Diatom metabarcoding applied to large scale monitoring networks: Optimization of bioinformatics strategies using Mothur software

Sinziana Rivera, Valentin Vasselon, Agnes Bouchez, Frédéric Rimet

Diatom metabarcoding applied to large scale monitoring networks: Optimization of bioinformatics strategies using Mothur software

Sinziana F. RIVERA[1], Valentin VASSELON[2], Agnès BOUCHEZ[1], & Frédéric RIMET[1]

1: Université Savoie Mont-Blanc, INRA, UMR CARRTEL, 75bis av. de Corzent - CS 50511, FR-74200 Thonon-les-Bains, France

2: AFB, Pôle R&D « ECLA », INRA, UMR CARRTEL, 75bis av. de Corzent - CS 50511, FR-74200 Thonon-les-Bains, France

Corresponding author: sinzianaflorina@gmail.com

Highlights (3-5 bullets points, maximum 85 characters, including spaces, per bullet point)

- DNA-metabarcoding for diatom biomonitoring was tested on 447 French river sites
- We identified bioinformatics strategies giving the closest result to microscopy
- Non clustering strategy (ISU) requires less computing power and gave good results
- A loose taxonomic assignment, rather than a stringent one, was recommended
- The indices of microscopy and ISU deviate by less than 1 point for 72% of the 447 sites

Abbreviations

HTS: high-throughput sequencing

ISU: individual sequence units

OTU: Operational taxonomical Unit

Abstract (400 words)

Benthic diatoms are routinely used as ecological indicators in rivers. A standardized methodology is based on biofilm sampling, species identification, and counting under microscope. DNA-metabarcoding is an alternative methodology that can identify species and assess their proportion based on high-throughput DNA sequencing. Sequence data is analyzed with bioinformatics tools, and several strategies can be chosen. The strategy choice can affect communities composition and structure, and therefore the resulting ecological assessment. We wanted to optimize the bioinformatics strategy to obtain the closest results to microscopy. This was done in the framework of the Mothur pipeline. Here, 447 samples from French rivers were analyzed in the monitoring context of the European Water Framework Directive. Samples were analyzed both with DNA metabarcoding and microscopy. A usual bioinformatics strategy in Mothur includes clustering DNA-sequences into Operational Taxonomic Units (OTUs). Different algorithms exist for this. From a subsample of 142 samples, we showed that some strategies (Furthest neighbor) gave closer results to microscopy than others (Opticlust) in terms of community structure and diatom index values. However, we showed that OTU clustering was not necessary for

ecological monitoring: Direct taxonomic assignment of individual sequence units (ISU) gave similar results to those obtained in microscopy. Interestingly, direct assignment enabled the detection of more species 2 to 3 times faster in terms of computation time compared to the OTU strategy. However, it remained important to remove low quality and chimeric sequences; if not, biomonitoring results differed greatly from microscopy. We showed that it was preferable to have a loose taxonomical identification threshold instead of a stringent one. This allowed detecting more species, which could participate in the index calculation and increased its performance. Indeed, in diatoms, phylogenetically neighbor species often have similar ecologies, and this explains why it is preferable, in a biomonitoring framework, to identify more species with less stringency instead of identifying few species with stringency. Finally, the best strategy (direct assignment of filtered ISU with a loose taxonomical threshold of 60%) was applied to the 447 samples covering a large diversity of ecological qualities. These data were then used to produce quality index values, using a quantification correction factor taking into account species biovolumes. Compared to microscopy, the DNA-based method assigned the same quality class for 66% of the samples, and 72% of the samples had an index value (ranging from 0 to 20) with less than one point difference from microscopy.

Keywords

Diatoms, DNA metabarcoding, biomonitoring, bioinformatics

1. Introduction

Diatoms are ubiquitous unicellular microalgae routinely used as biological indicators of water quality in Europe as part of the Water Framework Directive (European commission, 2000). Current standard methods for water quality assessment using diatoms are based on the characterization of environmental assemblages that are subsequently used to calculate biotic indices (e.g. Rimet, 2012). Indices calculation requires identification of diatom taxa to the species level based on the morphology of their frustule. This microscopic identification is challenging because it requires a strong taxonomic expertise and is time-consuming thus limiting the use of diatoms for routine monitoring.

Diatom DNA metabarcoding was developed in recent years as an alternative method for diatom identification (e.g. Kermarrec et al., 2013; Zimmermann et al., 2015). By combining a DNA barcode reference database and bioinformatics processing steps, it is possible to obtain taxonomic lists that can be used for ecological assessment from high-throughput sequencing (HTS) data. To prepare these molecular inventories, traditional bioinformatics procedures, such as in implemented in Mothur (Schloss et al., 2009), usually involve several steps: a) sample demultiplexing, b) assembly of paired-end reads, c) removing poor quality sequences, d) sequence dereplication into ISU, e) detecting and removing chimeras, f) clustering of remaining reads into operational taxonomic units (OTUs) based on their genetic similarity using a similarity threshold, and g) taxonomic assignment of each OTU using a taxonomic assignment threshold.

Different programs have been developed to process HTS data like Mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010), UPARSE (Edgar, 2013), DADA2 (Callahan et al., 2016), etc. For each pipeline, the user has to choose a number of settings (e.g. length and filtering criteria, removal of chimeric sequences, choice of a clustering algorithm, selection of a taxonomic assignment threshold for OTUs, etc.). The choice of these settings is not straightforward and may affect the final molecular inventories, which may in turn change the ecological assessment. For instance, prior work showed that the choice of bioinformatics treatment strongly affects final molecular inventories in eukaryotes and the ecological conclusions for marine nematodes and for diatoms (Majaneva et al., 2015; Holovachov et al., 2017; Tapolczai et al. 2019, respectively).

Furthermore, the similarity matrix generation step towards the creation of OTUs requires considerable storage space and running time (Al-Neama et al., 2014). Bioinformatics can create molecular diatom inventories for biomonitoring purposes, but this process has not yet been standardized so the impact of the settings choice on the final ecological assessment is still unknown despite the work of many scientists (Leese et al., 2016). Fortunately, several studies have shown the potential of diatom metabarcoding for water quality assessment at small (Kermarrec et al., 2014; Zimmermann et al., 2015; Visco et al., 2015; Vasselon et al. 2017a) and regional scales (Apothéloz-Perret-Gentil et al., 2017; Rivera et al., 2018a; Vasselon et al., 2017b).

Until now, the bioinformatics treatment of the various diatom metabarcoding studies carried out in French rivers and lakes was performed using Mothur software (Schloss et al., 2009). Here, precise settings were used, and the classical bioinformatics treatment used has been described (Keck et al., 2018). The aim of this study was to optimize some of these settings in order to obtain metabarcoding assessment results as close as possible to those obtained through microscopy which is the reference methodology for water managers at present. Furthermore, since computational power is still a drawback —especially for large data sets coming from monitoring studies— we wanted to know if we could simplify bioinformatics treatment to be faster. In this sense, we ran tests by changing the following settings: a) confidence threshold for taxonomic assignment of DNA sequences called Individual Sequence Units (ISU, after Esling et al., 2015), b) clustering OTU methods, and c) confidence threshold for the taxonomic assignment of OTU data.

We made these tests on 142 diatoms samples collected in 2016 from rivers from the French WFD monitoring network and compared the results to the morphological data. This comparison helped to select the nearest results to the microscopic analyses. We then applied this strategy on a larger set of 447 diatom samples (305 samples from 2017 combined to the 142 samples from 2016) and combined those with the ones from 2016. Finally, we attempt to refine molecular inventories by considering the biovolumes of species. This made the HTS data more similar to microscope analyses after this transformation (Vasselon et al., 2018). These different strategies were compared to microscopy in terms of taxonomic composition, community structure, and biotic indices.

2. Methodology

2.1. Study sites and sampling

In order to test diatom metabarcoding on a large geographical scale, 447 diatom samples were collected from the French river monitoring network that is composed of seven main basins (Adour-Garonne, Artois-Picardie, Loire-Bretagne, Rhin-Meuse, Rhône-Méditerranée, Corse, and Seine-Normandie). Samples were collected only from mainland during two sampling campaigns held in 2016 and 2017 resulting in 142 and 305 samples, respectively (Figure 1). Sampling sites are part of the national river monitoring network funded by the Water Agencies and are monitored every year through microscope diatom analyses. Only a part of the monitoring network was analyzed. The site selection has been validated by experts at regional agencies (*Direction Régional de l'Environnement, de l'Aménagement et du Logement*). Sampling sites were located mainly in rivers presenting marked pollution gradients and in rivers weakly impacted by anthropogenic pressures (references sites). Finally, the entire river network of the eastern administrative divisions (départments) Ain, Jura, Haute-Savoie, Savoie, Rhône and Loire was sampled because it covers a large range of habitats (alpine, lowlands, agriculture, forest and densely urbanized), human densities, and pollution levels.

Diatom sampling was performed following the French standard NFT 90 354 (Afnor, 2007) and the European standard (Afnor, 2014a). Briefly, diatoms were collected from at least five stones from the

fast-flowing parts of the rivers. The upper surface of the stones was scrapped using a toothbrush to collect the biofilms containing diatoms. Biofilms were then fixed with ethanol (90%) to give a final concentration of at least 70%. Samples were stored in the dark at 7°C until molecular and microscope analyses.

2.2. Morphological analysis

Diatom valves were cleaned from environmental samples using 40% H_2O_2 and HCl. Clean valves were mounted in resin (Naphrax©). At least 400 valves from each sample were counted and identified using light microscopes (1000× magnification) according to European (Afnor, 2010) and French (Afnor, 2007) standards. The abundances of all observed taxa were expressed as relative counts. Identification to species level was done based on European floras such as Krammer and Lange-Bertalot (1986), Krammer and Lange-Bertalot (1988), Krammer and Lange-Bertalot (1991a), Krammer and Lange-Bertalot (1991b), Reichardt (1997), Lange-Bertalot et al.(2017) and according to the European standard Afnor (2014b). A list of the taxa and their relative abundances was produced for each of the samples. Morphological analyses were performed by private agencies following inter-calibration standards for diatom counting.

2.3. Molecular analysis

DNA extraction was performed twice. Samples from the first sampling campaign (2016) were extracted using the GenElute TM-LPA protocol described in Chonova *et al*. (2016). Several samples from this campaign could not be amplified because they were loaded with humic acids known to be PCR inhibitors. As a result, non-amplified samples and samples from the second sampling campaign (2017) were extracted using the commercial DNA extraction kit Macheray-Nagel NucleoSpin® Soil kit (MN-Soil) including a column purification step to remove PCR inhibitors. For each sample, 2 ml of biofilm was centrifuged at 13,000 rpm for 30 min at 4°C. After centrifugation, the supernatant containing ethanol was removed, and the pellet was used as a starter for DNA extraction. Extractions were performed following the manufacturer's instructions. Some authors (Deiner et al., 2015) have shown the impact of extraction protocols on biodiversity assessment in rivers, but others (Vasselon et al., 2017a) showed that the choice of the extraction method has no impact on the diatom indices calculated for quality assessment even if relative abundances of some taxa can be slightly affected by the methods. For sequencing all samples in a single Illumina Miseq run, HTS libraries were prepared using two successive PCR steps as described in Keck et al. (2018):

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 et al., 2017b) allowing one to focus on a short fragment of the r*bcL* plastid gene (312 bp). Half of the P5 (CTTTCCCTACACGACGCTCTTCCGATCT) and P7 (GGAGTTCAGACGTGTGCTCTTCCGATCT) Illumina adapters were included to the 5′ part of the r*bcL* forward and reverse primers, respectively. PCR1 amplifications were performed in a final volume of 25 µl following mix and reaction conditions used previously (Vasselon et al., 2017a, b) except for the number of amplification cycles which was set to 33.

PCR2: The three PCR1 replicates prepared for each DNA sample were pooled and sent to the "GenoToul Genomics and Transcriptomics" platform (GeT - PlaGe, Auzeville, France) where subsequent laboratory preparations were performed. PCR1 amplicons were purified and used as templates in the PCR2 that used Illumina tailed primers targeting the half of P5 and P7 sequences. Finally, all generated PCR2 amplicons were dual-indexed and pooled into a single tube. The final pool was sequenced on an Illumina MiSeq platform using the V3 paired-end sequencing kit (250 bp × 2). Raw sequencing data is available on https://data.inra.fr/dataset.xhtml?persistentId=doi%3A10.15454%2F9EG5Z4

2.4. Sequencing data processing

Sequencing data processing was conducted in two stages. We first tested 16 bioinformatics strategies to produce diatom floristic lists for the 142 samples collected in 2016. We used a Dell Precision, Tower 7910 workstation (16 processors, 2.60 GHz, 64 Go RAM). Second, the bioinformatic strategy showing the nearest results to microscopy was adopted to produce diatom floristic lists for the 447 samples sequenced during this study (campaigns in 2016 and 2017). Bioinformatics treatment was performed in Mothur software (Schloss et al., 2009) based on the bioinformatics treatment presented previously (Keck et al., 2018) and summarized in Figure 2.

Classical sequence data processing

The Genotoul sequencing platform (GeT-PlaGe, Auzeville, France) provides for Miseq sequencing demultiplexed and overlapped *fastq* files. They are the starting point of our bioinformatics treatment. For each *fastq* file, DNA reads are filtered by length and quality according to the following criteria: minimum length = 250 bp, Phred quality score >23 over a moving window of 25 bp, maximum 1 mismatch in forward primer sequence, homopolymers <8 bp. In addition, any sequences containing ambiguous base calls are removed (maxambig=0). Then, all the resulting *fasta* files are combined and de-replicated to keep only unique sequences (ISU) with read abundances >2. This step enables to remove low abundant reads mainly related to sequencing and PCR errors, with the added benefit of saving processing time during the next steps of the bioinformatics treatment.

Next, the *Vsearch* algorithm detects and removes chimeric DNA sequences. Then, taxonomic assignment of ISU is performed using the naive Bayesian method (Wang et al., 2007) with a confidence score threshold of 85% (i.e. in a bootstrap, the percentage of times that the sequence must match to the same taxonomy in order to be assigned a definitive taxonomic name), and the DNA reference library for diatoms Diat.barcode (formerly called R-Syst::diatom in Rimet et al., 2016). Only the DNA sequences belonging to diatoms (Bacillariophyta) are kept for further analysis. Subsequently, a similarity distance matrix is generated using the *dist.seqs* command. Based on this distance matrix, sequences belonging to closely related groups are clustered into OTUs using the furthest neighbor algorithm at a 95% similarity level. OTUs containing one single sequence (singletons) are removed, and a list of the OTUs and their relative abundances is produced for each of the samples based on read abundances per OTU.

Molecular taxa lists are then created by providing a taxonomy to each OTUs using the *classify.otu* command with a consensus confidence threshold of 80% (i.e. consensus taxonomy of ISU within each OTU) (Schloss et al., 2009). Finally, a DNA representative sequence is determined for each OTU using the *get.oturep* command in Mothur. Based on this workflow, sequencing data from the first sampling campaign (2016) was processed by changing different settings at different levels of the original workflow as described below.

2.4.1. Test on taxonomic assignment threshold of filtered ISU

The different tests were performed on 142 demultiplexed and overlapped *fastq* files delivered by the GeT-PlaGe sequencing platform (paired sequences overlap > 140 bp and mismatches < 0.1 %). Quality filter conditions for each *fastq* file remained equal to the classical bioinformatics treatment described previously except that the min length changed from 250 to 280 pb. After quality filtering, dereplication, and chimera removal, the resulting ISU were assigned a taxonomy using the Diat.barcode library (version 7 updated in May 2017 available at: https://www6.inra.fr/carrtel-collection_eng/Barcoding-database/) and the naïve Bayesian method (Wang et al., 2007). We tested three taxonomic assignment thresholds from loose stringency to high stringency: 60% (loose), 70% (intermediate), and 85% (high). A list of taxa and their relative abundances based on read abundances was produced for each of the samples for each taxonomic assignment threshold (60 inventory, 70 inventory and 85 inventory) (Figure 2). Several different methods are available in Mothur to assign a taxonomy to the sequences. We

selected the Bayesian method because of its accuracy and its swiftness (Wang et al., 2007) and also because is the default taxonomical assignment method proposed by Mothur.

2.4.2. Test on clustering sequences into OTUs

A similarity distance matrix was generated using the *dist.seqs* command for each *fasta* file resulting from the taxonomic assignment of the ISU at different taxonomic thresholds (60, 70, and 85%). Based on these distance matrices, reads were clustered into OTUs at a 95% similarity level. Two clustering algorithms were tested: Furthest Neighbor and OptiClust. While Mothur proposed several different algorithms to cluster DNA sequences into OTUs (Opticlust, average neighbor, furthest neighbor, nearest neighbor, Vsearch agc and Vsearch dgc), we chose to compare only these two algorithms. This is mainly because the Furthest neighbor has been used so far to generate diatom molecular inventories in previous studies (Vasselon et al., 2017b; Keck et al., 2018; Rivera et al. 2018a; 2018b) and because OPtiClust is a relatively new algorithm that can create more robust OTUs than other clustering methods (e.g. average neighbor, furthest neighbor, nearest neighbor, Vsearch agc, Vsearch dgc, Usearch agc, Usearch dgc, Sumaclust and Swarm) (Westcott and Schloss, 2017). Furthermore, OptiClust is the default clustering algorithm proposed by Mothur. After clustering, OTUs containing one-single sequence (singletons) were removed. A list of the OTUs and their relative abundances—based on read abundances per OTU— was produced for each of the samples for each clustering method. The results were compared to microscopy in terms of community structure.

2.4.3. Test on taxonomic assignment of OTUs

Molecular taxa lists were created for each clustering method by getting a consensus taxonomy for each OTU. This was done by using the *classify.otu* command. Two taxonomic assignment thresholds were tested: 60% (loose stringency) and 80% (high stringency). A list of taxa and their relative abundances based on read abundances was produced for each taxonomic assignment threshold for each clustering method (Inv.60_60_F, Inv.60_80_F, Inv.70_60_F, Inv.70_80_F, Inv.85_60_F, Inv.85_80_F, (Inv.60_60_O, Inv.60_80_O, Inv.70_60_O, Inv.70_80_O, Inv.85_60_O, Inv.85_80_O) (Figure 2).

2.4.4. Test on taxonomic assignment of raw ISU

Next, we tried to avoid sequence filtering and sequence clustering into OTUs to see if bioinformatics treatment could be simplified and generate molecular inventories for biomonitoring purposes. Here, we used the 142 *fastq* files provided by the platform and conducted a de-replication step skipping the quality filters and removing the chimeras. The resulting ISU were then assigned a taxonomy at a stringent threshold of 85% using the naive Bayesian method (Wang et al., 2007) and the Diat.barcode library (version 7 updated in May 2017). A list of taxa and their relative abundances based on raw ISU abundances was produced for each of the samples (Raw inventory) (Figure 2) and compared to the morphological inventory.

2.5. Comparison of bioinformatics strategies to microscopy

2.5.1. Comparison of diatom assemblages' structures of bioinformatic strategies to microscopy

The structure of the diatom assemblages obtained from both morphological and molecular approaches for each bioinformatics treatment was compared using a Mantel test (Pearson correlation coefficient). Diatom assemblages were expressed in relative abundance of species in each sample (relative

abundances based on frustules counts for microscopy and sequences reads for molecular data). Diatom assemblages obtained with microscopy and the 16 different bioinformatics strategies were compared with Bray-Curtis distance to produce distance matrices. These distances matrices were then used to perform Mantel tests between the morphological and the molecular floristic inventories (statistical software PAST 3.14, (Hammer *et al*., 2001)).

2.5.2. *Comparison of th*e water quality assessment

The molecular inventories resulting from each bioinformatics treatment as well as morphological inventories were used to calculate the IPS diatom index ("Indice de Polluosensibilité spécifique") (Cemagref, 1982). This diatom index is widely used in Europe and elsewhere for river quality assessment (Rimet, 2012). It classifies the ecological quality of water courses into five categories via a scale that ranges from 1 to 20 (1 - 4.9: *bad*; 5 - 8.9: *poor*; 9 - 12.9: *moderate*; 13 - 16.9: *good*; 17 - 20: *very good*). IPS was calculated via the OMNIDIA software version 6.0 (Lecointe et al., 1993).

A Spearman correlation test was performed between the molecular IPS scores obtained with each bioinformatics treatment and the morphological IPS scores to assess the effect of bioinformatics on water quality assessment. These analyses were performed in the statistical software R (version 3.5.2) using the R Stats Package (R Core Team, 2018).

2.6 Application of the best strategy to a large set of diatom samplings

The best bioinformatics strategy was identified as the one showing the highest correlations obtained with the Mantel tests (assemblages' structures) and the highest correlation for water quality assessment (IPS diatom indices). We first applied this bioinformatics strategy to assess the ecological quality of all the 447 samples collected in 2016 and 2017.

Second, we transformed the sequence abundances with the correction factor adapted to diatoms (Vasselon et al., 2018) to make the relative abundances of species from the molecular inventories more similar to those obtained with microscopy. Indeed, microscope analyses are based on frustule counts and do not consider the biovolum of species in the abundance assessment of species; sequence abundances from HTS depends on species biovolumes and their proportions in the sample (Vasselon et al., 2018). This factor considers the biovolume of species. We propose the modification below:

$$CFv2 = 10^{0.0703*(\log(species\ biovolume))^{2,4908}}$$

We then modified the sequence abundances:

$$modified\ sequence\ abundance = \frac{sequence\ abundance}{CFv2}$$

We calculated the diatom indices based on these modified sequences using OMNIDIA software version 6.0 (Lecointe et al., 1993). We then compared the results of the quality assessment obtained with unmodified and modified (CFv2) sequence abundances. Slopes obtained from the linear regressions between microscopic and molecular diatom indices (with or without transformation with CFv2) were compared in R software using the library lsmeans (for the ANOVA, we used the "anova" function, and for slopes comparison we used the "pairs" function). We used libraries psych and data.table for correlation coefficient comparison ("paired.r" function).

3. Results

For microscope analyses, 841 taxa —mostly identified at species level— were observed for a total of 364.398 frustules for the 447 samples from 2016 and 2017 sampling campaigns. For the molecular analyses, 20,588,593 sequences were obtained from 3 different runs (one for the samples carried out in 2016 and two for the samples carried out in 2017); the three runs were of good quality and could be used for subsequent analyses.

3.1 Comparison of diatom species compositions

The dominant taxa detected with the 16 bioinformatics strategies were similar. However, there were important differences in the proportions of taxa after the taxonomic assignment of OTUs created with OptiClust algorithm. Indeed, *Achnanthidium* sp., *Gomphonema* sp., *Achnanthidum pyrenaicum* and *Nitszchia* sp. were detected in greater proportions compared to the other bioinformatics strategies (Figure 3).

The number of detected species varied across bioinformatics strategies (Figure 4). Taxonomic assignment of raw ISU resulted in the detection of a higher number of species compared to the other treatments followed by the taxonomic assignment of filtered ISU at a threshold of 60%.

The proportion of unclassified sequences also varied across bioinformatics strategies. The greatest number of unclassified sequences was obtained with taxonomic assignment of raw ISU (Figure 4). The taxonomic assignment of filtered ISU at a threshold of 60% resulted in the smallest number of unclassified sequences compared to the remaining bioinformatics treatments. The number of unclassified sequences resulting from the taxonomic assignment of OTUs created with OptiClust algorithm were very different depending on the OTUs assignment threshold) (Figure 4).

3.2 Comparison of assemblages' structures

We tested 16 bioinformatic treatments and found that the molecular inventory resulting from the taxonomic assignment of filtered ISU at a threshold of 60% correlated better to the morphological inventory according to the Mantel test results (R_{60}= 0.60, Figure 5). The weakest correlation was with taxonomic assignment of OTUs created with OptiClust algorithm at an assignment threshold of 85 and 80% ($R_{85\_80\_O}$= 0.37, Figure 5).

The number of generated OTUs differed depending on the clustering algorithm and the sequence taxonomic assignment threshold. The furthest neighbor created fewer OTUs than OptiClust (Figure 6) and allowed taxonomic assignment of a greater number of taxa (Figure 4). Furthermore, the furthest neighbor provided a slightly better characterization of diatom communities than OptiClust.

3.3 Comparison of quality assessment

Morphological and molecular IPS scores obtained with each bioinformatic strategy were compared using Pearson's correlation coefficient. The best correlation was obtained with the IPS scores calculated from the molecular inventory resulting from the taxonomic assignment of filtered ISU at a loose threshold of 60% (IPS_60; R^2= 0.60) (Figure 7). The worst correlation was obtained with IPS values calculated from molecular inventories of raw ISU (IPS_Raw; R^2= 0.14).

Tables 1 and 2 summarize the results of the statistical analyses given above. Table 1 shows that when the stringency of the taxonomic assignment threshold increases from 60% to 85%, the number of unclassified sequences increased. In contrast, the number of detected species, together with the correlation between metabarcoding and microscopy diatom assemblages as well as IPS scores decreased. Table 2 shows that the number of unclassified sequences is lower for filtered ISU and Furthest Neighbor strategies, while the number of detected species is higher for raw and filtered ISU strategies. Correlation between diatom indices obtained in microscopy and metabarcoding is higher with filtered ISU and Furthest Neighbor strategies. Correlation between diatom assemblages obtained in metabarcoding and microscopy is lower with the Opticlust strategy. Finally, when comparing calculation times, we observe that the filtered ISU strategy is two times longer than the raw ISU strategy and that the Furthest Neighbor and Opticlust strategies are at least five times longer than the raw ISU strategy.

3.4 Application of the best bioinformatics strategy to a large set of diatom samplings

The best bioinformatics strategy was the one based on filtered ISU with the loose taxonomic assignment threshold (60%). We applied this selected strategy to calculate the IPS values for all 447 samples (campaigns 2016 and 2017) (Figure 8). We also transformed the quantification of the molecular inventories with CFv2 (based on species biovolumes) and calculated the IPS values for all samples again. We then compared these two strategies: the correlation coefficients to microscopy of both methods are not significantly different ($p>0.05$); however, the slope of the data transformed with CFv2 is significantly higher (ANOVA, $p<0.001$; slope comparison $p < 0.001$).

We then compared the water quality classes obtained from microscope counts to the quality classes obtained with this bioinformatics strategy: one is based on non-transformed data (Table 3a), and the other is based on data transformed with the correction factor CFv2 (Table 3b). 64% of the samples were assigned to the same quality class with the untransformed data; 66% were in the same quality class with the transformed data (CFv2).

4. Discussion

4.1 Diatom species compositions obtained in microscopy differed from those obtained in metabarcoding

Of all the produced inventories, the one obtained with microscopy appears to be the most distinct from all others produced with metabarcoding in terms of number of detected taxa and in terms of relative abundances of the taxa. Even if most of the dominant species detected in metabarcoding were the same than those observed in microscopy, there was a difference in terms of abundances. The dominant species observed in microscopy were small species such as *Achnanthidum minutissimum, A. pyrenaicum*, and *Amphora pediculus*; those in metabarcoding had large biovolumes such as *Melosira varians*. This is because diatom taxa abundance is calculated from the number of DNA reads in metabarcoding, whereas in microscopy it is calculated from the number of individuals (frustules). The number of copies of the *rbc*L gene is correlated to cell biovolume; hence, metabarcoding overestimates the abundances of big species compared to small ones in comparison to morphology (Vasselon et al., 2018). To limit this difference, a correction factor was proposed (Vasselon et al., 2018) to transform the proportion of sequences to enable a better comparability between morphological and molecular inventories. This correction factor was applied in the framework of this study allowing a

better assessment of the relative abundance of species obtained with HTS in a more similar way to microscopy (see section 4.3).

The overall number of species determined in microscopy was much greater than the number of species detected in metabarcoding. This is due to several reasons. First, diatom frustules from dead cells in the collected biofilms can be detected in microscopy but not in metabarcoding because the DNA is already degraded. This has already been observed by Kermarrec et al. (2014) in rivers and by Rivera et al. (2018) in lakes. Second, the reference barcoding library is incomplete. Indeed, a significant proportion of species observed in microscopy could not be detected in metabarcoding because their barcode was not present in the Diato.barcode (version 7) despite a significant effort to complete it (Rimet et al., 2018). Third, microscope determinations were carried out by people from different laboratories with potentially differing identification skills as already shown in inter-calibration exercises (Kahlert et al., 2009). This artificially increase the number of species detected in microscopy. Fourth, resolution of the *rbc*L barcode (312 bp) might not be sufficient to distinguish all taxa. In some cases, we can probably only identify taxa at genus level. Fifth, the sequencing depth might not be sufficient to properly detect the full diatom diversity—especially regarding low abundant and small taxa. This is not a problem for water quality assessment since biotic indices values mostly depend on abundant taxa, but this may impact the number of species detected (Zaheer et al., 2018). Regardless of the bioinformatics strategy used, these reasons make microscopic and metabarcoding analyses different.
However, we could have obtained an opposite result where the number of species detected with metabarcoding may be larger than microscopy. Indeed, the presence of persisting free-floating DNA (extracellular DNA) coming from diatoms cells living in the upper part of the sampling sites may distort the results since this free DNA will be detected in metabarcoding but not in microscopy. Furthermore, microscopy might not be sufficient to detect all the biodiversity present in the sample since morphological counts are limited to 400 valves compared to metabarcoding which provides thousands of sequences for a single sample. In our study we analysed 364.398 frustules vs. 20,588.593 sequences, the microscopic depth is 56 times lower. Despite this, microscopy is still the gold standard for water managers at present.

4.2 Compared to microscopy, some bioinformatics strategies gave more similar assemblage structures and water quality assessments

To the best of our knowledge, apart from the study of Tapolczai et al. (2019), there are no studies comparing different bioinformatics treatments of diatom sequencing data for monitoring purposes. Here, we compared 16 bioinformatics strategies to microscopy in terms of diatom assemblages' structure and water quality assessment. We noted differences in terms of species detected, community structures, and water quality depending on the strategy selected. Some of the tested strategies should be avoided while others are preferred to keep our results comparable to microscopy.

4.2.1 Which OTU clustering algorithm was the best?

Molecular inventories resulting from the taxonomic assignment of OTU data created with the Furthest neighbor algorithm gave the most similar results to microscope inventories in terms of structure of diatom assemblages and water quality assessment compared to the OptiClust algorithm. Opticlust is widely used in virology (e.g. Romano et al. (2017)), medicine (e.g. Wong et al. (2017)), and ecology (Probandt et al., 2018), and few studies have assessed its capacities compared to other algorithms (Westcott and Schloss, 2017). These results indicate that the Furthest Neighbor is recommended in our case, which confirms a previous decision to use it for diatom metabarcoding in biomonitoring (Keck et al., 2018) and Rivera et al. (2018b). However, in another ecological context, a recent study using OptiClust as clustering algorithm provided coherent results between morphological and molecular water

quality assessment using diatoms (Mortágua et al., 2019). The results of our work show that the recommended OTU assignment threshold for the establishment of molecular inventories using the Furthest Neighbor as a clustering algorithm has a less stringent taxonomic threshold (60% for both sequences and OTU assignments).

4.2.2 Was OTU clustering necessary for diatom biomonitoring?

We observed good correlation between microscope inventories and inventories obtained from bioinformatic strategies calculating OTUs. However, we observed same good correlations between morphological and molecular IPS scores with the simple strategy using filtered ISU and a loose taxonomic assignment threshold of 60%. This shows that we can bypass the OTU calculation step to establish a molecular-based inventory for biotic indices. This saves time and computing power during bioinformatics data processing because the similarity distance matrix calculation is avoided. This result is in the same line of the strategies followed by recent pipelines that do not cluster sequences into OTUs like DADA2 (Callahan et al., 2016) where the authors show that OTUs underutilize the quality of modern sequencing (like Illumina technology) by "precluding the possibility of resolving fine-scale variation"; this variation can be important for ecological studies. Moreover, the number of taxa taxonomically assigned with filtered ISU was higher than with OTUs strategies. These additional taxa were important to consider because this strategy produces a water quality assessment that is closer to that obtained with the microscope.

4.2.3 Was ISU filtering necessary for diatom biomonitoring?

The diatom assemblages could be nicely characterized by simplifying to the extreme sequence processing and directly assigning the ISU without any quality filters (raw data). The results were quite comparable to those obtained in microscopy. However, IPS scores resulting from this bioinformatic strategy were badly correlated to the IPS morphological scores compared to all other strategies. Taxonomic assignment of filtered ISU showed a slightly better correlation to microscopy in terms of structure and water quality assessment regardless of the taxonomic assignment threshold selected. This means that for biomonitoring purposes, sequences must be filtered in terms of quality, length, and chimeras should be removed. If not, the results are far from what is expected by standardized biomonitoring approaches currently based on microscopy.

4.2.4 Shall we select a stringent or a loose taxonomic assignment threshold?

The taxonomic assignment thresholds (minimum percentage of times that a sequence must match the same taxonomy in order to be assigned) played an important role in the final molecular inventories. Loose assignment thresholds imply a greater ability to detect species from an environmental sample but with a higher probability of misallocation of the taxonomic name. On the other hand, with a stringent assignment threshold, the ability to detect species from an environmental will be reduced because the individuals in the environment will be assigned only if they are very similar to those in the reference database (in terms of barcode sequence). In return, we will be more confident in the identification. Indeed, the number of detected species decreased when the stringency of the taxonomic assignment threshold increased (60, 70 to 85%; see Table 1). Similarly, the correlation between diatom assemblages obtained via metabarcoding and microscopy decreased. The same was observed for diatom indices. This

indicates that flexibility is important for an efficient identification, and thus the assignment threshold should remain loose (i.e. 60%).

These results should be seen in the perspective of phylogeny and ecology of diatoms: phylogenetically related diatom species have a better chance of sharing similar ecologies (Keck et al., 2016). In particular, one can predict the ecology of unassigned sequences from the ecology of their phylogenetically-related species (Keck et al., 2018). In our case, we showed that it is preferable to have a rather flexible identification in a biomonitoring framework (loose threshold i.e., 60%), to detect more species, even if some may be badly identified. This makes it possible to give a species name to more environmental sequences, and thus to have a more robust diatomic index value, i.e., because it will be based on a larger number of environmental sequences. It is better to keep the sequences misidentified to a phylogenetically neighbor species than not identifying the species at all. This is because neighbor species usually share the same pollution sensitivities, and such information is important to keep for diatom index calculations.

4.3 Application to a large monitoring scale

In order to calculate the diatom indices values (IPS) on the large monitoring data set of 447 samples, we selected the filtered ISU strategy because it gave the most similar results to microscopy; and we chose the loose (60%) taxonomic assignment threshold since it gave the best results. The correlation between the IPS values obtained in metabarcoding and microscopy was high (R^2: 69%). This correlation was similar when we transformed the sequence abundances with a correction factor that considers biovolumes of species (Vasselon et al., 2018); however, the slope of the correlation was closer to 1, which made this last strategy even more comparable to microscopy. The percentage of sampling sites sharing the same quality class between microscopy and metabarcoding was high (64% for non-transformed data and 66% for CFv2 transformed data); 72% of cases had an index value difference between microscopy and metabarcoding less than 1 point (the IPS ranged from 0 to 20 points). Our metabarcoding results are much more similar to microscopy than prior biomonitoring works (e.g., Vasselon et al., 2017b; Rivera et al., 2018). This was made possible by progressive methodological developments in different areas: barcode selection (Kermarrec et al., 2013), DNA extraction methodology (Vasselon et al., 2017a), update of *rbc*L primers (Vasselon et al., 2017b), quantification correction factors based on species biovolumes (Vasselon et al., 2018), and completion of the reference database (F. Rimet et al., 2018).

5. Conclusions and perspectives

These optimizations demonstrate how metabarcoding can complement or even replace microscopic analyses for biomonitoring (e.g., Hering et al., 2018), but some work remains. First, reference barcoding libraries are still incomplete (Weigand et al., 2019), and a concerted international effort is needed such as the Diat.barcode initiative (Rimet et al., 2018; international initiative to curate and complete a reference library of barcodes for diatoms). Hopefully, protocols will soon be transferred to water managers and companies in charge of aquatic ecosystem monitoring (Hering et al., 2018). This process started for diatoms according to acceptance from the European Standardization Committee of protocols for diatom sampling and reference barcoding libraries (CEN, 2018a; CEN, 2018b), but all other items in the workflow still need to be standardized. The DNA-based methods for diatom water quality assessment will enter the era of routine use and will surely change the way water managers work (Keck et al., 2017).

Acknowledgements

This study was funded by the AFB (Agence Française pour la Biodiversité). We thank the French Water Agencies, the DREAL (Direction Régionale de l'Environnement, de l'Aménagement et du Logement) and the private offices (Aquabio, Aquascop, Eurofins, Grebe, SAGE) for sampling collection and morphological analyses. We also thank Cécile Chardon for the DNA extraction, PCR, and library preparation. We thank the COST action DNAqua-Net (CA 15219) funded by the European Union for helpful support. We also thank the "SYNAQUA" project supported by the European Cross-Border Cooperation Program (Interreg France-Switzerland 2014-2020) that funded the study of Swiss border rivers by a European grant (ERDF, European Regional Development Fund) and a Swiss grant (from the Cantons of Valais, Geneva, Vaud and the Swiss Confederation).

Supplementary data

Floristic lists: Rivera, Sinziana; Vasselon, Valentin; Chardon, Cécile; Jacas, Louis; Guéguen, Julie; Bouchez, Agnès; Rimet, Frederic, 2018, "Bioindication diatomées : comparaison microscopie / barcoding ADN. Listes floristiques, indices IBD. Projet AFB numéro 15000239 / A30.", https://doi.org/10.15454/WNI6FQ, Portail Data Inra, V1

https://data.inra.fr/dataset.xhtml?persistentId=doi%3A10.15454%2FWNI6FQ

Raw sequencing data (Fastq files): Rivera, Sinziana; Vasselon, Valentin; Chardon, Cécile; Guéguen, Julie; Bouchez, Agnès; Rimet, Frédéric, 2018, "Bioindication diatomées : comparaison microscopie / barcoding ADN. Données brutes Fastq 2016 + 2017, Test des différentes stratégies bioinfo sur données 2016. Projet AFB numéro 15000239 / A30.", https://doi.org/10.15454/9EG5Z4, Portail Data Inra, V1

https://data.inra.fr/dataset.xhtml?persistentId=doi%3A10.15454%2F9EG5Z4

Bioinformatic pipeline (Mothur) selected for the calculation of the índices values of the 447 sites: Rivera, Sinziana; Vasselon, Valentin; Chardon, Cécile; Jacas, Louis; Guéguen, Julie; Bouchez, Agnès; Rimet, Frédéric, 2019, "Bioindication diatomées : comparaison microscopie / barcoding ADN. Pipeline MOTHUR selectionné", https://doi.org/10.15454/1OTGWL, Portail Data Inra, V1

https://data.inra.fr/dataset.xhtml?persistentId=doi%3A10.15454%2F1OTGWL

References

Afnor, 2014a. EN 13946. Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers. Afnor 1–17.

Afnor, 2014b. EN 14407 - Water quality Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. CEN Stand. 1–13.

Afnor, 2010. Guide pour l'étude, l'échantillonnage et l'analyse en laboratoire du phytobenthos dans les cours d'eau peu profonds. NF EN 15708. Afnor 1–21.

Afnor, 2007. NF T90-354. Qualité de l'eau - Détermination de l'Indice Biologique Diatomées (IBD). Afnor 1–79.

Al-Neama, M., Reda, N., Ghaleb, F., 2014. An Improved Distance Matrix Computation Algorithm for Multicore Clusters. BioMed Res. Int. 2014, 406178. https://doi.org/10.1155/2014/406178

562 Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., Pawlowski, J., 2017. Taxonomy-
563 free molecular diatom index for high-throughput eDNA biomonitoring. Mol. Ecol. Resour. 17,
564 1231–1242.
565 Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High
566 resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583.
567 https://doi.org/10.1038/nmeth.3869
568 Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña,
569 A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E.,
570 Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh,
571 P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows
572 analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336.
573 https://doi.org/10.1038/nmeth.f.303
574 Cemagref, 1982. Etude des methodes biologiques quantitative d'appreciation de la qualite des eaux.
575 CEN, 2018a. Water quality - CEN/TR 17245 - Technical report for the routine sampling of benthic
576 diatoms from rivers and lakes adapted for metabarcoding analyses. CEN Stand. 1–8.
577 CEN, 2018b. Water quality - CEN/TR 17244 - Technical report for the management of diatom barcodes
578 1–11.
579 Chonova, T., Keck, F., Labanowski, J., Montuelle, B., Rimet, F., Bouchez, A., 2016. Separate treatment of
580 hospital and urban wastewaters: a real scale comparison of effluents and their effect on
581 microbial communities. Sci. Total Environ. 542, 965–975.
582 Deiner, K., Walser, J.-C., Mächler, E., Altermatt, F., 2015. Choice of capture and extraction methods
583 affect detection of freshwater biodiversity from environmental DNA. Biol. Conserv., Special
584 Issue: Environmental DNA: A powerful new tool for biological conservation 183, 53–63.
585 https://doi.org/10.1016/j.biocon.2014.11.018
586 Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods
587 10, 996–998. https://doi.org/10.1038/nmeth.2604
588 Esling, P., Lejzerowicz, F., Pawlowski, J., 2015. Accurate multiplexing and filtering for high-throughput
589 amplicon-sequencing. Nucleic Acids Res. 43, 2513–2524. https://doi.org/10.1093/nar/gkv107
590 European Commission, 2000. Directive 2000/60/EC of the European Parliament and of the Council of
591 23rd October 2000 establishing a framework for Community action in the field of water policy.
592 Off. J. Eur. Communities 327, 1–72.
593 Hammer, Ø., Harper, D.A.T., Ryan, P.D., 2001. PAST: Paleontological Statistics Software Package for
594 Education and Data Analysis. Palaeontologia Electronica, 4: 1-9.
595 Hering, D., Borja, A., Jones, J.I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B.,
596 Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., Kelly, M.,
597 2018. Implementation options for DNA-based identification into ecological status assessment
598 under the European Water Framework Directive. Water Res. 138, 192–205.
599 https://doi.org/10.1016/j.watres.2018.03.003
600 Holovachov, O., Haenel, Q., Bourlat, S.J., Jondelius, U., 2017. Taxonomy assignment approach
601 determines the efficiency of identification of OTUs in marine nematodes. R. Soc. Open Sci. 4.
602 https://doi.org/10.1098/rsos.170315
603 Kahlert, M., Albert, R.L., Anttila, E.L., Bengtsson, R., Bigler, C., Eskola, T., Galman, V., Gottschalk, S.,
604 Herlitz, E., Jarlman, A., Kasperoviciene, J., Kokocinski, M., Luup, H., Miettinen, J., Paunksnyte, I.,
605 Piirsoo, K., Quintana, I., Raunio, J., Sandell, B., Simola, H., Sundberg, I., Vilbaste, S., Weckstrom,
606 J., 2009. Harmonization is more important than experience-results of the first Nordic-Baltic
607 diatom intercalibration exercise 2007 (stream monitoring). J. Appl. Phycol. 21, 471–482.
608 Keck, F., Bouchez, A., Franc, A., Rimet, F., 2016. Linking phylogenetic similarity and pollution sensitivity
609 to develop ecological assessment methods: a test with river diatoms. J. Appl. Ecol. 53, 856–864.

610 Keck, F., Vasselon, V., Rimet, F., Bouchez, A., Kahlert, M., 2018. Boosting DNA metabarcoding for
611 biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles.
612 Mol. Ecol. Resour. 0. https://doi.org/10.1111/1755-0998.12919
613 Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., Bouchez, A., 2017. Freshwater biomonitoring in the
614 Information Age. Front. Ecol. Environ. 1–9.
615 Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J.-F., Bouchez, A., 2014. A next-
616 generation sequencing approach to river biomonitoring using benthic diatoms. Freshw. Sci. 33,
617 349–363.
618 Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.-F., Bouchez, A., 2013. Next-generation
619 sequencing to inventory taxonomic diversity in eukaryotic communities: A test for freshwater
620 diatoms. Mol. Ecol. Resour. 13, 607–619.
621 Krammer, K., Lange-Bertalot, H., 1991a. Bacillariophyceae 4. Teil: Achnanthaceae. Kritische Erganzungen
622 zu Navicula (Lineolatae) und Gomphonema. Gesamtliteraturverzeichnis Teil 4.
623 Krammer, K., Lange-Bertalot, H., 1991b. Bacillariophyceae 3. Teil: Centrales, Fragilariaceae, Eunotiaceae.
624 Krammer, K., Lange-Bertalot, H., 1988. Bacillariophyceae 2. Teil: Bacillariaceae, Epithemiaceae,
625 Surirellaceae.
626 Krammer, K., Lange-Bertalot, H., 1986. Bacillariophyceae 1. Teil: Naviculaceae.
627 Lange-Bertalot, H., Hofmann, G., Werum, M., Cantonati, M., 2017. Freshwater benthic diatoms of
628 central Europe: over 800 common species used in ecological assessment. English edition with
629 updates taxonomy and added species., Koltz Botanical Books, Schmitten-Oberreifenderg,
630 Germany. ed. Cantonati, M., Kelly, M.G., Lange-Bertalot, H.
631 Lecointe, C., Coste, M., Prygiel, J., 1993. "Omnidia": Software for taxonomy, calculation of diatom
632 indices and inventories managment. Hydrobiologia 269/270, 509–513.
633 Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., Mergen, P., Pawlowski, J.,
634 Piggott, J., Rimet, F., Steinke, D., Taberlet, P., Weigand, A., Abarenkov, K., Beja, P., Bervoets, L.,
635 Björnsdóttir, S., Boets, P., Boggero, A., Bones, A., Borja, Á., Bruce, K., Bursić, V., Carlsson, J.,
636 Čiampor, F., Čiamporová-Zatovičová, Z., Coissac, E., Costa, F., Costache, M., Creer, S., Csabai, Z.,
637 Deiner, K., DelValls, Á., Drakare, S., Duarte, S., Eleršek, T., Fazi, S., Fišer, C., Flot, J.-F., Fonseca,
638 V., Fontaneto, D., Grabowski, M., Graf, W., Guðbrandsson, J., Hellström, M., Hershkovitz, Y.,
639 Hollingsworth, P., Japoshvili, B., Jones, J., Kahlert, M., Stroil, B.K., Kasapidis, P., Kelly, M., Kelly-
640 Quinn, M., Keskin, E., Kõljalg, U., Ljubešić, Z., Maček, I., Mächler, E., Mahon, A., Marečková, M.,
641 Mejdandzic, M., Mircheva, G., Montagna, M., Moritz, C., Mulk, V., Naumoski, A., Navodaru, I.,
642 Padisák, J., Pálsson, S., Panksep, K., Penev, L., Petrusek, A., Pfannkuchen, M., Primmer, C.,
643 Rinkevich, B., Rotter, A., Schmidt-Kloiber, A., Segurado, P., Speksnijder, A., Stoev, P., Strand, M.,
644 Šulčius, S., Sundberg, P., Traugott, M., Tsigenopoulos, C., Turon, X., Valentini, A., Hoorn, B. van
645 der, Várbíró, G., Hadjilyra, M.V., Viguri, J., Vitonytė, I., Vogler, A., Vrålstad, T., Wägele, W.,
646 Wenne, R., Winding, A., Woodward, G., Zegura, B., Zimmermann, J., 2016. DNAqua-Net:
647 Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in
648 Europe. Res. Ideas Outcomes 2, e11321. https://doi.org/10.3897/rio.2.e11321
649 Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M., 2014. Swarm: Robust and fast clustering
650 method for amplicon-based studies. PeerJ 2, e593.
651 Majaneva, M., Hyytiäinen, K., Varvio, S.L., Nagai, S., Blomster, J., 2015. Bioinformatic amplicon read
652 processing strategies strongly affect eukaryotic diversity and the taxonomic composition of
653 communities. PLOS ONE 10, e0130035. https://doi.org/10.1371/journal.pone.0130035
654 Mortágua, A., Vasselon V., Oliveira R., Elias C., Chardon C., Bouchez A., Rimet F., João Feio M., Almeida S.
655 2019. Applicability of DNA metabarcoding approach in the bioassessment of Portuguese rivers
656 using diatoms. Ecological Indicators 106: 105470.
657 Probandt, D., Eickhorst, T., Ellrott, A., Amann, R., Knittel, K., 2018. Microbial life on a sand grain: from
658 bulk sediment to single grains. Isme J. 12, 623–633. https://doi.org/10.1038/ismej.2017.197

R Core Team, 2018. R: A language and environment for statistical computing., R Foundation for Statistical Computing, Vienna, Austria.

Reichardt, E., 1997. Taxonomic revision of the species complex involving Gomphonema pumilum (Bacillariophyceae). Nova Hedwig. 65, 99–129.

Rimet, F., 2012. Recent views on river pollution and diatoms. Hydrobiologia 683, 1–24.

Rimet, Frederic, Abarca, N., Bouchez, A., Kusber, W.-H., Jahn, R., Kahlert, M., Keck, F., Kelly, M.G., Mann, D.G., Piuz, A., 2018. The potential of High-Throughput Sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. Fottea 18, 37–54.

Rimet, F., Abarca, N., Bouchez, A., Kusber, W.H., Jahn, R., Kahlert, M., Keck, F., Kelly, M.G., Mann, D.G., Piuz, A., Trobajo, R., Tapolczai, K., Vasselon, V., Zimmermann, J., 2018. The potential of high throughput sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. Fottea. https://doi.org/doi 10.5507/fot.2017.013

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A., Bouchez, A., 2016. R-Syst: diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. Database 2016.

Rimet, F., Gusev, E., Kahlert, M., Kelly, M., Kulikovskiy, M., Maltsev, Y., Mann, D., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., Bouchez, A., 2018. Diat.barcode, an open-access barcode library for diatoms. https://doi.org/10.15454/TOMBYZ

Rivera, S.F., Vasselon, V., Ballorain, K., Carpentier, A., Wetzel, C.E., Ector, L., Bouchez, A., Rimet, F., 2018. DNA metabarcoding and microscopic analyses of sea turtles biofilms: Complementary to understand turtle behavior. PloS One 13, e0195770.

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., Rimet, F., 2018. Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. Hydrobiologia 807, 37–51.

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ 4: e2584.

Romano, K.A., Martinez-del Campo, A., Kasahara, K., Chittim, C.L., Vivas, E.I., Amador-Noguez, D., Balskus, E.P., Rey, F.E., 2017. Metabolic, epigenetic, and transgenerational effects of gut bacterial choline consumption. Cell Host Microbe 22, 279–290. https://doi.org/10.1016/j.chom.2017.07.021

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., 2009. Introducing Mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Tapolczai K., Valentin 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.

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 Ecol. Evol. 9, 1060–1069.

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

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

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. Environ. Sci. Technol. 49, 7597–7605.

709 Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of
710 rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267.
711 Weigand, H., Beermann, A.J., Čiampor, F., Costa, F.O., Csabai, Z., Duarte, S., Geiger, M.F., Grabowski, M.,
712 Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A.M., Willassen, E., Wyler, S.A., Bouchez,
713 A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, S., Dijkstra, K.-D.B., Eisendle, U., Freyhof, J.,
714 Gadawski, P., Graf, W., Haegerbaeumer, A., van der Hoorn, B.B., Japoshvili, B., Keresztes, L.,
715 Keskin, E., Leese, F., Macher, J.N., Mamos, T., Paz, G., Pešić, V., Pfannkuchen, D.M.,
716 Pfannkuchen, M.A., Price, B.W., Rinkevich, B., Teixeira, M.A.L., Várbíró, G., Ekrem, T., 2019.
717 DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and
718 recommendations for future work. Sci. Total Environ. 678, 499–524.
719 https://doi.org/10.1016/j.scitotenv.2019.04.247
720 Westcott, S.L., Schloss, P.D., 2017. OptiClust, an improved method for assigning amplicon-based
721 sequence data to operational Taxonomic Units. mSphere 2, e00073-17.
722 https://doi.org/10.1128/mSphereDirect.00073-17
723 Wong, S.H., Zhao, L., Zhang, X., Nakatsu, G., Han, J., Xu, W., Xiao, X., Kwong, T.N.Y., Tsoi, H., Wu, W.K.K.,
724 Zeng, B., Chan, F.K.L., Sung, J.J.Y., Wei, H., Yu, J., 2017. Gavage of fecal samples from patients
725 with colorectal cancer promotes intestinal carcinogenesis in germ-free and conventional mice.
726 Gastroenterology 153, 1621-+. https://doi.org/10.1053/j.gastro.2017.08.022
727 Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S.R., Marinier, E., Van Domselaar, G., Belk, K.E., Morley, P.S.,
728 McAllister, T.A., 2018. Impact of sequencing depth on the characterization of the microbiome
729 and resistome. Sci. Rep. 8, 5890. https://doi.org/10.1038/s41598-018-24280-8
730 Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., Gemeinholzer, B., 2015. Metabarcoding vs.
731 morphological identification to assess diatom diversity in environmental studies. Mol. Ecol.
732 Resour. 15, 526–542.
733

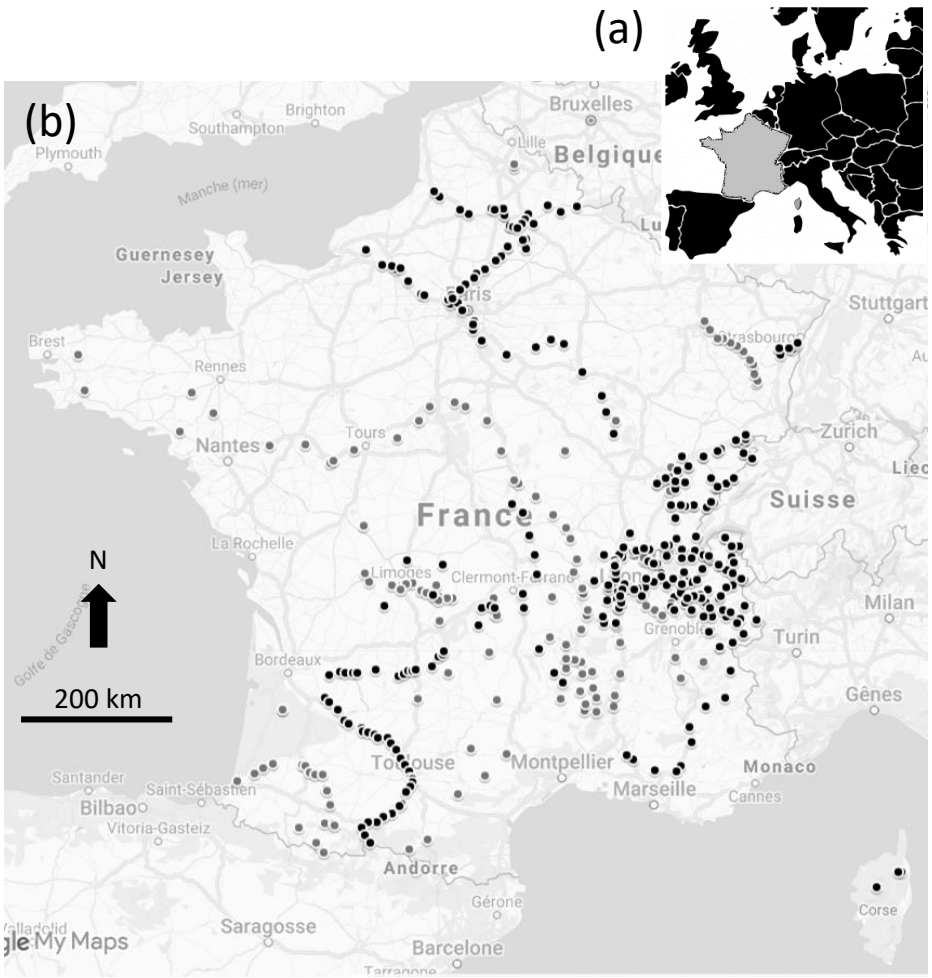


Figure 1. Geographic location of France in Europe (a) and geographic location of the sampling sites in France (b). Grey dots indicate sites sampled in 2016, black dots indicate sites sampled in 2017.

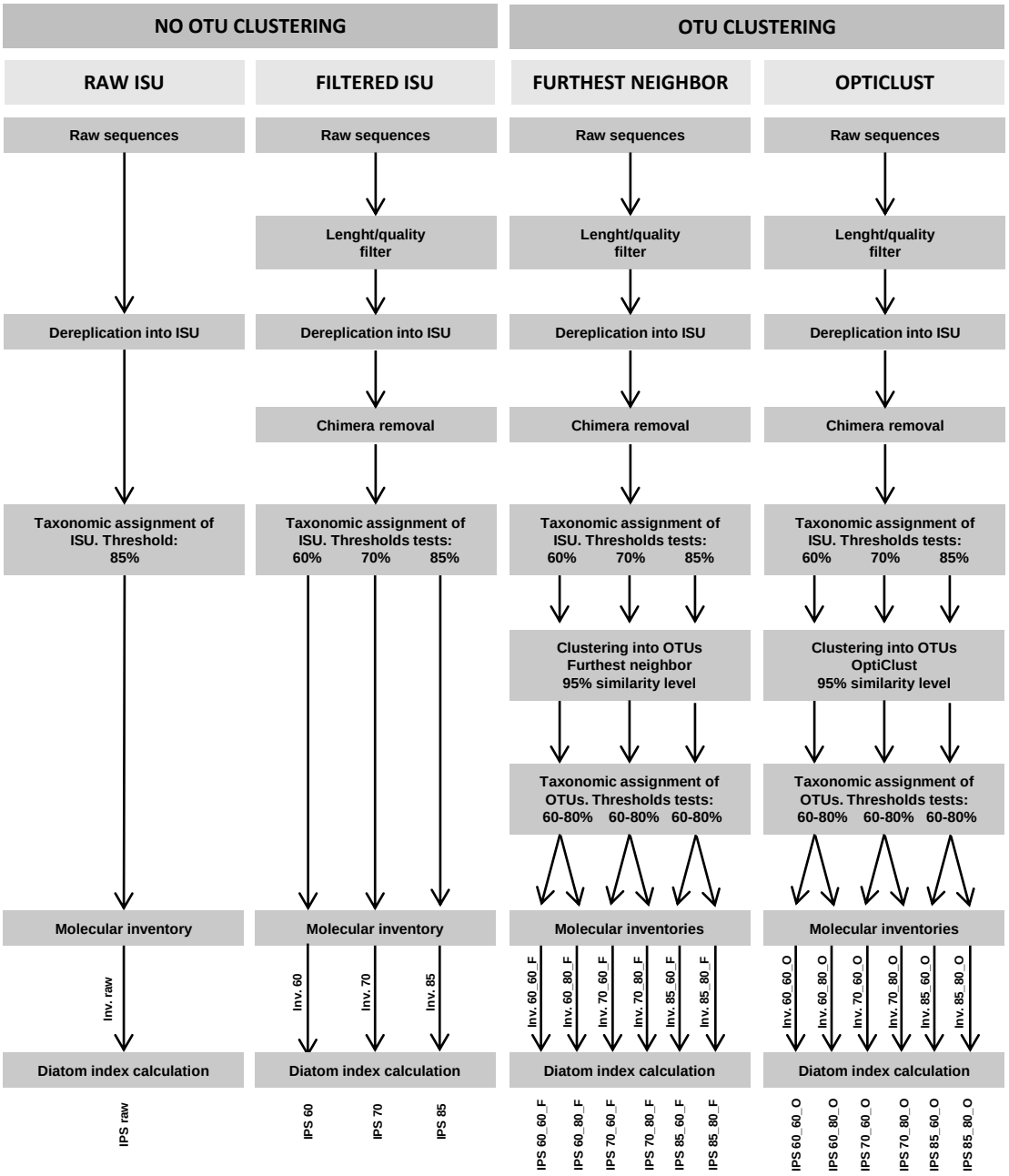


Figure 2. Overview of the 16 bioinformatic strategies tested. The bioinformatics strategy used in Keck et al. (2018) corresponds to Inv.85_80_F. Detailed descriptions of each bioinformtics strategy is given in section 2,4

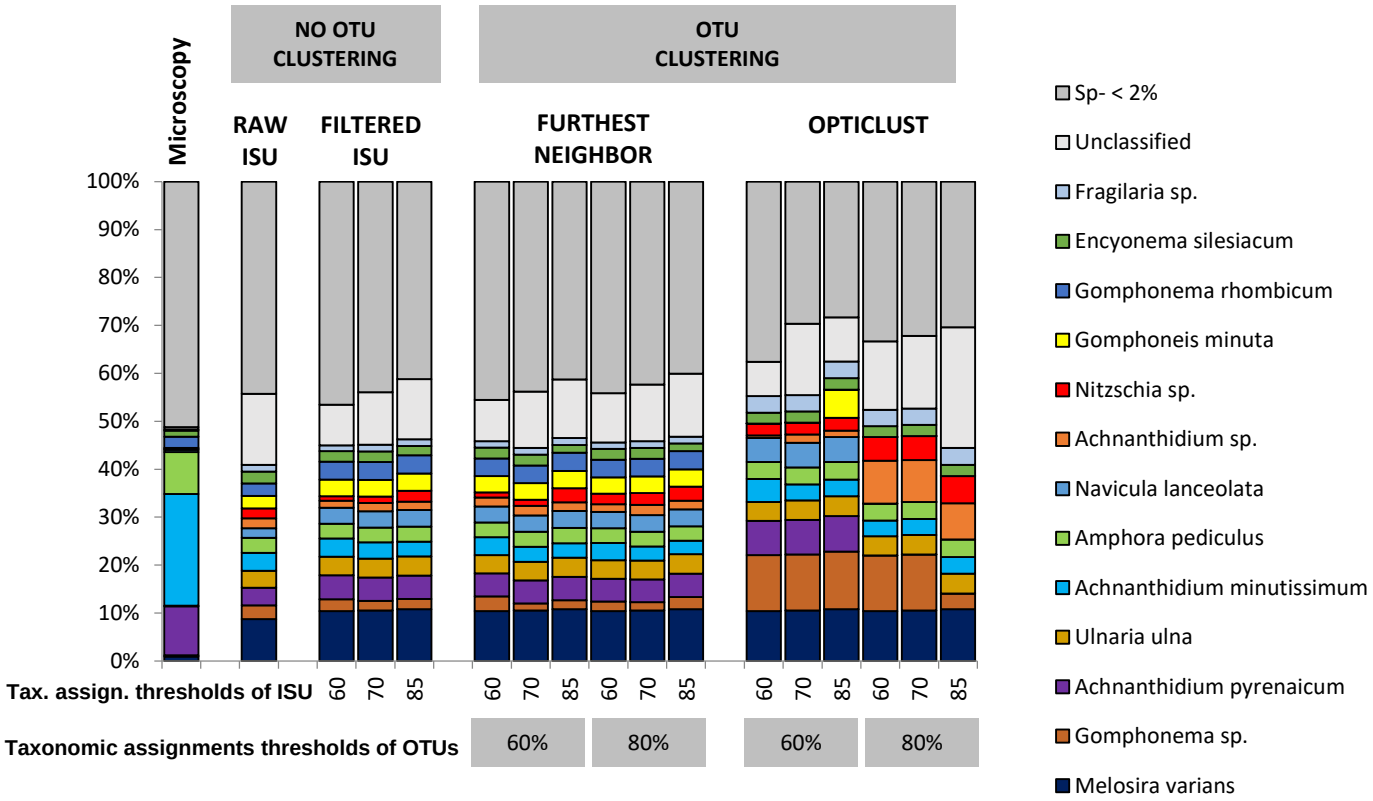


Figure 3. Relative abundances of diatom taxa (genera and species) detected with microscopy and with the 16 bioinformatics strategies. Only taxa with proportions over 2% are given. Even if dominant taxa are similar between bioinformatics strategies there are considerable differences in the proportion of taxa obtained with the OptiClust algorithm compared to the other bioinformatics strategies. Microscopy also gives very different proportions of various taxa compared to all the bioinformatics strategies.

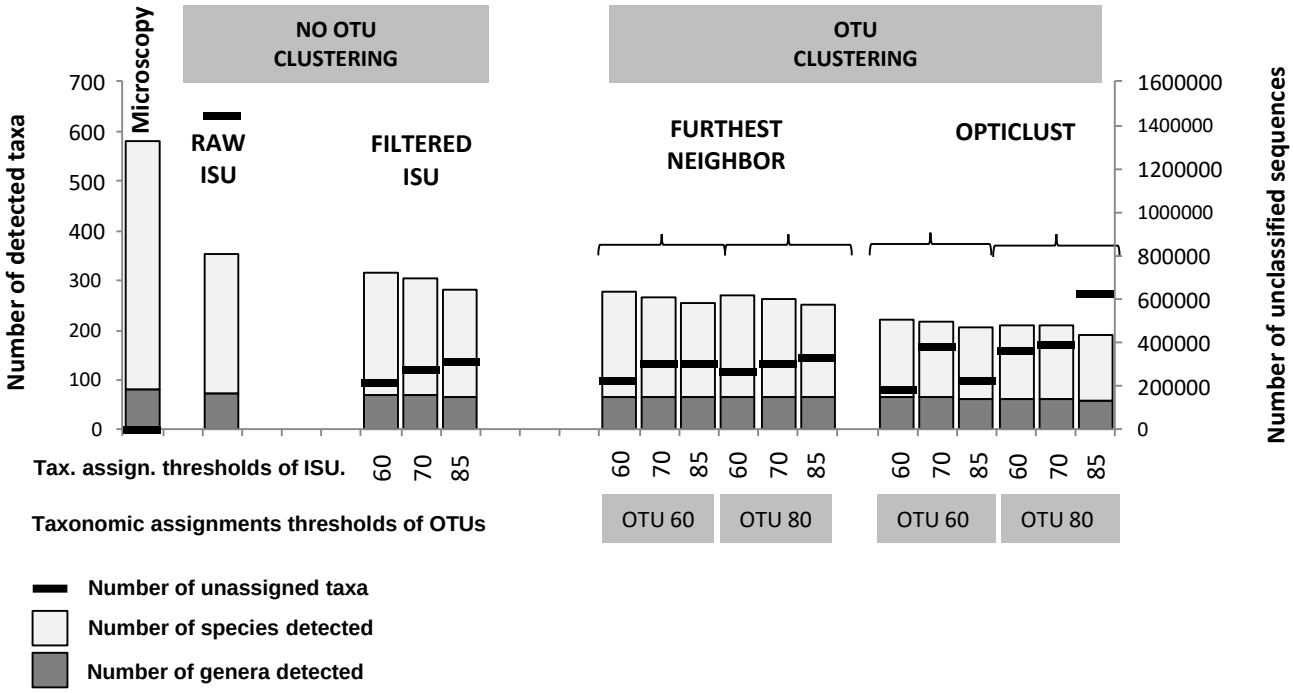


Figure 4. Number of unclassified and detected taxa (species and genera) with microscopy and with the 16 different bioinformatics strategies. Taxonomic assignment of raw sequences resulted in the higher number of unclassified sequences. For the filtered ISU and the Furthest neighbor strategies less unclassified sequences are obtained when taxonomic assignment thresholds are lower. For the OptiClust strategy the number of unclassified varies greatly with no clear pattern.

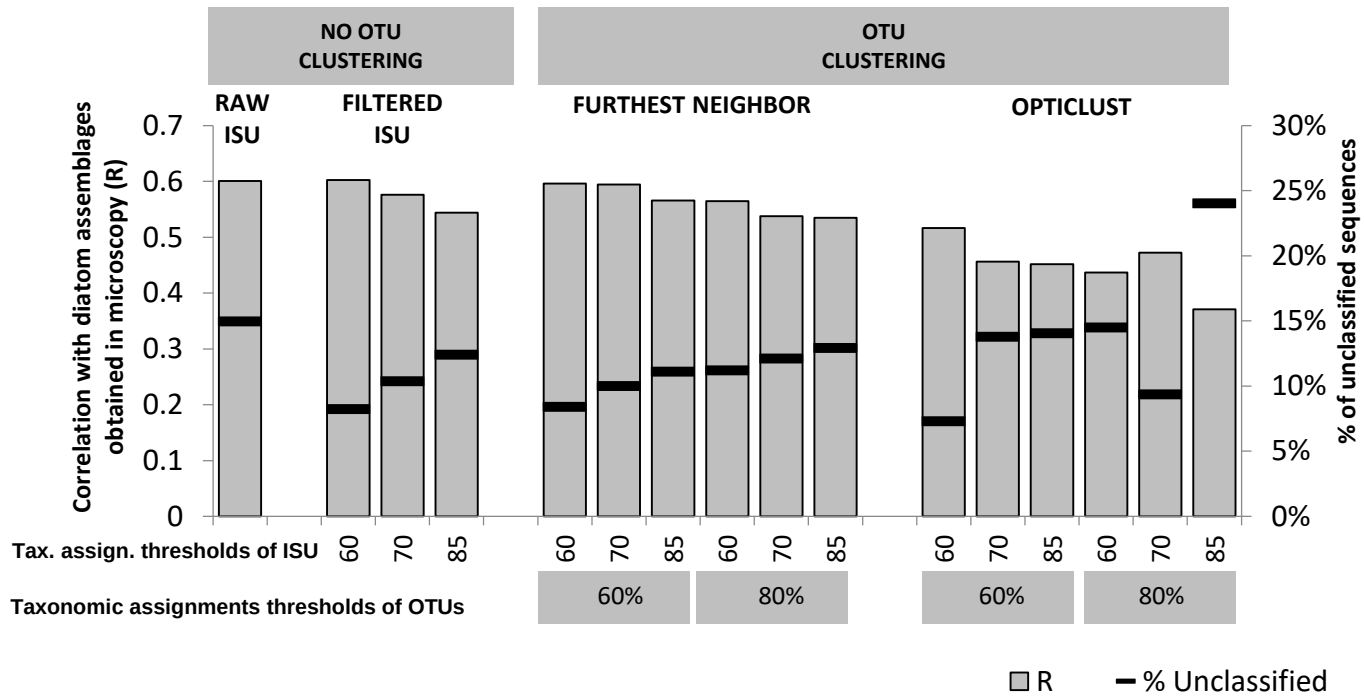


Figure 5. Comparison of diatom assemblages structures obtained with the 16 different bioinformatics strategies and microscopy. Diatoms assemblages for microscopy are expressed in relative abundances of frustules per species in each sample and for bioinformatics strategies they are expressed in relative abundances of sequences per species in each sample. R is the Pearson correlation coefficient calculated using a Mantel test, between microscopy and the bioinformatics strategy considered (Bray-Curtis distances). Note that OptiClust provided the weakest correlations with microscopy. ISU strategies give similar correlations than Furthest neighbor strategies.

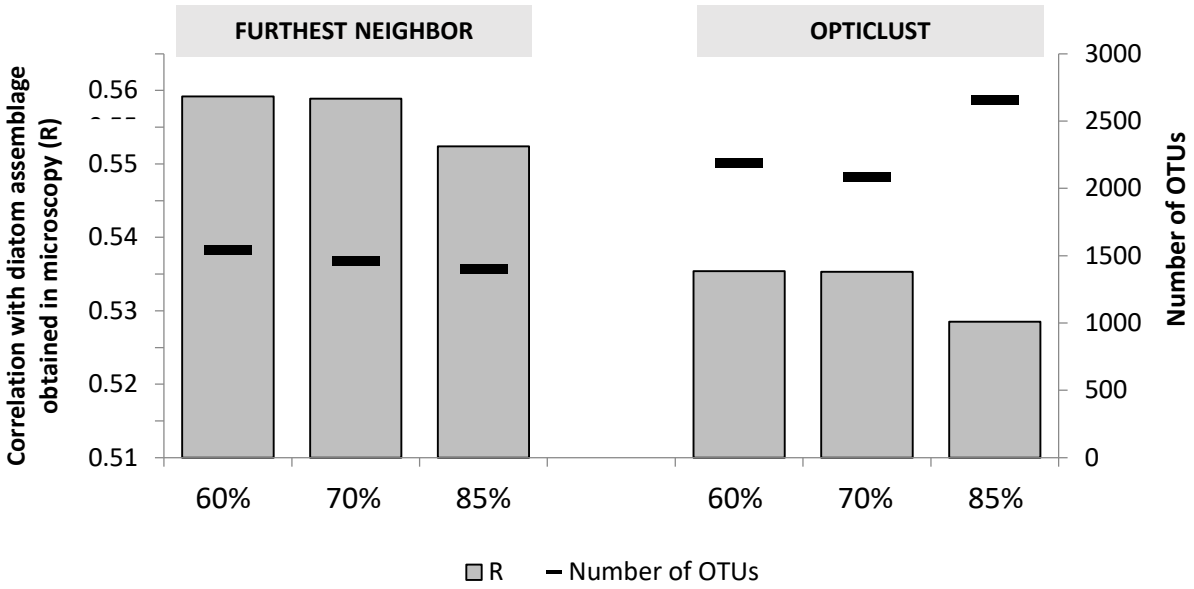


Figure 6 . Comparison of the diatom assemblages' structures obtained with the 6 different bioinformatics strategies based on OTUs clustering and microscopy. Diatom assemblages for OTUs clustering strategies are expressed in relative abundances of sequences per OTUs in each sample and for microscopy they are expressed in relative abundances of frustules per species in each sample. R is the Pearson correlation coefficient calculated using a Mantel test, between microscopy and the bioinformatics strategy considered (Bray-Curtis distances). The number of OTUs created with each pipeline is also indicated. Note that OptiClust provided the weakest correlation with microscopy and generated more OTUs than Furthest neighbor.

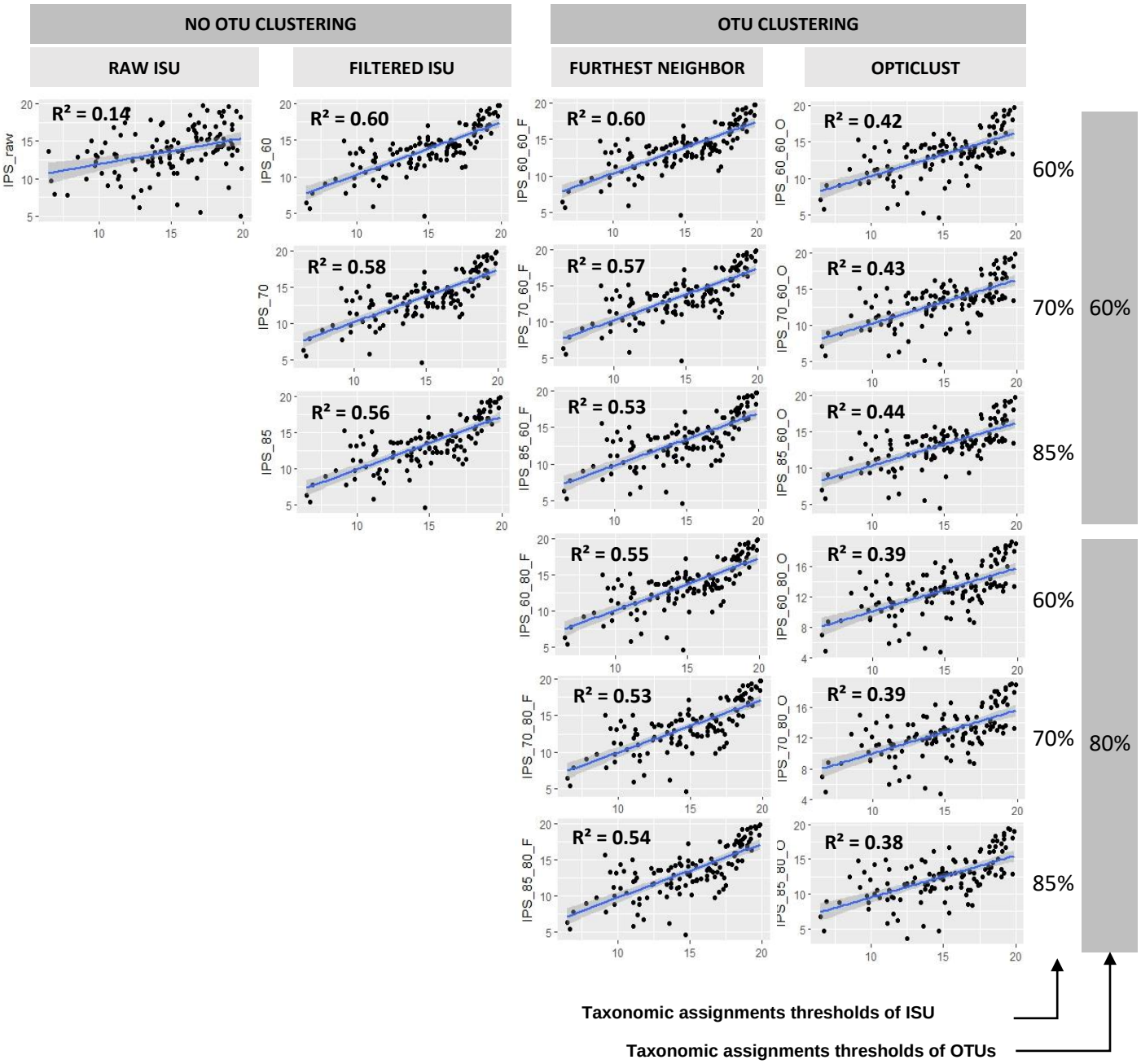
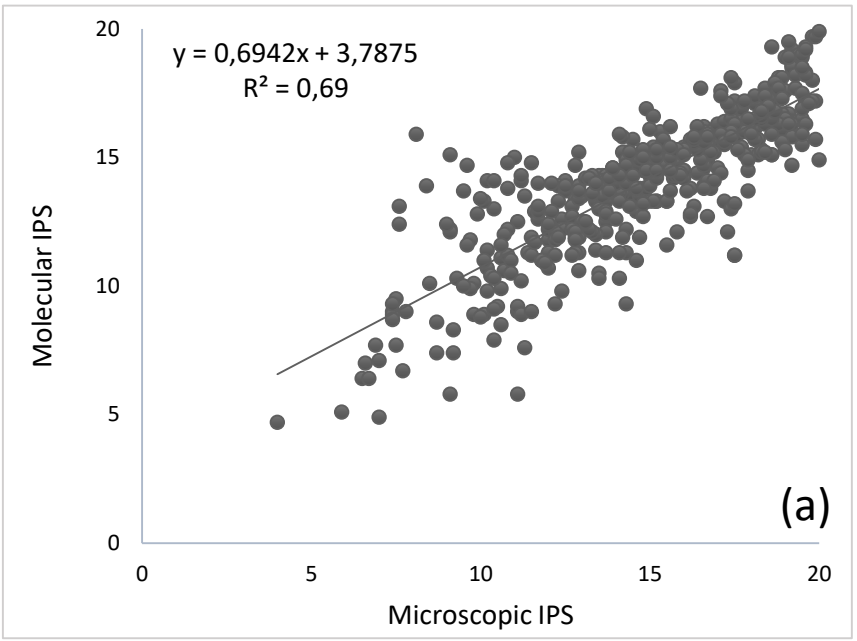


Figure 7 . Correlation between the water quality assessment obtained with microscopy (x axis) and the 16 different bioinformatic strategies (y axis). The biotic diatom index IPS (indice de Polluosensibilité Spécifique, Cemagref 1982) was calculated. IPS scores vary from 1 (bad quality status) to 20 (good quality status). IPS scores calculated from ISU with any quality filters (raw data) were poorly correlated to microscopy. Furthest neighbor and filtered ISU strategies provided similar results and were better correlated to microscopy than OptiClust.



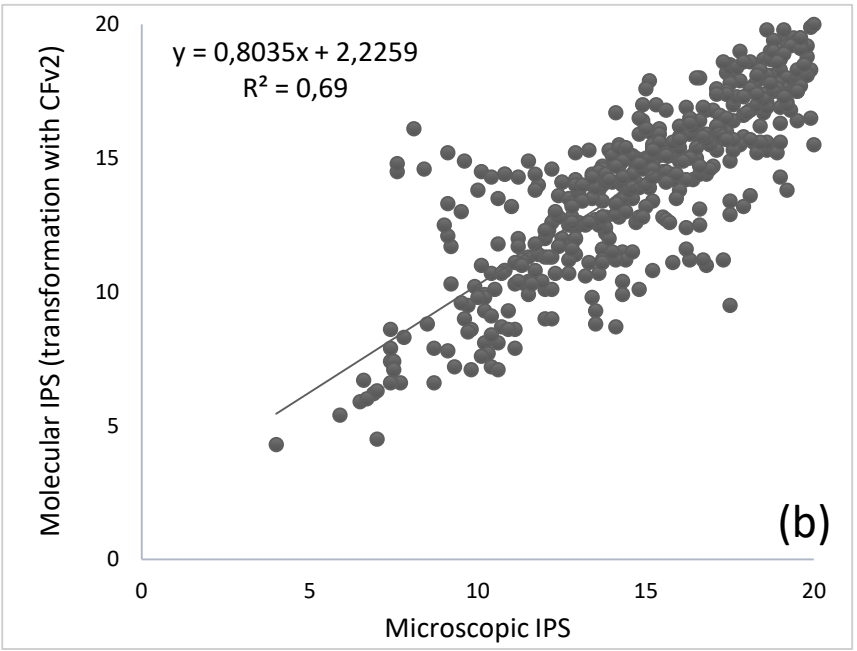


Figure 8. Correlation between morphological and molecular IPS scores. The best bioinformatics strategy was the one based on filtrated ISU with a taxonomic assignment of 60%. IPS scores vary from 1 (bad quality status) to 20 (good quality status). (a) Molecular IPS were calculated with untransformed data (b) Molecular IPS were calculated with data transformed with CFv2 (this transformation takes into account biovolume of species). Correlation within molecular and morphological IPS values from not transformed (a) and transformed data (b) was the same. However, the slope of the correlation is higher with transformed data (b).

Table 1. Summary table comparing taxonomic assignment thresholds of filtered sequences. Codes signification: "-": low, "~": intermediate, "+": high

	Taxonomic assignment threshold		
	60%	**70%**	**85%**
Number of unclassified sequences	**-**	**~**	**+**
Number of detected species	**+**	**~**	**-**
Correlation between diatom assemblages obtained in microscopy and in metabarcoding (relative abundances of diatom taxa)	**+**	**~**	**-**
Correlation between diatom assemblages obtained in microscopy and in metabarcoding (relative abundances of OTUs)	**+**	**~**	**-**
Correlation between IPS scores obtained in microscopy and metabarcoding	**+**	**~**	**-**

Table 2. Summary table comparing bioinformatics treatments. Codes signification: "-": low, "~": intermediate, "+": high, "++": very high, "n/a": not applicable. Calculation time is given as an indicative basis, calculation were carried out with a Dell Precision, Tower 7910 workstation (16 processors, 2.60 GHz, 64 Go RAM).

	Bioinformatics strategies			
	Raw ISU	**Filtered ISU**	**Furthest Neighbor**	**OptiClust**
Number of unclassified sequences	**+**	**-**	**-**	**~**
Number of detected species	**++**	**+**	**~**	**-**
Correlation between diatom assemblages obtained in metabarcoding (relative abundance of taxa) and microscopy	**+**	**~**	**~**	**-**
Correlation between diatom assemblages obtained in metabarcoding (relative abundance of OTUs) and microscopy	**n/a**	**n/a**	**+**	**+**
Correlation between microscopy and molecular IPS scores	**-**	**+**	**+**	**-**
Calculation time (computing hours)	**~3h30**	**~7h00**	**~ 19h00**	**~ 19h00**

Table 3: Confusion matrix comparing quality classes obtained with the diatom index IPS calculated from microscopy and from the best bioinformatics strategy (filtrated sequences, 60%). (a) Quality classes obtained with the best bioinformatics strategy when data are not transformed with species biovolumes, (b) quality classes obtained with the best bioinformatics strategy when data are transformed with species biovolumes using the correction factor CFv2. Quality classes boundaries: 1: bad quality [1; 5[, 2: poor quality [5; 9[, 3: moderate quality [9; 13[, 4: good quality [13; 17[, 5: high quality [17; 20].

(a)		**Quality classes obtained with the best bioinformatic strategy**				
		1	**2**	**3**	**4**	**5**
Quality classes obtained with microscopy	**1**	100.0	0.0	0.0	0.0	0.0
	2	4.5	54.5	27.3	13.6	0.0
	3	0.0	10.1	62.4	27.5	0.0
	4	0.0	0.0	16.8	82.6	0.5
	5	0.0	0.0	2.3	56.2	41.5

(b)		**Quality classes obtained with the best bioinformatic strategy and data were transformed with the biovolume correction factor CFv2**				
Quality classes obtained with microscopy		**1**	**2**	**3**	**4**	**5**
	1	100.0	0.0	0.0	0.0	0.0
	2	4.5	77.3	0.0	18.2	0.0
	3	0.0	14.7	59.6	25.7	0.0
	4	0.0	1.1	24.5	71.2	3.3
	5	0.0	0.0	2.3	35.4	62.3