites of poor ecological quality status, which probably impact the capacity of both taxonomy-free approaches to unravel statistically meaningful associations between poor-quality status and ASVs profiles (Mol-assign) or community structure (ML).

It is also important to note that the community of diatom is usually structured differently between poor and good quality sites. Poor ecological quality sites are generally represented by few very abundant opportunistic diatom species (Stevenson et al., 2010) while very good ecological quality sites are rather represented by communities with medium to high diversity (Hürlimann & Niederhauser, 2007; Whitton et al., 1991), except for very good quality sites in alpine rivers where the diversity is very low in Switzerland.

The structure of diatom community also has a great impact on the success of taxonomy-free methods. Sites with high diversity need to have a greater representation in the training data set to take into account the whole assemblage and the presence of the different species. It is interesting to note that in previous study (Apothéloz-Perret-Gentil et al., 2017) very good sites were not well predicted with the taxonomy-free approach, while in the present study the predictions were much more accurate. This could be explained by the fact that very good sites represent more than 50% of the training data set in this study, against less than 20% in the previous study. This result shows the importance of the training data set, which was very poorly represented for bad quality classes in this study. Indeed, if during the cross-validation process some species with high abundance on poor sites are not represented in the training data set, a lower abundance of this species on other sites will force the optima of the species to better values than if the species was also represented in poor sites.

Interestingly, when the two genetic markers are compared, a general slightly better correlation with morphological reference is observed with *rbcl* marker. It is possible that this is due to the fact that the *rbcl* marker is more taxonomically resolutive and the distinction of diatom and other species is more accurate as shown previously by Kermarrec et al. (2013). As shown by our BLAST results (Figure S1) it does not seem that the *rbcl* primers are more specific than the 18S V4 primers. However, it is important to note that most of unassigned hits in *rbcl* BLAST data sets could possibly also belong to diatoms but are not identified as such. This could be due to lack of higher-level taxonomic signal in *rbcl* barcode as well as to the high level of unassigned environmental sequences in the GenBank database. The hypothesis that *rbcl* primers are very specific to diatoms is supported by the assignment with the curated Diat.barcode database for which the amount of sequences assigned to diatom (not necessarily to species level) range between 37% and 99% with a median value of 93%.

An important advantage of using *rbcl* is its comprehensive reference database, which is much more developed than the one for the 18S marker (Rimet et al., 2019). Both the completeness of the *rbcl* database, and its high level of curation certainly explain the higher correlation of the taxonomy-based approach using this marker. Thus,

*rbcl* represents so far the ideal candidate for an implementation of metabarcoding methods for routine rivers monitoring, because the generated species lists are more exhaustive than the ones generated by targeting the 18S marker. A significant congruence in these taxonomic inventories is indeed very important to assure a backward and forward compatibility of diatom-based monitoring time series. Noteworthy, it will also support a smoother transition between arduous and labor-intensive morphological methods with faster and cost-effective molecular ones. Such methodological shift is strongly advocated and anticipated, to ensure a more continuous vision on sites subject to regulatory biomonitoring or a better monitoring of sites subject to restoration actions.

In conclusion, our study confirms the usefulness of diatom metabarcoding as a tool for the assessment of rivers ecological status that give results in line with the currently applied morphological approach. It shows that the taxonomy-free approaches perform as well or better than those based on taxonomic assignment of metabarcoding data when compared to morphological analyses used as reference. Our results also emphasize the importance of well-curated reference sequences database. Upon the sustained efforts to complete such reference databases, as in the case of *rbcl*, the taxonomy-based analytical approach can provide results similar to those of the taxonomy-free approaches. Finally, the performance of taxonomy-free approaches in recovering the reference values is highly depending on the balanced coverage of sites of contrasting ecological status.

Taken together, our results highlight the need to sustain ongoing efforts to build comprehensive reference databases. Such databases would be either composed of curated reference sequences for taxonomy-based approaches or composed of data sets containing both metabarcoding data and independently established ecological quality status for taxonomy-free approaches. In the future, it seems important that the taxonomy-free methods are benchmarked directly on environmental variables. At present, the traditional morphotaxonomic assessments are used as references to ecological quality status. However, such methods have also its own biological and technical biases and limits. Using environmental variables could help overcoming these biases and improve the sensitivity of taxonomy-free approaches.

The further developments of molecular diatom indices could also take advantage of the ecological information that may be embedded within high resolution DNA variants (e.g., ASVs). Indeed, diatom cryptic diversity revealed by metabarcoding could be leveraged for the establishment of new subspecies molecular bioindicators. Being more sensitive to small environmental variations would make metabarcoding even more attractive as a tool for future environmental biomonitoring.

Diatom metabarcoding offers many advantages in terms of cost and time effectiveness compared to the traditional approach. However, it also has some limitations, related to the incompleteness of reference databases, but also to the lack of standardized protocols. Solving these issues is of key importance for the implementation of diatom metabarcoding in routine monitoring.

## ACKNOWLEDGEMENTS

The SYNAQUA project was supported by the European Cross-Border Cooperation Program (Interreg France-Switzerland 2014–2020) and has thus received a European (European Regional Development Fund) and a Swiss Federal grant covering respectively 60% of the French total cost and 29% of the Swiss total cost. Funding was also provided by Swiss cantons (Valais, Geneva, Vaud). This work also benefitted from the support and discussion of other participants in SYNAQUA project, including Alina Pawlowska, Estelle Lefrançois, Philippe Blancher, Régis Vivien, Benoit Ferrari, and others. LAPG, TC and JP were supported by the Swiss National Science Foundation (grant 31003A\_179125). The authors also thank DNAqua-net COST Action CA15219 “Developing new genetic tools for bioassessment of aquatic ecosystems in Europe” funded by the European Union, for helpful discussions. We also thank Sonia Lacroix, Cécile Chardon and Louis Jacas for the DNA extraction, PCR, and library preparation of the French samples.


## AUTHOR CONTRIBUTION

L.A.P.G., A.B., F.R., J.P. conceived and design the experiment. L.A.P.G. processed Swiss samples and analysed all data with the help of T.C. for the ML algorithms. A.C. performed all the sampling and morphological work for Swiss samples. J.G., V.V., F.R., and A.B. collected and processed French samples. L.A.P.G., and J.P. took the lead in writing the manuscript. All authors provided critical feedback and helped to shape the final version of the manuscript.

## DATA AVAILABILITY STATEMENT

The raw sequencing data is available at the Short Read Archive public database under the accession PRJNA629002.

## ORCID

Laure Apothéloz-Perret-Gentil  <https://orcid.org/0000-0002-8592-3079>

## REFERENCES

- AFNOR (2014a). EN 13946 – Water quality – Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers. AFNOR.
- AFNOR (2014b). EN 14407 – Water quality – Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters | Engineering 360. AFNOR.
- AFNOR (2016). NF T90–354 | Qualité de l'eau – Échantillonnage, traitement et analyse de Diatomées benthiques en cours d'eau et canaux. AFNOR.
- 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*, 17(6), 1231–1242. <https://doi.org/10.1111/1755-0998.12668>
- Aylagas, E., Borja, Á., & Rodríguez-Ezpeleta, N. (2014). Environmental status assessment using DNA metabarcoding: Towards a genetics based marine biotic index (gAMBI). *PLoS One*, 9(3), e90529. <https://doi.org/10.1371/journal.pone.0090529>
- Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., Irigoien, X., Danovaro, R., & Rodríguez-Ezpeleta, N. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Marine Pollution Bulletin*, 114(2), 679–688. <https://doi.org/10.1016/j.marpolbul.2016.10.050>
- Bailet, B., Bouchez, A., Franc, A., Frigerio, J.-M., Keck, F., Karjalainen, S.-M., Rimet, F., Schneider, S., & Kahlert, M. (2019). Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe freshwater and consequences for ecological status. *Metabarcoding and Metagenomics*, 3, 21–35. <https://doi.org/10.3897/mbmg.3.34002>
- Bista, I., Carvalho, G. R., Walsh, K., Seymour, M., Hajibabaei, M., Lallias, D., Christmas, M., & Creer, S. (2017). Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications*, 8, 14087. <https://doi.org/10.1038/ncomms14087>
- Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wild-life biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29(6), 358–367. <https://doi.org/10.1016/j.tree.2014.04.003>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583. <https://doi.org/10.1038/nmeth.3869>
- Chariton, A. A., Stephenson, S., Morgan, M. J., Steven, A. D. L., Colloff, M. J., Court, L. N., & Hardy, C. M. (2015). Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental Pollution*, 203, 165–174. <https://doi.org/10.1016/j.envpol.2015.03.047>
- Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environmental Science & Technology*, 51(16), 9118–9126. <https://doi.org/10.1021/acs.est.7b01518>
- Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., & Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, 18(6), 1381–1391. <https://doi.org/10.1111/1755-0998.12926>
- Cordier, T., Frontalini, F., Cermakova, K., Apothéloz-Perret-Gentil, L., Treglia, M., Scantamburlo, E., Bonamin, V., & Pawlowski, J. (2019). Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Marine Environmental Research*, 146, 24–34. <https://doi.org/10.1016/j.marenvres.2018.12.009>
- Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21), 5872–5895. <https://doi.org/10.1111/mec.14350>
- Directive 2000, 60, EC (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 L*, 327, 1–72.
- Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R., Andruszkiewicz, E. A., Olesin, E., Hubbard, K., Montes, E., Otis, D., Muller-Karger, F. E., Chavez, F. P., Boehm, A. B., & Breitbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, 11(1), 1–9. <https://doi.org/10.1038/s41467-019-14105-1>
- Dufresne, Y., Lejzerowicz, F., Apothéloz-Perret-Gentil, L., Pawlowski, J., & Cordier, T. (2019). SLIM: A flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinformatics*, 20(1), 88. <https://doi.org/10.1186/s12859-019-2663-2>



- Egan, S. P., Grey, E., Olds, B., Feder, J. L., Ruggiero, S. T., Tanner, C. E., & Lodge, D. M. (2015). Rapid molecular detection of invasive species in ballast and harbor water by integrating environmental DNA and light transmission spectroscopy. *Environmental Science & Technology*, 49(7), 4113–4121. <https://doi.org/10.1021/es5058659>
- Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, 43(5), 2513–2524. <https://doi.org/10.1093/nar/gkv107>
- Feio, M. J., Serra, S. R. Q., Mortágua, A., Bouchez, A., Rimet, F., Vasselon, V., & Almeida, S. F. P. (2020). A taxonomy-free approach based on machine learning to assess the quality of rivers with diatoms. *Science of the Total Environment*, 722, 137900. <https://doi.org/10.1016/j.scitotenv.2020.137900>
- Frühe, L., Cordier, T., Dully, V., Breiner, H.-W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2020). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology*, 1–19. <https://doi.org/10.1111/mec.15434>
- Hofmann, G., Werum, M., & Lange-Bertalot, H. (2011). *Diatomeen im Süßwasser-Benthos von Mitteleuropa: Bestimmungsflores Kiesalgen für die ökologische Praxis; 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. État de l'environnement n° 0740* (p. 132 p.). Office fédéral de l'environnement.
- Keck, F., Vasselon, V., Rimet, F., Bouchez, A., & Kahlert, M. (2018). Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. *Molecular Ecology Resources*, 18(6), 1299–1309. <https://doi.org/10.1111/1755-0998.12919>
- 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>
- Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators*, 85, 1044–1057. <https://doi.org/10.1016/j.ecolind.2017.11.014>
- Kelly, M., Boonham, N., Juggins, S., Killie, P., Mann, D., Pass, D., Sapp, M., Sato, S., & Glover, R. (2018). A DNA based diatom metabarcoding approach for water framework directive classification of rivers. (No. SC140024/R). Environment Agency.
- Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J.-F., & Bouchez, A. (2014). A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, 33(1), 349–363. <https://doi.org/10.1086/675079>
- 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>
- Krammer, K., & Lange-Bertalot, H. (1986). *Bacillariophyceae*. Gustav Fischer Verlag.
- Lange-Bertalot, H. (2001). *Diatoms of the European inland waters and comparable habitats* (Vol. 2–4). ARG Gantner Verlag KG.
- 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*. Koeltz Scientific Books.
- Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. *Molecular Ecology*, 25(17), 4392–4406. <https://doi.org/10.1111/mec.13761>
- Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lear, G., & Pochon, X. (2018). A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. *Marine Pollution Bulletin*, 127, 97–107. <https://doi.org/10.1016/j.marpolbul.2017.11.042>
- Lefrançois, E., Apothéloz-Perret-Gentil, L., Blancher, P., Botreau, S., Chardon, C., Crepin, L., Cordier, T., Cordonier, A., Domaizon, I., Ferrari, B. J. D., Guéguen, J., Hustache, J.-C., Jacas, L., Jacquet, S., Lacroix, S., Mazenq, A.-L., Pawlowska, A., Perney, P., Pawlowski, J., ... Bouchez, A. (2018). Development and implementation of eco-genomic tools for aquatic ecosystem biomonitoring: The SYNAQUA French-Swiss program. *Environmental Science and Pollution Research International*, 25(34), 33858–33866. <https://doi.org/10.1007/s11356-018-2172-2>
- Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D., & Pawlowski, J. (2015). High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports*, 5, 13932. <https://doi.org/10.1038/srep13932>
- Mortágua, A., Vasselon, V., Oliveira, R., Elias, C., Chardon, C., Bouchez, A., Rimet, F., João Feio, M., & F.P. Almeida, S. (2019). Applicability of DNA metabarcoding approach in the bioassessment of Portuguese rivers using diatoms. *Ecological Indicators*, 106, 105470. <https://doi.org/10.1016/j.ecolind.2019.105470>
- Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, 6(1), 140. <https://doi.org/10.1186/s40168-018-0521-5>
- Paulson, J. N., Pop, M., & Bravo, H. C. (2013). metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. *Bioconductor Package*, 1, 191.
- Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: Assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources*, 14(6), 1129–1140. <https://doi.org/10.1111/1755-0998.12261>
- Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M. J., Filipe, A. F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Iwan Jones, J., Sagova-Mareckova, M., Moritz, C., ... Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *The Science of the Total Environment*, 637–638, 1295–1310. <https://doi.org/10.1016/j.scitotenv.2018.05.002>
- Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., Wales, N. A., Carøe, C., Campos, P. F., Schmidt, A. M. Z., Gilbert, M. T. P., Hansen, A. J., Orlando, L., & Willerslev, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660), 20130383. <https://doi.org/10.1098/rstb.2013.0383>
- Pérez-Burillo, J., Trobajo, R., Vasselon, V., Rimet, F., Bouchez, A., & Mann, D. G. (2020). Evaluation and sensitivity analysis of diatom DNA metabarcoding for WFD bioassessment of Mediterranean rivers. *Science of the Total Environment*, 138445. <https://doi.org/10.1016/j.scitotenv.2020.138445>
- Reichardt, E. (1999). Zur Revision der Gattung Gomphonema: Die Arten um *G. affine/insigne*, *G. angustatum/micropus*, *G. acuminatum* sowie gomphonemoides Diatomeen aus dem Oberoligozän in Böhmen. A.R.G. Gantner.
- Rimet, F., & Bouchez, A. (2012). Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. *Knowledge and Management of Aquatic Ecosystems*, 406(01), 1–12. <https://doi.org/10.1051/kmae/2012018>
- Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., Mann, D. G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., & Bouchez, A. (2019). Diat.barcode, an open-access curated

- barcode library for diatoms. *Scientific Reports*, 9(1), 1–12. <https://doi.org/10.1038/s41598-019-51500-6>
- Rivera, S. F., Vasselon, V., Bouchez, A., & Rimet, F. (2020). Diatom metabarcoding applied to large scale monitoring networks: Optimization of bioinformatics strategies using Mothur software. *Ecological Indicators*, 109, 105775. <https://doi.org/10.1016/j.ecolind.2019.105775>
- Stevenson, R. J., Pan, Y. D., & van Dam, H. (2010). *Assessing environmental conditions in rivers and streams with diatoms* (2nd ed., pp. 57–85). Retrieved from: CABDirect2. The Diatoms: Applications for the Environmental and Earth Sciences.
- Swiss Federal Council (1998). *Waters Protection Ordinance*. Swiss Federal Council. Retrieved from <https://www.admin.ch/opc/en/classified-compilation/19983281/index.html>, 814.201.
- Taberlet, P., Bonin, A., Coissac, E., & Zinger, L. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press.
- Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, 21(8), 1789–1793. <https://doi.org/10.1111/j.1365-294X.2012.05542.x>
- Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., & Vasselon, V. (2019). Diatom DNA metabarcoding for biomonitoring: Strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. *Frontiers in Ecology and Evolution*, 7, 409. <https://doi.org/10.3389/fevo.2019.00409>
- Tapolczai, K., Vasselon, V., Bouchez, A., Stenger-Kovács, C., Padisák, J., & Rimet, F. (2019). The impact of OTU sequence similarity threshold on diatom-based bioassessment: A case study of the rivers of Mayotte (France, Indian Ocean). *Ecology and Evolution*, 9(1), 166–179. <https://doi.org/10.1002/ece3.4701>
- Thomsen, P. F., Kielgast, J., Iversen, L. L., Møller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One*, 7(8), e41732. <https://doi.org/10.1371/journal.pone.0041732>
- Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.-M., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4), 929–942. <https://doi.org/10.1111/mec.13428>
- Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., Tapolczai, K., & 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. <https://doi.org/10.1086/690649>
- 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. <https://doi.org/10.1016/j.ecolind.2017.06.024>
- 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>
- Whitton, B. A., Rott, E., & Friedrich, G. (1991). Use of algae for monitoring rivers. *Journal of Applied Phycology*, 3(3), 287.
- Wright, E. S. (2016). Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal*, 8(1), 352. <https://doi.org/10.32614/RJ-2016-025>
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17. <https://doi.org/10.18637/jss.v077.i01>
- Zaiko, A., Samuiloviene, A., Ardura, A., & Garcia-Vazquez, E. (2015). Metabarcoding approach for nonindigenous species surveillance in marine coastal waters. *Marine Pollution Bulletin*, 100(1), 53–59. <https://doi.org/10.1016/j.marpolbul.2015.09.030>
- Zelinka, M., & Marvan, P. (1961). *Zur Präzisierung der biologischen Klassifikation der Reinheit fließender Gewässer*, (pp. 389–407). Archiv Für Hydrobiologie.
- 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.

**How to cite this article:** Apothéloz-Perret-Gentil L, Bouchez A, Cordier T, et al. Monitoring the ecological status of rivers with diatom eDNA metabarcoding: A comparison of taxonomic markers and analytical approaches for the inference of a molecular diatom index. *Mol Ecol* 2021;30:2959–2968. <https://doi.org/10.1111/mec.15646>