DR. FRANÇOIS KECK (Orcid ID : 0000-0002-3323-4167)

MR. VALENTIN VASSELON (Orcid ID : 0000-0001-5038-7918)

Article type : Resource Article

Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of OTUs' ecological profiles

F. Keck[1, 2], V. Vasselon[2], F. Rimet[2], A. Bouchez[2], and M. Kahlert[1]

[1] *Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P. O. Box 7050, 750 07 Uppsala, Sweden*

[2] *UMR CARRTEL, INRA, Université Savoie Mont Blanc, F-74200 Thonon, France*

Running title: Phylogenetics for DNA-based biomonitoring

Corresponding Author:

François Keck

Department of Aquatic Sciences and Assessment,

Swedish University of Agricultural Sciences,

P. O. Box 7050, 750 07 Uppsala, Sweden

francois.keck@gmail.com

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12919

Abstract

DNA metabarcoding has been introduced as a revolutionary way to identify organisms and monitor ecosystems. However, the potential of this approach for biomonitoring remains partially unfulfilled because a significant part of the sampled DNA cannot be affiliated to species due to incomplete reference libraries. Thus, biotic indices which are based on the estimated abundances of species in a community and their ecological profiles can be inaccurate. We propose to compute biotic indices using phylogenetic imputation of OTUs' ecological profiles (OTU-PITI approach). Firstly, OTUs sequences are inserted within a reference phylogeny. Secondly, OTUs' ecological profiles are estimated on the basis of their phylogenetic relationships with reference species whose ecology is known. Based on these ecological profiles, biotic indices can be computed using all available OTUs. Using freshwater diatoms as a case study, we show that short DNA barcodes can be placed accurately within a phylogeny and their ecological preferences estimated with a satisfactory level of precision. In light of these results, we tested the approach with a dataset of 139 environmental samples of benthic river diatoms for which the same biotic index (IPS) was calculated using (i) traditional microscopy, (ii) OTUs with taxonomic assignment approach, (iii) OTUs with phylogenetic estimation of ecological profiles (OTU-PITI), and (iv) OTU with taxonomic assignment completed by the phylogenetic approach (OTU-PITI) for unclassified OTUs. Using traditional microscopy as a reference, we found that the combination of the OTUs' taxonomic assignment completed by the phylogenetic method performed satisfactorily and substantially better than the other methods tested.

Keywords: Metabarcoding, Biomonitoring, Environmental DNA, Diatoms, Phylogenetic signal

Introduction

The protection and conservation of ecosystems requires managers to accurately assess the quality of the environment over time (Ibáñez, Caiola, Sharpe, & Trobajo, 2010). Ecologists have developed a wide set of biotic indices to monitor ecological impacts of human activities, based on the principle that anthropic pressures shape biological communities (Chapman, 1996). Hence, a large variety of indices are available to estimate environmental quality from the richness, diversity, structure, and functioning of biological communities.

Diatoms are unicellular eukaryotic algae encompassing a large taxonomic diversity (Round, Crawford, & Mann, 1990). Because they have a relatively short generation time and their communities respond strongly to changes in habitat quality, diatoms are recognized as powerful bioindicators of freshwater quality (Rimet, 2012; Stevenson, Yangdong, & Van Dam, 2010). Most of the diatom biotic indices are based on species autecology and are usually derived from the equation of Zelinka and Marvan (1961). For example, the IPS index (specific sensitivity value; Coste, 1982) which is used in this paper as a case study is defined in Equation 1, where a_i is the relative abundance of species i in the sample, $IPSV_i$ its indicator value (tolerance) and $IPSS_i$ its pollution sensitivity (optimum).

$$IPS = \frac{\sum_{i=1}^{n} a_i \times IPSV_i \times IPSS_i}{\sum_{i=1}^{n} a_i \times IPSV_i} \quad (1)$$

The estimation of diatom indices like the IPS index require an accurate taxonomic inventory of the community (Besse-Lototskaya, Verdonschot, Coste, & Van de Vijver, 2011). Diatom inventories are traditionally based on the morphological identification of several hundred individuals under microscope (Prygiel et al., 2002). Given the diversity of diatoms (Mann & Vanormelingen, 2013), this step is time-consuming, requires highly-qualified staff, and is prone to errors (Besse-Lototskaya, Verdonschot, & Sinkeldam, 2006). However, the development of methods to identify multiple taxa simultaneously from an environmental sample with standard genetic markers (DNA metabarcoding),

combined with high-throughput sequencing technologies (HTS) have enabled the production of fast and cost-effective taxonomical inventories of communities (Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012). Therefore, metabarcoding has been promoted as an attractive alternative to traditional identification using microscopy for biomonitoring (Baird & Hajibabaei, 2012). Recent studies have shown that molecular inventories of diatom communities can be used to calculate various biotic indices (Kermarrec et al., 2014; Rivera et al., 2017; Vasselon, Rimet, Tapolczai, & Bouchez, 2017; Visco et al., 2015).

The classical approach to compute ecological indices with metabarcoding data comprises of clustering DNA reads into operational taxonomic units (OTUs) and then assigning them a taxonomic name using a reference library (Kermarrec et al., 2014; Zimmermann, Glöckner, Jahn, Enke, & Gemeinholzer, 2015). Once the list of OTUs is converted into a taxonomic list, one can compute traditional bioassessment indices based on the ecological preferences of the species (IPSS and IPSV values in the case of IPS). However, this approach has some notable drawbacks, the most significant being that an important proportion of OTUs cannot be adequately classified into species because reference libraries are incomplete. Hence, a large part of the biological diversity unraveled by DNA methods is discarded and cannot be used for bioassessment purposes.

To circumvent this problem, it has been suggested to skip the conversion from DNA reads to taxonomic entities and work directly on molecular data (Keck, Vasselon, Tapolczai, Rimet, & Bouchez, 2017). In this respect, different strategies have been considered, including OTU-based indices (Apothéloz-Perret-Gentil et al., 2017) or the use of supervised machine-learning algorithms to process genetic inventories (Cordier et al., 2017). Alternatively, Keck *et al.* (2016; 2017) have suggested an approach based on the relationships existing between the phylogenetic position of species and their ecology (i.e., the phylogenetic signal, Blomberg, Garland, & Ives, 2003). The central idea is to combine an algorithm placing OTUs within a reference phylogeny and an algorithm to phylogenetically impute OTUs' ecological profiles (i.e., autecological values, here IPSS and IPSV) based on information available from neighbor species. This type of workflow has been introduced and successfully used by Kembel *et al.* (2012) to estimate 16S gene copy number and improve estimates

of organismal abundance in microbial communities. However, implementing this approach for ecological assessment with metabarcoding requires to assess if DNA reads produced by HTS are long and informative enough to be accurately placed in a reference phylogeny, and once inserted, if then the phylogenetic signal is strong enough to estimate precisely their ecological profiles.

In this paper, we aim to implement and test the phylogenetic approach (termed OTU-PITI; i.e., OTU Phylogenetic Insertion and Trait Imputation) for ecological assessment with metabarcoding data. We first compared the placement accuracy of short *rbcL* DNA reads (312 bp) as produced by HTS technologies with full length sequences of the *rbcL* gene. Second, we tested for the phylogenetic signal and performed a cross-validation procedure to assess whether phylogenetic imputation of species ecological profiles can be estimated from their phylogenetic positions. Finally, we tested the method with a dataset of 139 environmental river samples for which diatoms communities were analyzed using both microscopy and DNA metabarcoding. IPS indices based on the OTU-PITI approach and on taxonomically assigned OTUs were then compared to IPS indices based on classical microscopy.

Material and Methods

Reference phylogenetic tree reconstruction

The reference phylogenetic tree was reconstructed from the chloroplast *rbcL* gene coding for the RuBisCO enzyme. This gene is recognized for its good performances to differentiate diatoms species and is a popular marker both for phylogenetic and metabarcoding studies. However, *rbcL* may have a limited ability to recover deep phylogenetic relationships within diatom clades (Theriot, Ruck, Ashworth, Nakov, & Jansen, 2011). Therefore, we used the phylogeny of diatoms published by Theriot, Ashworth, Nakov, Ruck, & Jansen (2015) and based on seven genes (*SSU*, *atpB*, *psaA*, *psaB*, *psbA*, *psbC* and *rbcL*) as a fixed guide in the reconstruction process. We extracted 1380 *rbcL* sequences from the curated library R-Syst::diatom (Rimet et al., 2016). The sequences were aligned using MUSCLE (Edgar, 2004) and consensus sequences were computed using per-base majority rule

for 550 species. The new set of 550 sequences was then merged and re-aligned against the 208 sequences alignment of Theriot *et al.* (2015). Duplicated species were dropped, giving a final reference alignment of 604 species. The phylogeny was then reconstructed with RAxML 8.2.11 (Stamatakis, 2014) using the phylogenetic tree of Theriot *et al.* (2015) as a topological constraint, a substitution model GTR+G+I, 200 runs and 1000 bootstraps (Fig S1, Supporting Information). The tree was dated in relative time using PATHd8 (Britton, Anderson, Jacquet, Lundqvist, & Bremer, 2007).

Testing for short sequences placement

To test if a 312 bp *rbcL* barcode is sufficient to recover the phylogenetic position of the species, we sequentially dropped species from the reference phylogenetic tree and placed them using their reference barcode of 312 bp. The phylogenetic placement was performed using the Evolutionary Placement Algorithm (EPA; Berger, Krompass, & Stamatakis, 2011) implemented in RAxML. To assess the quality of the barcode placement, we measured the distance between the insertion point of the full-length reference sequence (~1500 bp) edge and the insertion point of the placed barcode sequence (312 bp). This distance is expressed as the number of nodes located on the path which connects the two insertion points. Ideally, the barcode sequence is placed at the same location as the full-length sequence and the node distance is zero.

Testing for phylogenetic estimation of autecological values

We tested the phylogenetic signal for pollution sensitivity values (IPSS) and the indicator values (IPSV) of the species using Pagel's λ (Pagel, 1999) computed with the R package *phylosignal* (Keck, Rimet, Bouchez, & Franc, 2016). We used a leave-one-out cross-validation (LOOCV) procedure to test whether the IPSS and IPSV values can be estimated accurately from their phylogenetic position. The analysis was performed on a subset of 237 species which were found both in the reference phylogenetic tree and the IPS database. We sequentially estimated the IPSS and IPSV values of each species, given its phylogenetic position (as estimated using the barcode sequence; see above), and the known autecological values of the other species in the tree. The prediction was done using the

framework introduced by Bruggeman, Heringa, and Brandt (2009) which estimates the phylogenetic covariance matrix parameters under a given evolution model and use it to impute the missing data as the best linear unbiased predictions (Ho, Si, & Ané, 2014). We used the implementation available in the R package *rphylopars* (Goolsby, Bruggeman, & Ané, 2017) to test 6 different phylogenetic models: Brownian motion (*BM*), Ornstein-Uhlenbeck (*OU*), Early-Burst (*EB*), *lambda*, *delta* and *kappa* (see Goolsby et al., 2017 for details on the tested models). Additionally we used an ad-hoc non-phylogenetic model (*star*) which assumes that the best estimate for a missing value is given by the mean of all observations. The performances of the different models were assessed using the LOOCV mean squared error (MSE) and compared using pairwise Wilcoxon signed-rank test on squared error distributions with Bonferroni correction for multiple comparisons. The absolute error (i.e., the absolute value of the difference between the estimated and the true trait value) was used to investigate model prediction errors in detail.

Sample collection

A total of 139 benthic diatom samples were collected from rivers as part of the 2016 French monitoring campaign for water quality assessment (Fig. S2, Supporting information). Benthic diatoms communities were collected by scraping biofilms from at least 5 submerged stones using a toothbrush, as recommended by the European standard (European Committee for Standardization, 2016). Immediately after collection, each sampled biofilm was homogenized and divided into 2 subsamples to perform the molecular and morphological approaches. Each subsample was transferred into 50 mL Falcon tubes and preserved with a final concentration of at least 70 % of ethanol.

Morphological approach

Sample preparation, species identification and counting were performed by offices responsible for the ecological assessment of French rivers in the context of the Water Framework Directive. Benthic samples were treated using 40 % H_2O_2 and HCl according to the European standard (European Committee for Standardization, 2014). Resulting diatom samples were mounted in Naphrax and used

to obtain permanent slides for analysis by microscopy. A minimum of 400 diatoms valves were determined using standard European floras (European Committee for Standardization, 2014).

Molecular laboratory methods

The preserved biofilm samples were centrifuged at 17,000 g during 30 minutes and the supernatant containing ethanol discarded. Total genomic DNA was extracted from the pellet using a non-commercial method based on Sigma-Aldrich GenEluteTM-LPA DNA precipitation, as described and recommended previously for diatom metabarcoding (Chonova et al., 2016; Vasselon, Domaizon, Rimet, Kahlert, & Bouchez, 2017). In order to have technical replicates, two subsamples of each DNA extracts were used for subsequent PCR amplification and HTS, for a total of 278 DNA samples (139 x 2) sequenced. To enable the sequencing of all samples in a single Illumina run, 2 successive PCR were performed to prepare HTS libraries. (i) PCR1: DNA extracts were amplified in triplicate using the equimolar mixes of Diat_rbcL_708F_1, 708F_2, 708F_3 and R3_1, R3_2 as forward and reverse primers respectively (Vasselon, Rimet, et al., 2017), allowing to focus a short fragment of the *rbcL* plastid gene (312 bp). Half of the P5 (CTTTCCCTACACGACGCTCTTCCGATCT) and P7 (GGAGTTCAGACGTGTGCTCTTCCGATCT) Illumina adapters were included to the 5' part of the *rbcL* forward and reverse primers respectively. PCR1 amplifications were performed in a final volume of 25µL following mix and reaction conditions used in Vasselon *et al.* (2017), except the number of amplification cycles which was set to 33. (ii) PCR2: the 3 PCR1 replicates prepared for each DNA sample were pooled and sent to the "GenoToul Genomics and Transcriptomics" facility (GeT-PlaGe, Auzeville, France) where subsequent laboratory preparations were performed. PCR1 amplicons were purified and used as templates in the PCR2 which used Illumina-tailed primers targeting the half of P5 and P7 sequences. Finally, all generated 278 PCR2 amplicons were dual indexed and pooled into a single tube. Final pool was sequenced on an Illumina Miseq platform using the V3 paired-end sequencing kit (250 bp × 2).

HTS data analyses

Demultiplexed and overlapped Miseq data were delivered by the GeT-PlaGe sequencing platform (paired sequences overlap > 140 bp and mismatches < 0.1 %), resulting in 278 fastq files. A quality filtering was performed using Mothur software (Schloss et al., 2009) to remove DNA reads with: Phred quality score < 23 over a moving window = 25 bp, primer sequence mismatch > 1, homopolymer > 8 bp, ambiguous base > 0. Chimeras were removed using the Uchime algorithm (Edgar, Haas, Clemente, Quince, & Knight, 2011) available in Mothur. Then, all the fastq files were combined and de-replicated in order to keep only unique sequences with DNA read abundance > 2. Using the R-Syst::diatom library (Rimet et al., 2016) and the naïve Bayesian method (Wang, Garrity, Tiedje, & Cole, 2007), taxonomy was assigned to each DNA read with a confidence threshold > 85 %. DNA reads assigned to the Bacillariophyta phylum were clustered into OTUs using a distance similarity threshold of 95 % as described in Vasselon, Domaizon, et al. (2017). For each sample, the 2 replicates were merged and only the OTUs shared by both replicates were conserved in order to remove unrepresentative and spurious OTUs. The taxonomy of OTUs were defined as the consensus taxonomy of DNA reads (threshold > 80). A DNA representative sequence was determined for each OTU using the *Get.oturep* command in Mothur.

Biotic indices

We computed four biotic indices for 139 sites, all based on the IPS index (Coste, 1982). The first index IPS-MICROTAXO was computed from the relative abundances of the species estimated using traditional microscopy. The second, IPS-DNATAXO was computed from the relative abundance of the OTUs after they were classified into species using Mothur. Since the IPSS and IPSV values are inherited from the taxonomical affiliation, the fraction of unclassified OTUs cannot be used for this index. The third index, IPS-DNAPHYLO takes into account all the OTUs. For this index the IPSS and IPSV values are phylogenetically imputed. OTUs were placed within the reference phylogenetic tree using their representative sequence (most abundant sequence) and the EPA algorithm. The IPSS and IPSV values of each OTU were estimated using *rphylopars* with the best evolution model selected at

the cross-validation step (see above). Finally, the fourth index, IPS-DNAHYBRID is a combination of IPS-DNATAXO and IPS-DNAPHYLO: species IPSS and IPSV are used for OTUs which can be classified into species using Mothur, while the unclassified fraction of OTUs is used with phylogenetically imputed IPSS and IPSV values. DNA-based indices were compared to IPS-MICROTAXO using the mean squared error (MSE) and pairwise Wilcoxon signed-rank test on squared error.

Results

Quality of read placements

Overall, barcode sequences allowed to place species accurately within the reference phylogeny (Fig. 1, Table S1, Supporting information). About 45% (272) of the species were placed exactly at the same location as the full-length sequence. Most of the species (508; 84%) were placed at a short distance, ≤ 3 nodes from the reference target. Only a few species were not placed correctly within the reference phylogeny (35 species; 5.8% at ≥ 10 nodes from the reference targets).

Quality of autecological values estimation

We found a significant phylogenetic signal for IPSS ($\lambda = 0.67$; p-value < 0.001) and for IPSV ($\lambda = 0.52$; p-value < 0.001). For IPSS, five phylogenetic models (*lambda*, *delta*, *kappa*, *BM* and *EB*) produced better predictions (lower MSE, p-values < 0.001) than the non-phylogenetic *star* model (Fig. 2; Table S2, Supporting information). The best model with the lowest MSE was the *lambda* model which exhibited a 30% decrease of MSE compared to the *star* model. For IPSV, five phylogenetic models produced lower MSE than the *star* model but differences were not significant (p-values > 0.05).

The estimated IPSS values for each species are mapped onto the reference phylogenetic tree in Fig. 3 and can be compared with the true IPSS values. For 150 (63%) of the species represented in green in Fig. 3, the absolute error was found to be low (≤ 1), indicating a good prediction. The absolute error

was ranging from 1 to 2 for 79 (33%) of the species (represented in orange), indicating a poor prediction quality. Finally, for a few species (8; 3%) the prediction quality was found to be very poor (absolute error > 2).

Morphological analysis

A total of 534 species were determined using microscopy. The dominant species were *Achnanthidium minutissimum* (14% of the valves determined), *Achnanthidium pyrenaicum* (6%), *Amphora pediculus* (5%), *Achnanthidium delmontii* (5%) and *Eolimna minima* (4%). For these 5 dominant taxa a reference barcode is present in the R-Syst::diatom library, except for *Achnanthidium delmontii.* Among the 100 most frequently determined species, 38 have a barcode in R-Syst::diatom and among the 534 species, only 114 species have a DNA barcode.

HTS analysis

The Illumina Miseq sequencing produced a total of 11,249,428 x 2 DNA reads. After all the bioinformatics processes, the OTU list obtained for the 139 samples included 682 OTUs composed by 3,033,967 DNA reads. After the taxonomic assignment, 362 OTUs were identified at the genus level (77 % of DNA reads) and 205 at the species level (58 % of DNA reads). Final molecular taxonomic list contained 28 families, 53 genera and 102 diatom species. The final list of OTUs with their taxonomic assignment and DNA representative sequence is available in the Supporting information (Table S3).

Performances of biotic indices

The distribution of IPS-MICROTAXO scores was left skewed with a majority of high rated sites. DNA-based indices scores exhibited unimodal distribution with a restricted variability (few sites low rated and high rated). This was particularly true for IPS-DNAPHYLO which showed a variance of 0.12, much lower than the variance of IPS-MICROTAXO ($s^2 = 0.5$). The indices were significantly correlated with each other (Fig. 4; all correlations > 0.49 and all p-values < 0.001). When comparing DNA-based indices with IPS-MICROTAXO, IPS-DNAHYBRID appeared to be the best index with

the highest correlation (r = 0.74) and the lowest MSE (0.33). IPS-DNATAXO and IPS-DNAPHYLO exhibited similar correlation with IPS-MICROTAXO (r = 0.69 and r = 0.70, respectively) and similar MSE (0.45 and 0.43). Wilcoxon tests detected no difference between the squared error distribution of IPS-DNATAXO and IPS-DNAPHYLO (p-value = 1) but both methods were significantly outperformed by IPS-DNAHYBRID (p-values < 0.05).

Discussion

DNA metabarcoding appears to be a promising alternative to the traditional methods of characterizing biodiversity and assessing environmental quality. However, the massive quantities of genetic data produced by HTS challenge ecologists to think differently about the way biotic indices are computed (Keck et al., 2017). In this paper, we have introduced a new method based on phylogeny to compute biotic indices from DNA reads generated by metabarcoding workflows. The phylogenetic method is in line with the recent developments in taxonomy-free approaches for bioassessment which aim to bypass taxonomic reference libraries in order to maximize the genetic information taken into account (e.g., Apothéloz-Perret-Gentil et al., 2017; Cordier et al., 2017). The phylogenetic OTU-PITI approach has sound theoretical grounds, because, the imputation of missing values is based on the phylogenetic signal (i.e., the non-independence among species trait values because of their phylogenetic relatedness) which is a direct consequence of Darwin's principle of descent with modification (Felsenstein, 1985).

The OTU-PITI approach is based on two main steps: first, the placement of DNA reads within the phylogeny and second, the estimation of their ecological values. We found that short *rbcL* marker (312 bp) gives satisfactory results with most of the species barcodes placed exactly or very close to their reference position. This is consistent with the benchmark results obtained by Berger *et al.* (2011) on short sequences (200 ± 60 bp) in the original publication of the Evolutionary Placement Algorithm. However, some species could not be placed correctly by the EPA (Fig. 1; see Table S1, Supporting information for the detailed list). The performances of the EPA for a given sequence may depend on

many factors like the choice of genetic marker, the length of the sequence, and the presence of closely-related taxa in the reference tree. In our case, it seems that wrong placements often involve species isolated in the phylogeny. Thus, increasing the phylogenetic coverage of underrepresented taxa may help to improve the placement of these species. Obviously, longer DNA reads capture more historical signal. Hence, the quality of reads insertion is also expected to improve as read lengths produced by HTS will increase (Tedersoo, Tooming-Klunderud, & Anslan, 2017). Finally, it should be noted that most of the species which were wrongly placed are marine (e.g., *Guinardia striata*, *Stephanopyxis turris*) and therefore will not impede the computation of freshwater biotic indices like the IPS Index.

Diatoms pollution sensitivity (IPSS) exhibited a significant phylogenetic signal and was much better predicted using a phylogenetic model (*lambda*) than using the ad-hoc non-phylogenetic *star* model. These results are consistent with the presence of phylogenetic signal for ecological optima and pollution sensitivity (Keck et al., 2016). Nonetheless, some species were very poorly predicted (e.g., *Nitzschia soratensis*, *Terpsinoë musica*). Incorrect estimation can be the result of a wrong placement of the species within the phylogeny. For example, the high absolute error found for *Lemnicola hungarica* (2.55) could be explained by a rough phylogenetic placement of this species (node distance = 11). It is also clear that trait imputation is less effective when closely related species exhibit very contrasted trait values and therefore strongly depart from the underlying model of evolution (overdispersion). For example, *Halamphora oligotraphenta* and *Halamphora veneta* are two closely related species with very different ecological preferences, the former living in oligotrophic freshwater, while the latter is found in eutrophic habitats (Levkov, 2009). As a result, the pollution sensitivity values of these two species are incorrectly predicted (Fig. 3). Overdispersion can be the consequence of recent evolutionary events and selection under active constraints like convergent evolution or character displacement. Unlike IPSS, the species tolerance value IPSV was poorly predicted by the phylogenetic models. The fact that the ad-hoc model (*star*) performed as good as the best phylogenetic model (*lambda*) reflects a low, yet significant phylogenetic signal, or a signal which cannot be appropriately modelled with the tested phylogenetic methods. A weak signal can be the

result of trait instability and lability over time (Blomberg et al., 2003). With IPSV being an approximate and partial measure of diatoms realized niche volume, its variability may be more related to interspecific interactions and non-genetic effects.

The two IPS-derived indices implementing the OTU-PITI approach (IPS-DNAPHYLO and IPS-DNAHYBRID) were strongly correlated to the index estimated from microscopy. However, IPS-DNAPHYLO had a restricted range of values, with a tendency to overestimate the score of bad quality sites and underestimate the score of good quality sites. This tendency is likely to be caused by the phylogenetic imputation algorithm which has been shown to be a form of kriging (Cressie, 1993) in a phylogenetic context (Ho et al., 2014). As an inverse distance weighting method, kriging is subject to a smoothing effect and does not reproduce the histogram of the sample data (Isaaks & Srivastava, 1989). One solution could be to estimate the strength of smoothing from the LOOCV data and apply a correction factor to the OTUs estimated ecological values. With the true ecological values being more reliable than the phylogenetically imputed ones, we advocate for the use of the DNAHYBRID which benefits from the ecological values of the assigned species if available while using 100% of the OTUs via the OTU-PITI approach. In our study, the IPS-DNAHYBRID is the molecular index correlating the best with the index based on microscopy. Despite a deviation of the residuals around the 1:1 line, this method has the lowest error rate (MSE = 0.32).

The OTU-PITI approach solves two important problems that scientists and environmental managers recurrently face when using metabarcoding data to compute biotic indices. The first problem is the incompleteness of reference libraries connecting DNA barcode sequences to taxonomic names. An incomplete library strongly limits the proportion of OTUs which can be taxonomically assigned and used for indices calculation. In this study, the reference library covered 21% of the species detected using microscopy. As a consequence, the proportion of OTUs assigned at species level was only 30%, similar to the proportions obtained in previous diatom studies (e.g., Apothéloz-Perret-Gentil *et al.* 2017; Rivera *et al.* 2017; Vasselon *et al.* 2017b respectively 35%, 23% and 35.7%). Additionally, the OTU-PITI approach offers a convenient solution to the incompleteness of ecological libraries connecting taxa and ecological values. Some species, detected either by microscopy or DNA, do not

have IPSS and IPSV values. Therefore, they cannot be used to compute the IPS index. For example, in this study, 9 OTUs were assigned to species which were not found in the IPS library. This is often the result of taxonomic names discrepancies among libraries (synonyms, misspellings), but in some cases autecological information can simply be missing. The OTU-PITI allowed to estimate IPSS and IPSV values for these OTUs and include them in the calculation of IPS. This feature is particularly interesting for the implementation of biotic indices including a restricted number of taxa, or to extend the use of well-established indices to new habitats and new regions with endemic taxa. However, practitioners must keep in mind that the phylogenetic estimation of ecological profile is error prone and associated with uncertainty. For example, the case of closely related species with different ecological profiles can be problematic as an unexpected cryptic diversity has been recently described within many diatom species complexes (Mann & Evans, 2007) and we are still in the beginning of understanding their autecology (Vanelslander et al., 2009). Nonetheless, with a large number of samples and OTUs, our results suggest that the phylogenetic signal is strong enough to improve water quality assessment on average.

Two other taxonomy-free approaches to compute molecular indices have been recently introduced in the literature. Firstly, Apothéloz-Perret-Gentil *et al.* (2017) proposed to assign ecological values directly to OTUs. Secondly, Cordier *et al.* (2017) investigated the use of supervised machine-learning regression to infer indices values from lists of OTUs. The main advantage of the OTU-PITI over these two approaches is that it does not require the collection of chemical and physical measurements to train or calibrate the model which makes it a ready-to-use tool, not restricted to the geographical area of the training data. Conversely, a well-trained machine learning classifier used within its geographical scope will probably outperform the OTU-PITI approach. As advocated by Keck *et al.* (2017), OTU-PITI, OTU-based indices and machine learning are complementary tools which should make it possible to make better use of genetic data in the future.

Our knowledge of biodiversity is very uneven. Microscopic organisms, which include diatoms, are extremely diversified and largely unknown. Thus, OTU-PITI may be a very useful way to address this paucity of biodiversity knowledge, pending the availability of comprehensive taxonomic and

ecological libraries. Here we have shown that this approach can be successfully applied to use the unclassified DNA material which is normally discarded from biotic indices computation. The range of applications of the OTU-PITI is large: the method can be applied to any biotic index and any group of biological indicator, provided that an accurate phylogeny is available. Moreover, traits values can be modeled and estimated within phylogenetic multivariate frameworks (Clavel, Escarguel, & Merceron, 2015; Goolsby et al., 2017). Multiple biological traits and functional groups come with several advantages compared to biotic autecological indices (Bonada, Prat, Resh, & Statzner, 2006; Tapolczai, Bouchez, Stenger-Kovács, Padisák, & Rimet, 2016). Thus, the OTU-PITI approach could be a way to integrate the immense diversity revealed by metabarcoding and move closer towards functional biomonitoring of the environment.

Acknowledgment

We thank Diego Fontaneto and Francis Burdon for providing perceptive comments on this manuscript. This article is based upon work from COST Action DNAqua-Net (CA15219), supported by the COST (European Cooperation in Science and Technology) program. We thank Cécile Chardon who performed the DNA extractions, amplifications, and the preparation of libraries. Samplings and microscopic analyses were carried out by the Dreal (Direction Régionale de l'Environnement de l'aménagement et du Logement) Aquitaine (D. Sagnet), Auvergne (F. Véry), Bourgogne (V. Peeters), Bretagne (G. Gicquiaud), Centre (S. Saadat, C. Karabaghli), Franche-Comté (E. Parmentier), Limousin (J.-M. Vouters), Lorraine (D. Heudre), Midi Pyrénées (E. Seigneur), Pays de la Loire (D. Guillard), Rhône-Alpes (R. Chavaux), Normandie (F. Petel), Nord Pas de Calais (N. Zydek), and the private officies Aquabio (R. Marcel, B. Fontan), Asconit (L. Kermarrec, E. Ponton), Sage (A. Rolland, J.-P. Vulliet, C. Geret). We thank the French Water Agencies, Artois-Picardie (C. Lesniak), Rhône-Méditerranée et Corse (L. Imbert, F. Repellini), Adour-Garonne (M. Durand, J.-P. Rebillard, M. Saut), Rhin-Meuse (J.-L. Matte, G. Demortier), Loire-Bretagne (J. Durocher), Seine-Normandie (M. Berdoulay) who funded the microscopical analyses. The AFB (Agence Française de la Biodiversité) funded the sequencing, which was realized in the GeT-PlaGe sequencing plateform. We

thank the Swedish Agency for Marine and Water Management for a contribution via the program Environmental monitoring (project 2014-16, Development of the diatom barcoding method for freshwater).

References

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

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

Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology*, *60*(3), 291–302. https://doi.org/10.1093/sysbio/syr010

Besse-Lototskaya, A., Verdonschot, P. F., 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

Besse-Lototskaya, A., Verdonschot, P. F. M., & Sinkeldam, J. A. (2006). Uncertainty in diatom assessment: sampling, identification and counting variation. *Hydrobiologia*, *566*(1), 247–260. https://doi.org/10.1007/s10750-006-0092-5

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, *57*(4), 717–745.

Bonada, N., Prat, N., Resh, V. H., & Statzner, B. (2006). Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology*, *51*, 495–523. https://doi.org/10.1146/annurev.ento.51.110104.151124

Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S., & Bremer, K. (2007). Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*, *56*(5), 741–752. https://doi.org/10.1080/10635150701613783

Bruggeman, J., Heringa, J., & Brandt, B. W. (2009). PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, *37*(Web Server issue), W179–W184. https://doi.org/10.1093/nar/gkp370

Chapman, D. V. (1996). *Water quality assessments: a guide to the use of biota, sediments and water in environmental monitoring*. E & Fn Spon London. Retrieved from http://wwwlive.who.int/entity/water_sanitation_health/resourcesquality/watqualassess.pdf

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*, 965–975. https://doi.org/10.1016/j.scitotenv.2015.10.161

Clavel, J., Escarguel, G., & Merceron, G. (2015). mvmorph: an r package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, *6*(11), 1311–1319. https://doi.org/10.1111/2041-210X.12420

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., … 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

Coste, M. (1982). *Étude des méthodes biologiques d'appréciation quantitative de la qualité des eaux* (p. 218). Cemagref.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: Wiley.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

European Committee for Standardization. (2014). *Water quality - Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters.* (No. European Standard NF EN 14407) (pp. 1–13). Brussels.

European Committee for Standardization. (2016). *Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers* (No. European Standard NF EN 13946) (pp. 1–18). Brussels.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, *125*(1), 1–15.

Goolsby, E. W., Bruggeman, J., & Ané, C. (2017). Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, *8*(1), 22–27. https://doi.org/10.1111/2041-210X.12612

Ho, T., Si, L., & Ané, C. (2014). A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic Biology*, *63*(3), 397–408. https://doi.org/10.1093/sysbio/syu005

Ibáñez, C., Caiola, N., Sharpe, P., & Trobajo, R. (2010). Ecological indicators to assess the health of river ecosystems. In S. E. Jørgensen, F.-L. Xu, & R. Costanza (Eds.), *Handbook of Ecological Indicators for Assessment of Ecosystem Health* (2nd ed., pp. 447–464). Boca Raton, Florida: CRC Press.

Isaaks, E. H., & Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. New York: Oxford University Press.

Keck, F., Rimet, F., Bouchez, A., & Franc, A. (2016). phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecology and Evolution*, *6*(9), 2774–2780. https://doi.org/10.1002/ece3.2051

Keck, F., Rimet, F., Franc, A., & Bouchez, A. (2016). Phylogenetic signal in diatom ecology: Perspectives for aquatic ecosystems biomonitoring. *Ecological Applications*, *26*(3), 861–872. https://doi.org/10.1890/14-1966

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

Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Computational Biology*, *8*(10), e1002743. https://doi.org/10.1371/journal.pcbi.1002743

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

Levkov, Z. (2009). *Amphora sensu lato*. (H. Lange-Bertalot, Ed.). Gantner Verlag.

Mann, D. G., & Evans, K. M. (2007). Molecular genetics and the neglected art of diatomics. In J. Brodie & J. Lewis (Eds.), *Unravelling the algae: the past, present, and future of algal systematics* (pp. 231–265). Boca Raton, Florida: CRC Press. Retrieved from http://books.google.fr/books?hl=fr&lr=&id=-YEYFhgUBsQC&oi=fnd&pg=PA231&dq=molecular+genetics+and+the+neglected+art+of+diatomics&ots=JuYPRbKgNm&sig=sq3Ll7QwK4YGoUCiTNK_Ud1A3Gw

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

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884.

Prygiel, J., Carpentier, P., Almeida, S., Coste, M., Druart, J.-C., Ector, L., … Ledeganck, P. (2002). Determination of the biological diatom index (IBD NF T 90–354): results of an intercomparison exercise. *Journal of Applied Phycology*, *14*(1), 27–39. https://doi.org/10.1023/A:1015277207328

Rimet, F. (2012). Recent views on river pollution and diatoms. *Hydrobiologia*, *683*(1), 1–24. https://doi.org/10.1007/s10750-011-0949-0

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*, *2016*, baw016. https://doi.org/10.1093/database/baw016

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., & Rimet, F. (2017). Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia*. https://doi.org/10.1007/s10750-017-3381-2

Round, F. E., Crawford, R. M., & Mann, D. G. (1990). *The diatoms: biology and morphology of the genera*. Cambridge, UK: Cambridge University Press.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, btu033. https://doi.org/10.1093/bioinformatics/btu033

Stevenson, R. J., Yangdong, P., & Van Dam, H. (2010). Assessing environmental conditions in rivers and streams with diatoms. In J. P. Smol & E. F. Stoermer (Eds.), *The Diatoms: Applications for the Environmental and Earth Sciences* (2nd ed., pp. 55–85). Cambridge University Press. Retrieved from https://doi.org/10.1017/CBO9780511763175.005

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., Bouchez, A., Stenger-Kovács, C., Padisák, J., & Rimet, F. (2016). Trait-based ecological classifications for benthic algae: review and perspectives. *Hydrobiologia*, *776*(1), 1–17. https://doi.org/10.1007/s10750-016-2736-4

Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2017). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist*. https://doi.org/10.1111/nph.14776

Theriot, E. C., Ashworth, M. P., Nakov, T., Ruck, E., & Jansen, R. K. (2015). Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Molecular Phylogenetics and Evolution*, *89*, 28–36. https://doi.org/10.1016/j.ympev.2015.03.012

Theriot, E. C., Ruck, E., Ashworth, M., Nakov, T., & Jansen, R. K. (2011). Status of the pursuit of the diatom phylogeny: Are traditional views and new molecular paradigms really that different? In J. Seckbach & J. P. Kociolek (Eds.), *The Diatom World* (pp. 119–142). New York, USA: Springer. Retrieved from https://doi.org/10.1007/978-94-007-1327-7_5

Vanelslander, B., Créach, V., Vanormelingen, P., Ernst, A., Chepurnov, V. A., Sahan, E., … Sabbe, K. (2009). Ecological Differentiation Between Sympatric Pseudocryptic Species in the Estuarine Benthic Diatom Navicula Phyllepta (bacillariophyceae). *Journal of Phycology*, *45*(6), 1278–1289. https://doi.org/10.1111/j.1529-8817.2009.00762.x

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

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

Zelinka, M., & Marvan, P. (1961). Zur präzisierung der biologischen klassifikation der reinheit fließender gewässer. *Archiv Für 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

Data accessibility

Illumina Miseq raw data and diatom microscopic counts can be accessed at doi: 10.5281/zenodo.1005186. *rbcL* sequences alignment and phylogenetic tree can be accessed at doi: 10.5281/zenodo.998556. The code to reproduce the analyses is available at doi: 10.5281/zenodo.1228696.

Author Contributions

FK and VV conceived the study and performed the data analyses. FK wrote the paper with significant contributions from all authors. All authors gave final approval for publication.

Figures

Fig. 1 Histogram showing the placement accuracy of the 604 species from the reference tree using 312 bp *rbcL* barcode sequences and the EPA algorithm.

Fig. 2 Barplots showing the LOOCV mean squared error of each model for IPSS and IPSV estimation.

Fig. 3 Phylogenetic tree representing 236 diatoms species for which both phylogenetic position and IPSS value were available. For each species, true IPSS value is represented as a point, while its estimated IPSS value is represented as a dash. Low absolute errors (≤ 1) are represented in green, medium absolute errors (> 1 and ≤ 2) in orange and high absolute errors (> 2) in red.

Fig. 4 Distributions and relationships between the 4 indices computed for 139 environmental samples. Diagonal: Histograms of the distribution of each index expressed as frequencies. Lower triangle: Scatterplots showing the relationships between the indices. The dashed lines represent the full equivalence between the indices. Upper triangle: correlation (cor) and mean squared error (MSE) between the indices.

Supporting Information

Fig. S1 Phylogenetic tree with bootstrap support values. The topological constraint used for the reconstruction is highlighted in red.

Fig. S2 Map of sampling sites.

Table S1 Node distances between references and barcode placements, detailed per species.

Table S2 Leave one out cross-validation results for IPSS and IPSV phylogenetic imputation (best model), detailed per species.

Table S3 List of OTUs, number of copies, taxonomic affiliations and DNA representative sequences.