Science of the Total Environment 745 (2020) 140948

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Diatom DNA metabarcoding for ecological assessment: Comparison among bioinformatics pipelines used in six European countries reveals the need for standardization

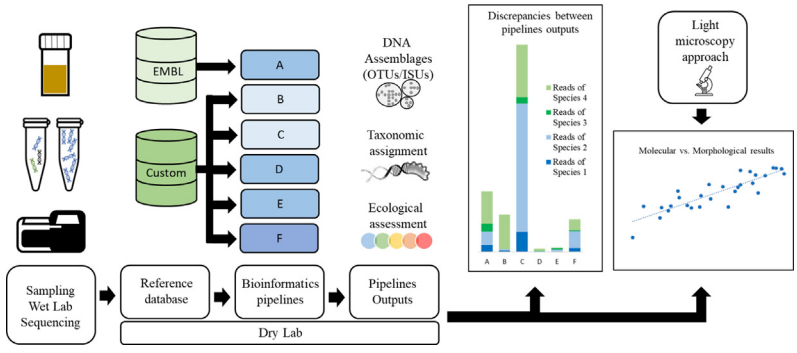
Bonnie Bailet [a,*], Laure Apothéloz-Perret-Gentil [b], Ana Baričević [c], Teofana Chonova [d,e], Alain Franc [f], Jean-Marc Frigerio [f], Martyn Kelly [g,h], Demetrio Mora [i], Martin Pfannkuchen [c], Sebastian Proft [i], Mathieu Ramon [j], Valentin Vasselon [k], Jonas Zimmermann [i], Maria Kahlert [a]

[a] Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, PO Box 7050, SE – 750 07 Uppsala, Sweden
[b] Department of Genetics and Evolution, University of Geneva, CH-1211 Geneva, Switzerland
[c] Center for Marine Research, Ruđer Bos'ković Institute, Rovinj, Croatia
[d] Research Department for Limnology, Mondsee, Faculty of Biology, University of Innsbruck, Mondsee, Austria
[e] CARRTEL, French National Institute for Agricultural Research (INRA), University of Savoie Mont Blanc, 75 bis avenue de Corzent, 74200 Thonon-les-Bains, France
[f] BioGeCo, French National Institute for Agricultural Research (INRA), 69 route d'Arcachon, 33610 Cesta, France
[g] Bowburn Consultancy, 11 Monteigne Drive, Bowburn, Durham DH6 5QB, UK
[h] School of Geography, University of Nottingham, Nottingham NG7 2RD, UK
[i] Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195 Berlin, Germany
[j] Fera Science Ltd, Sand Hutton, York YO41 1LZ, UK
[k] AFB, Pôle R&D "ECLA", INRA, UMR CARRTEL, 75bis av. de Corzent – CS 50511, FR-74200 Thonon-les-Bains, France

HIGHLIGHTS

- Bioinformatics pipelines used for diatom metabarcoding in six European countries are compared
- A common curated reference database does not guarantee detection of the same species assemblages
- Variation in filtering, clustering and taxonomic assignment drive discrepancies in the outputs
- Decisions made in bioinformatics analyses have an impact on environmental assessment
- Low reproducibility of outputs from the pipelines highlights the need for standardization.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 2 April 2020
Received in revised form 9 July 2020
Accepted 11 July 2020
Available online 18 July 2020

Editor: Sergi Sabater

ABSTRACT

Ecological assessment of lakes and rivers using benthic diatom assemblages currently requires considerable taxonomic expertise to identify species using light microscopy. This traditional approach is also time-consuming. Diatom metabarcoding is a promising alternative and there is increasing interest in using this approach for routine assessment. However, until now, analysis protocols for diatom metabarcoding have been developed and optimised by research groups working in isolation. The diversity of existing bioinformatics methods highlights the need for an assessment of the performance and comparability of results of different methods. The aim of this study was to test the correspondence of outputs from six bioinformatics pipelines currently in use for diatom

* Corresponding author.
E-mail addresses: bonnie.bailet@slu.se (B. Bailet), Laure.Perret-Gentil@unige.ch (L. Apothéloz-Perret-Gentil), ana.baricevic@cim.irb.hr (A. Baričević), teofana.chonova@inrae.fr (T. Chonova), alain.franc@inra.fr (A. Franc), Jean-Marc.Frigerio@inrae.fr (J.-M. Frigerio), MGkelly@bowburn-consultancy.co.uk (M. Kelly), d.mora@bgbm.org (D. Mora), pfannkuchen@cim.irb.hr (M. Pfannkuchen), Mathieu.Ramon@fera.co.uk (M. Ramon), j.zimmermann@bgbm.org (J. Zimmermann), maria.kahlert@slu.se (M. Kahlert).

https://doi.org/10.1016/j.scitotenv.2020.140948

Keywords:
Bacillariophyta
Biomonitoring
Metabarcoding
Morphological identification
*rbc*L
18S-V4

metabarcoding in different European countries. Raw sequence data from 29 biofilm samples were treated by each of the bioinformatics pipelines, five of them using the same curated reference database. The outputs of the pipelines were compared in terms of sequence unit assemblages, taxonomic assignment, biotic index score and ecological assessment outcomes. The three last components were also compared to outputs from traditional light microscopy, which is currently accepted for ecological assessment of phytobenthos, as required by the Water Framework Directive. We also tested the performance of the pipelines on the two DNA markers (*rbc*L and 18S-V4) that are currently used by the working groups participating in this study. The sequence unit assemblages produced by different pipelines showed significant differences in terms of assigned and unassigned read numbers and sequence unit numbers. When comparing the taxonomic assignments at genus and species level, correspondence of the taxonomic assemblages between pipelines was weak. Most discrepancies were linked to differential detection or quantification of taxa, despite the use of the same reference database. Subsequent calculation of biotic index scores also showed significant differences between approaches, which were reflected in the final ecological assessment. Use of the *rbc*L marker always resulted in better correlation among molecular datasets and also in results closer to these generated using traditional microscopy. This study shows that decisions made in pipeline design have implications for the dataset's structure and the taxonomic assemblage, which in turn may affect biotic index calculation and ecological assessment. There is a need to define best-practice bioinformatics parameters in order to ensure the best representation of diatom assemblages. Only the use of similar parameters will ensure the compatibility of data from different working groups. The future of diatom metabarcoding for ecological assessment may also lie in the development of new metrics using, for example, presence/absence instead of relative abundance data.

1. Introduction

Diatoms are valuable indicators of ecological status of water bodies (Smol and Stoermer, 2010) and are widely used as proxies for "phytobenthos", part of the "Macrophyte and Phytobenthos" biological quality element specified for ecological status assessment in the European Water Framework Directive (WFD: European Commission, 2000). The current approach to environmental quality assessment using diatoms is to identify and count diatom valves from a biofilm sample using light microscopy. The relative abundance of species found in the sample are then used to calculate a biotic index which is translated into ecological status classes (ranging from high to bad status) (Kelly et al., 2009; Kelly et al., 2014). This approach, however, is time-consuming and requires extensive taxonomic expertise. Variations in taxonomic expertise or concepts can lead to significant discrepancies in the taxa lists produced in different laboratories (Kahlert et al., 2008).

There is, as a result, considerable interest in alternative methods, with a strong focus on the development of DNA metabarcoding with data generated by High-Throughput Sequencing (HTS). The metabarcoding method could potentially provide a faster and cheaper way of identifying species at higher taxonomic resolution than is possible using the traditional method. Diatom metabarcoding has developed rapidly over the past decade (Kermarrec et al., 2013, Kermarrec et al., 2014, Visco et al., 2015, Zimmermann et al., 2015, Apothéloz-Perret-Gentil et al., 2017, Siegwald et al., 2017, Vasselon et al., 2017b, Bailet et al., 2019, Mortágua et al., 2019, Chonova et al., 2019, Rivera et al., 2020) and has the potential to be applied routinely for ecological assessment of streams at national scale (Kelly et al., 2018a, 2018b). However, as well as the diversity of existing methods and protocols within and among countries, there is still a lack of standardization and harmonization (Tapolczai et al., 2019b) hampering the establishment of a robust approach at the European scale to meet WFD requirements. Several methodological "hot-spots" such as reference databases (Weigand et al., 2019), quantification, stabilisation of technologies and protocols still require development. Over the years, developments such as evaluation of different barcode regions (Kermarrec et al., 2013) or different DNA extraction methods (Vasselon et al., 2017a), and also the extension of reference barcode databases (Rimet et al., 2018; Rimet et al., 2019) have allowed a consensus to develop, from which standards might be derived (CEN, 2018a; CEN, 2018b).

There is currently a great variety of bioinformatics pipelines available for diatom metabarcoding in Europe. Some use software packages such as Qiime (Caporaso et al., 2010) or Mothur (Schloss et al., 2009) which are wrappers around a family of different methods for both supervised and unsupervised clustering. Others are script-based processes such as MetBaN (Proft et al., 2017) and Diagno-syst (Frigerio et al., 2016). Until now, bioinformatics pipelines for diatom metabarcoding have been developed and optimised by groups working in relative isolation. There is a need to evaluate how these pipelines perform in comparison to one another, and if any have clear advantages over the others. Metabarcoding usually includes bioinformatics steps for the preprocessing of the sequences (demultiplexing of samples, paired-end fragment assemblage) and for cleaning the HTS data (quality filtering, chimera removal). At each stage, developers are free to select appropriate parameters such as filtering criteria or chimera removal algorithms. Furthermore, several strategies have been applied to link sequences to reference databases. These include clustering of similar sequences into OTUs (Operational Taxonomic Units) (Zimmermann et al., 2015; Vasselon et al., 2017b; Mora et al., 2019), or into ISUs (Individual Sequence Units) (Visco et al., 2015; Rivera et al., 2020), strict supervised clustering (Bailet et al., 2019) or taxonomy-free approaches (Apothéloz-Perret-Gentil et al., 2017). The diversity of bioinformatics analyses which results has been shown to have an impact on the taxonomic inventories produced from pipelines and in turn, on ecological assessment outcomes when using diatoms (Tapolczai et al., 2019a; Rivera et al., 2020), in turn, complicates inter-study comparisons.

In this study we provide a comparison of six bioinformatics procedures, implementing different pipelines currently proposed for ecological assessment with diatoms in six European countries (Croatia, France, Germany, Switzerland, Sweden and United Kingdom). Each pipeline was used to treat HTS data from 29 biofilm samples from lakes and rivers of Scandinavia, using a standardized reference database to minimize bias introduced from this source. We compared the sequence unit (OTUs and ISUs) assemblages produced, as well as the taxonomic assignment at genus and species level, the biotic index scores and the ecological status assessment. The main focus of our study was to compare the performance of different bioinformatics pipelines to each other. We also compared the performance of each bioinformatics pipeline to that of traditional light microscopy to illustrate possible consequences of pipeline differences when using them for ecological assessment. We compared two different DNA barcode markers commonly used for diatom metabarcoding: a fragment of the plastid gene *rbc*L and the V4 region of the nuclear-encoded 18S gene.

We specifically tested five hypotheses: i) different pipelines generate different assemblages of sequence units, ii) this, in turn, creates discrepancies in taxonomic assignment at genus and species level (in terms of presence/absence and relative abundances), iii) as a result of these taxonomic discrepancies, significantly different index scores are generated by different pipelines, iv) the differences in index scores lead to divergence of the ecological assessments between pipelines and methods; and, v) the bioinformatics analyses will provide higher taxa abundance and diversity when using the DNA marker they were optimised for. Finally, we consider the bioinformatics choices that are likely to be responsible for the observed discrepancies and the implications for environmental assessment.

2. Material and methods

2.1. Process outline

This study aimed to perform the entire bioinformatics analyses designed by each of the six research groups currently developing diatom metabarcoding in Europe on the same set of raw sequences, and to compare the outcomes. DNA amplification and sequencing of all samples was performed at the University of Geneva in order to exclude variation in these steps from the study. A single curated reference database was selected to avoid any bias from this source, and all except one pipeline used it for taxonomic assignment. The raw FASTQ files, tags sequences, primer sequences and the cleaned reference databases were sent to all six working groups to run through their pipelines. The bioinformatics steps included demultiplexing (removing tags bases and associating the sequences to their original sample), the creation of 'contigs' (merging the two pair-end sequences generated by MiSeq technology), cleaning the dataset (length, quality check etc.), sequence unit clustering, and taxonomic assignment. Finally, we calculated the Specific Pollution Sensitivity Index ("Indice de Polluosensibilité Spécifique" IPS: Cemagref, 1982) for all samples using the taxa lists created with each pipeline, from which an ecological status class could be derived. For simplicity, we used the Swedish method for ecological classification (Kahlert et al., 2007), as all samples originated from Northern countries, and the Swedish system recognises one water type and a single reference (i.e. expected) IPS value only. All results were compiled for data analysis at sequence unit level, assigned taxonomy level and ecological assessment level. Each analysis was performed in parallel on the datasets generated with the *rbc*L and with the 18S-V4 markers.

2.2. Sample selection

A dataset of 29 biofilm samples collected from rivers and lakes of Nordic countries (Sweden, Finland, and Norway) was used for this study (subset of the samples used in Bailet et al., 2019, Supplementary S8). The selected samples cover a broad ecological gradient, with total phosphorus ranging from 0.2 to 433 μg/l, and pH ranging from 4.6 to 8.6 (Supplementary S1). All samples were collected in autumn from submerged hard substrata following the European standard for diatom sampling (EN 13946:2014, CEN, 2014) and preserved with 97% ethanol (final concentration approximately 70%) to protect the DNA from degradation in long term storage (Stein et al., 2013).

2.3. Morphological analysis

Preparation, identification and counting for the diatom analyses using light microscopy were performed using European and Swedish standards (SS-EN 13946:2014 and SS-EN 14407:2014, SIS stöd, 2014a, 2014b, Jarlman et al., 2016). Briefly, samples were oxidized with hydrogen peroxide (H_2O_2) and mounted with Naphrax (Brunel Microscope Ltd) to permanently fix the sample material on glass slides. At least 400 valves per sample were identified using a light microscope (1000× magnification) using standard literature (Jarlman et al., 2016). To allow for a fair comparison between the molecular and the traditional method, we used a light microscopy assessment of the samples that only included taxa that were represented in our cleaned reference database. The assessment aims at species identification; however, diatom identification to species level is not always possible. Hence, in this study we use the term "taxa" which combines different taxonomic levels (for example genus and species levels). However, we also analyse assignment to genus or species level, respectively, and in this context we use these exact terms.

2.4. DNA extraction and sequencing

Pellets of biofilms were prepared by centrifuging (at 13,000 rpm for 30 min) between 2 and 4 ml of the initial biofilm suspension. DNA extraction was performed on the pellets using the Macherey Nagel NucleoSpin® Soil kit (MN-Soil) following the manufacturer's instructions (e.g. Vasselon et al., 2017a). Two regions were amplified with primers specific to diatoms. First, the hypervariable region V4 of the 18S rRNA gene was amplified using specific primers optimised for diatoms metabarcoding with Illumina sequencing by Visco et al. (2015). PCR amplifications were performed as described in Visco et al. (2015). The second marker used was a 312 bp fragment of the plastid *rbc*L gene using primers and PCR conditions according to Vasselon et al. (2017b), except that 2 μl of DNA extract was used and 35 PCR cycles were performed. For both markers, we used a unique combination of forward and reverse tagged primers with individual tags composed of 8 nucleotides attached at each primer's 5′-extremities (Esling et al., 2015). Several different forward and reverse tagged primers were designed to enable the multiplexing of all PCR products in a unique sequencing library for each marker. Three PCR replicates were performed for each sample and were then pooled and quantified with capillary electrophoresis using QIAxcel instrument (Qiagen). Equimolar concentrations of PCR products were pooled for each library and purified using the High Pure PCR Product Purification kit (Roche Applied Science). Library preparation was performed using the Illumina TruSeq® DNA PCR-Free Library Preparation Kit. The libraries were then quantified with qPCR using KAPA Library Quantification Kit and sequenced on a MiSeq instrument using paired-end sequencing for 500 cycles with Standard kit v2. Sequencing was performed at the University of Geneva on an Illumina MiSeq sequencer.

2.5. The different tested pipelines

2.5.1. Pipeline: Mothur ("ISUs script"); used in France (FR)

The pipeline was run entirely under Mothur version 1.43.0 software (Schloss et al., 2009). The first step with MiSeq data was to check the integrity of the forward and reverse primers, then to assemble the paired-end reads (make.contigs), to demultiplex the samples and remove the primer sequences (make.contigs) as well as removing reads with more than one mismatch in the primer sequence (make.contigs). A filtering step (screen.seqs) excluded any reads with an overlapping region shorter than 100 bp between the forward and reverse fragments, below a set read length range (263 + − 10 bp for the *rbc*L marker and 286 + − 10 bp for the 18S-V4 marker), with a Phred quality score below 23 over a moving window of 25 bp, or with any ambiguous base ("N") and homopolymer over 8 bp. The dereplication step (unique.seqs) kept unique sequences and their read abundance in the sample. Sequences encountered only once across the dataset (singletons) were removed (split.abund). Then an alignment of the reads was performed (align.seqs, default algorithm: Needleman-Wunsch) and poorly aligned sequences were removed (screen.seqs) according to an optimised start and end location and a set length range (as previously, 263 + − 10 bp for the *rbc*L marker and 286 + − 10 bp for the 18S-V4 marker). The sequences were trimmed to the preferred alignment length and a second dereplication step was applied to the aligned, trimmed and filtered sequences. Potential sequencing errors were

removed (pre.cluster) based on the sequence abundances (rare sequences that are within 1 mismatch from the most abundant sequences were removed) and potential chimeras were detected with the VSEARCH algorithm (chimera.vsearch) and removed (remove.seqs). At this point, all Individual Sequence Units (ISUs) with an abundance of only 1 read (or "singleton") were again removed (split.abund, cutoff = 1) and a taxonomic assignment was done using a naïve Bayesian method (the Wang method: Wang et al., 2007) (align.seqs, cutoff = 60%). DNA reads unassigned to diatom classes (Fragilariophyceae, Bacillariophyceae, Mediophyceae, Coscinodiscophyceae) were removed (get.lineage).

2.5.2. Pipeline: Mothur ("strict alignment" script); used in Croatia (HR)

This pipeline was run entirely in the Mothur 1.42.0 software (Schloss et al., 2009) and mainly followed the process of the previous Mothur script with a few modifications (Fig. 1). First, the demultiplexing and primer removal were combined with the assembling of the paired-end reads with default parameters and trimming the sequences to only the overlapping section (make.contigs). The filtering step (screen.seqs) excluded any reads with an overlap ≤172 bp (minoverlap = 172) between the F and R fragment for the *rbc*L marker and an overlap ≤150 bp (minoverlap = 150) for the 18S-V4 marker. The dereplication step was the same (unique.seqs), but there was no removal of singletons. An alignment of the reads was performed (align.seqs, default algorithm: Needleman-Wunsch) and poorly aligned sequences were removed (screen.seqs) applying the reference database similarity parameter (minsim = 100). Sequences were also filtered according to an optimised start and end position (*rbc*L marker: start = 71, end = 243 and 18S marker: optimize = start-end-minlength). As for the previous pipeline, the dereplication step was performed again in case some identical sequences appeared after screening the aligned sequences (unique.seqs). Potential chimeras were detected with the uchime algorithm (chimera.uchime) and removed (remove.seqs). The taxonomic assignment (classify.seqs) was done using the Wang method (Wang et al., 2007) on the unique sequences (ISUs) with cutoff = 0 to return a full taxonomy for every sequence, regardless of the bootstrap value for that taxonomic assignment. In contrast to the previously described Mothur script, OTU clustering, consensus taxonomy assignment or

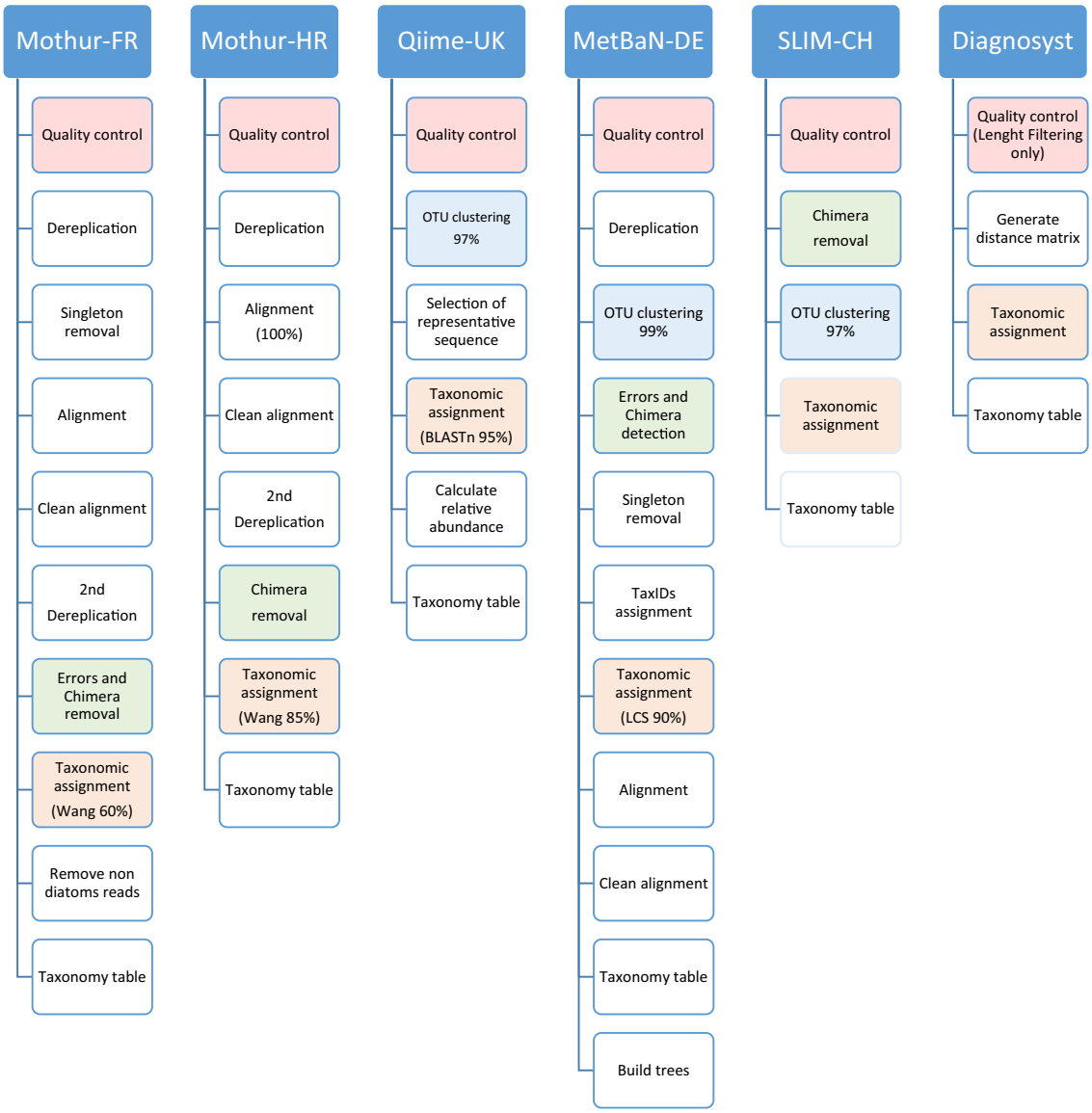


Fig. 1. Overview of the bioinformatics pipelines tested. A detailed description of each bioinformatics pipeline is given in Section 2.5. The pipelines' name combined with the country implementing the script in this study are used to refer to each bioinformatics analysis from hereon (ex: Mothur pipeline used by the working group in France: Mothur-FR). Country codes: France (FR), Croatia (HR), United Kingdom (UK), Germany (DE), Switzerland (CH) and Sweden (SE).

sample normalization to minimum read number were not performed in this pipeline.

2.5.3. Pipeline: Qiime; used in the United Kingdom (UK)

This pipeline is currently used in the United Kingdom for ecological status assessment based on diatoms assemblages and is separated into two parts: quality control and taxonomic assignment, as described in Kelly et al. (2018a). This official pipeline does not include a demultiplexing step because the sequencing platform (MiSeq) usually performs this during the routine analysis, therefore, a step of demultiplexing has been added specifically for this study (Fig. 1). This extra step, absent from Kelly et al. (2018a), uses the make.contig function of Mothur with an oligo file (linking barcodes and samples ID). The quality control stage included four steps: removal of PCR primers from both strands using Cutadapt v1.9.1 (Martin, 2011), trimming poor quality 3′ ends of sequences from both strands using Sickle v1.33 (Joshi and Fass, 2011), assembling paired-end reads using PEAR v0.9.6 (Zhang et al., 2013) and removal of any sequences with a quality score lower than 30 and shorter than 250 bp using Sickle v1.33. The taxonomic assignment part was mainly done with the Qiime platform (www.qiime.org) and included four steps: OTU de novo clustering at 97% similarity with UCLUST (Qiime) (Edgar, 2010), selection of the most abundant sequence as representative sequence for each OTU (Qiime), taxonomic assignment of each representative sequence using BLASTn (Qiime) with 95% of sequence identity threshold and calculation of relative abundance for each taxa present in the sample (without minimum abundance threshold value). Any sample with <3000 reads was excluded from the final dataset. In this case, the MR8 sample was removed. This pipeline has been designed to optimize the use of sequences generated on the *rbc*L marker and is currently limited to this DNA marker; thus, no output data on the 18S-V4 marker was generated.

2.5.4. Pipeline: MetBaN; used in Germany (DE)

The MetBaN pipeline (Proft et al., 2017) uses a phylogeny-based approach for taxa delimitation by combining phylogenetic and metabarcoding packages. It is split into three separate scripts i) downloading the EMBL (European Molecular Biology Laboratory) reference database ii) performing *in-silico* PCR to create a sub-sampled reference database and iii) the core analysis. The current version of MetBaN cannot run the taxonomic assignment with an external reference database and thus was the only pipeline in this study that did not use the custom diat.barcode reference database. The MetBaN core script included pre-processing of sequences (demultiplexing, assembling paired-end reads, cleaning, and taxonomic assignment), alignment of sequences and also construction of a phylogenetic tree to check and improve the accuracy of the taxonomic assignment (Fig. 1). The pre-processing steps were conducted using tools from the OBITools package. Assembly of paired-end reads was done using the Illuminapairedend command (with a minimum alignment score of 40) after which the obigrep command removes any unpaired reads. Demultiplexing was performed using the ngsfilter script and dereplication with the obiuniq script, after which reads were filtered to keep only the reads with a length of >150 bp. OTU clustering and cleaning of potential chimeras and errors was executed using the obiclean command with a threshold ratio of 0.05 and an allowed difference between sequences of one bp. Finally, the ecotag command was used to perform the taxonomic assignment ("Longest Common Subsequence" algorithm) based on the EMBL reference database, with a minimum identity of 90%. After that, the script created a phylogenetic tree for each taxonomic ID ("taxid") of interest in the dataset (e. g., the taxid of Bacillariophyta is 2836). Sequences were aligned using MAFFT (Katoh and Standley, 2013) and T-Coffee (Di Tommaso et al., 2011), and Maximum Likelihood trees were built using RAxML (Stamatakis, 2014) (with 10,000 bootstraps runs). The tree is used to manually refine the taxonomic assessment precision, to take into account the phylogeny in case of taxonomic conflict during the assignment and for sequences without taxonomic assignment. MetBaN can also produce Krona plots to visualise the distribution of found species.

2.5.5. Pipeline: Diagno-syst; used in Sweden (SE)

Diagno-syst (Frigerio et al., 2016) is a French stand-alone program implementing a supervised clustering algorithm for taxonomic assignment. This pipeline has been designed to use the full capacity of High-Performance Computing and hence avoids any heuristics classically developed to avoid memory and time limits in computing. It keeps the pre-processing of sequences to a minimum, favouring a straight taxonomic assignment algorithm instead. The pipeline is implemented in two different programs: MPI-disseq (written in C and MPI) and Diagno-syst (written in Python). It starts with demultiplexed FASTQ files of single-end reads, with the tag sequences removed. Since MiSeq data was used in this study, the assembly of paired-end reads was made using Mothur (Schloss et al., 2009) following the same steps as the Mothur ("ISUs script") described above. Pre-processing of the HTS data was limited to filtering the sequences by their length, to keep reads between 300 and 315 bp with *rbc*L marker and between 320 and 340 bp with the 18S-V4 marker. Demultiplexing, primers and tags removal are performed with Cutadapt (Martin, 2011). The first step of taxonomic assignment was to calculate all pair-wise distances between sample sequences ("queries") and reference database sequences ("references") using the MPI-disseq program in C (parallel implementation of a Smith-Waterman algorithm with Message Passing Interface MPI). This was done at the French National Computing Center (IDRIS) on a Blue Gene Q hyper-parallel machine (1024 cores). Taxonomic assignment was then completed using the Diagno-syst Python program which listed, for each query, all the reference barcodes that were at a distance lower than a given barcoding distance (defined in terms of bp differences). The taxonomic name was assigned to the read only if all references at the same distance (or lower) had the same name in the reference database. The pipeline also implemented a sliding barcoding gap parameter: this aspect provided, in one run, taxonomic assignments at different bp distances. In this study, we chose a range between 4 and 10 bp gaps to encompass the 97% identity usually used for species level identification. Any assignment with >10 bp differences was not kept for further analysis, which should also exclude any potential chimera or sequencing errors.

2.5.6. Pipeline: SLIM, used in Switzerland (CH)

The SLIM pipeline is a web application (Dufresne et al., 2019) with a graphic user interface and implements several modules from well-established software sources. All the cleaning steps in this study were done using the VSEARCH algorithms (Rognes et al., 2016). First, raw FASTQ reads were filtered by removing any sequence with a mean quality score of 30 or higher, with ambiguous bases or any mismatch in the primer sequences. Then, paired-end reads were assembled using the PEAR algorithm implemented in PANDAseq (Masella et al., 2012). Chimera detection and removal was performed using the –uchime_denovo algorithm and OTU clustering at 97% similarity was performed using the –cluster_size algorithm with default parameters (distance based greedy clustering). The approach to ecological assessment developed by Apothéloz-Perret-Gentil et al. (2017) is based on OTU structure and not on taxonomic assignment; thus, the pipeline does not typically include that step. However, for this study, the taxonomic assignment was performed using the assignment function of VSEARCH tool (–usearch_global), and OTUs sharing the same assignment were then grouped.

For simplicity, we refer to each bioinformatics pipeline from here on using an abbreviation with the software's name and the country of the working group using the pipeline's script in this study (e.g.: the Diagno-syst pipeline with the script used in Sweden will be referred to as "Diagnosyst-SE") (Fig. 1).

2.6. Common cleaned database

To the best of our knowledge, we used the most comprehensive curated reference database for diatoms available, diat.barcode, for both the *rbc*L and 18S-V4 marker (Rimet et al., 2019). In order to compare the performances of DNA markers, we chose to keep all the barcodes for species that were represented in both reference databases. This choice, while limiting the species detection potential, allowed for objective interpretation of markers without the bias of using different reference databases. Our earlier studies had shown that one of the main causes of different performance of markers was the incompleteness of reference databases (Bailet et al., 2019). The comparison of two DNA markers can only be fair if the databases used with each marker and pipeline are equal, and this is not the case for those currently available. The final cleaned databases contained 237 species names (509 sequences in the *rbc*L database and 412 sequences in the 18S-V4 database) and is available in the Zenodo repository (https://doi.org/10.5281/zenodo.3885810).

All the tested pipelines, except MetBaN-DE, performed the taxonomic assignment with this cleaned reference database. The LCS taxonomic assignment (ecotag function) with MetBaN currently runs with the EMBL reference database only.

2.7. Index calculations and quality classes

IPS values were calculated from all morphological and molecular taxa lists using the indicator value and sensitivity values from the OMNIDIA software (version 6.0.6) (Lecointe et al., 1993). We then used the IPS scores to infer an ecological status class. In order to compare the impact that taxa lists created by different pipelines might have on estimates of ecological status directly, irrespective of differences in ecological class boundaries between countries, we used the Swedish class boundaries for all datasets (Kahlert et al., 2007).

2.8. Data analyses

2.8.1. Comparison of DNA assemblage's structures

To detect any significant differences in the pipelines' outcomes, we first looked at the diversity of sequence units (OTUs and ISUs) produced by each pipeline, namely the number of sequence units produced and the range of reads that were clustered together. We also looked at the number of singletons and the number of unclassified reads present in each pipeline's outcome. Then, to assess the differences in sequence unit structure, we computed the Bray-Curtis distance matrix of each pipelines' sequence units' output using the vegdist function (vegan package, version 2.4.2) and ran a Procrustes analysis between each pair of pipeline outputs using the protest function (vegan package, version 2.4.2) in R (version 3.6.0) (R Core Team, 2014).

2.8.2. Comparison of taxonomic assemblages

In order to compare the performance of the different bioinformatics pipelines at taxonomic assignment, we removed all taxa not included in our custom reference database from the MetBaN-DE molecular inventories (run with the EMBL database). We looked at the taxonomic assignment produced by each pipeline in terms of presence/absence and relative abundance of taxa. The taxa lists obtained by taxonomic assignment of molecular data were also compared to the taxa lists obtained by light microscopy. We first compared the proportions of the different genera found in each dataset with an NMDS analysis using the isomds function (vegan package, version 2.4.2) and with a Multivariate ANOVA with the adonis function (permutations = 9999, method = "bray"; vegan package, version 2.4.2). With taxonomic assignment at the species level (transformed into relative abundances), we generated Euclidean distance dissimilarity matrices using the function vegdist (vegan package, version 2.4.2) between each molecular inventory. A Mantel test allowed us to assess the correlation between matrices. We also ran a SIMPER analysis (Clarke, 1993) using the PAST software (version 2.15b) (Hammer et al., 2001) and an Indicator Species Analysis (IndVal) (Dufrene and Legendre, 1997) using PC-ORD (version 6.08) (McCune and Mefford, 2011) to assess which taxa were driving the observed differences. The SIMPER analysis calculates the contribution of each taxon (%) to the dissimilarity between two groups of samples based on a Bray-Curtis dissimilarity matrix. The Indicator Species Analysis is used to identify which taxa are characteristic of particular groups of samples. The analysis produces an "indicator value" (IndVal), ranging from 0 to 100, based on the species' relative abundance in a sample and its relative frequency of occurrence in the different groups of samples. The analysis also calculates the statistical significance (at 0.05) of the IndVal to identify those taxa whose distribution is responsible for differences between groups of samples within a dataset. We calculated IndVal for each diatom taxon and considered that taxa with an IndVal of 50 and higher were strong indicators. Finally, we built a heatmap of the read numbers of each taxa (75 dominant species only) in all the pipelines using the ComplexHeatmap package (version 2.2.0) in R. We grouped taxa according to their phylogeny in order to visualise which taxonomic groups have good or weak resolution in metabarcoding compared to that found by light microscopy.

2.8.3. Comparison of ecological assessment outcomes

In order to assess the impact on the environmental assessment, we first looked at the correlations between IPS index scores calculated for each pipeline, and also to the IPS scores calculated from the light microscopy data. Then, we compared the ecological status classes ("high", "good", "moderate", "poor", "bad") derived from the IPS scores of the molecular and microscopy datasets: the five different classes were transformed into factorial data (i.e. "bad" = 1 and high = "5") and we used a Spearman's rank correlation test to compare the datasets (function CORREL in Microsoft Office Excel). We also looked at the number of HTS samples that either overestimated or underestimated ecological status when compared to classes derived from light microscopy.

2.8.4. Comparison of DNA markers

Finally, in order to assess the influence of the chosen DNA marker (*rbc*L and 18S-V4) on the performance of the pipeline, we examined the number of taxa shared by the different datasets. We also compared the r-squared values between IPS scores computed from metabarcoding and microscopy data.

3. Results

After sequencing, 7,386,592 DNA sequences were obtained from the *rbc*L marker and 6,279,138 DNA sequences from the 18S-V4 marker. Both runs were of good quality and could be used for further bioinformatics analysis. 284 taxa names were identified from HTS across all molecular datasets. 7173 valves were counted using light microscopy and 320 taxa were identified from the 29 biofilm samples. In order to make a fair comparison between microscopy and molecular datasets, only taxa represented in the reference database were kept and the final microscopy taxonomy included 53 taxa (14 of which were not identified using the molecular methods even with barcodes in the reference database). The taxa assemblages of the partial and full microscopy datasets remained similar to one another in terms of biotic index calculations (Supplementary S2).

3.1. Assemblages structure of the DNA datasets

The number of sequence units in the molecular datasets ranged from 159 to 542,871 in the *rbc*L datasets and from 94 to 24,707 in the 18S-V4 datasets. The lowest number of sequence units was produced with the Mothur-HR pipeline, while the Qiime-UK pipeline produced the highest number.

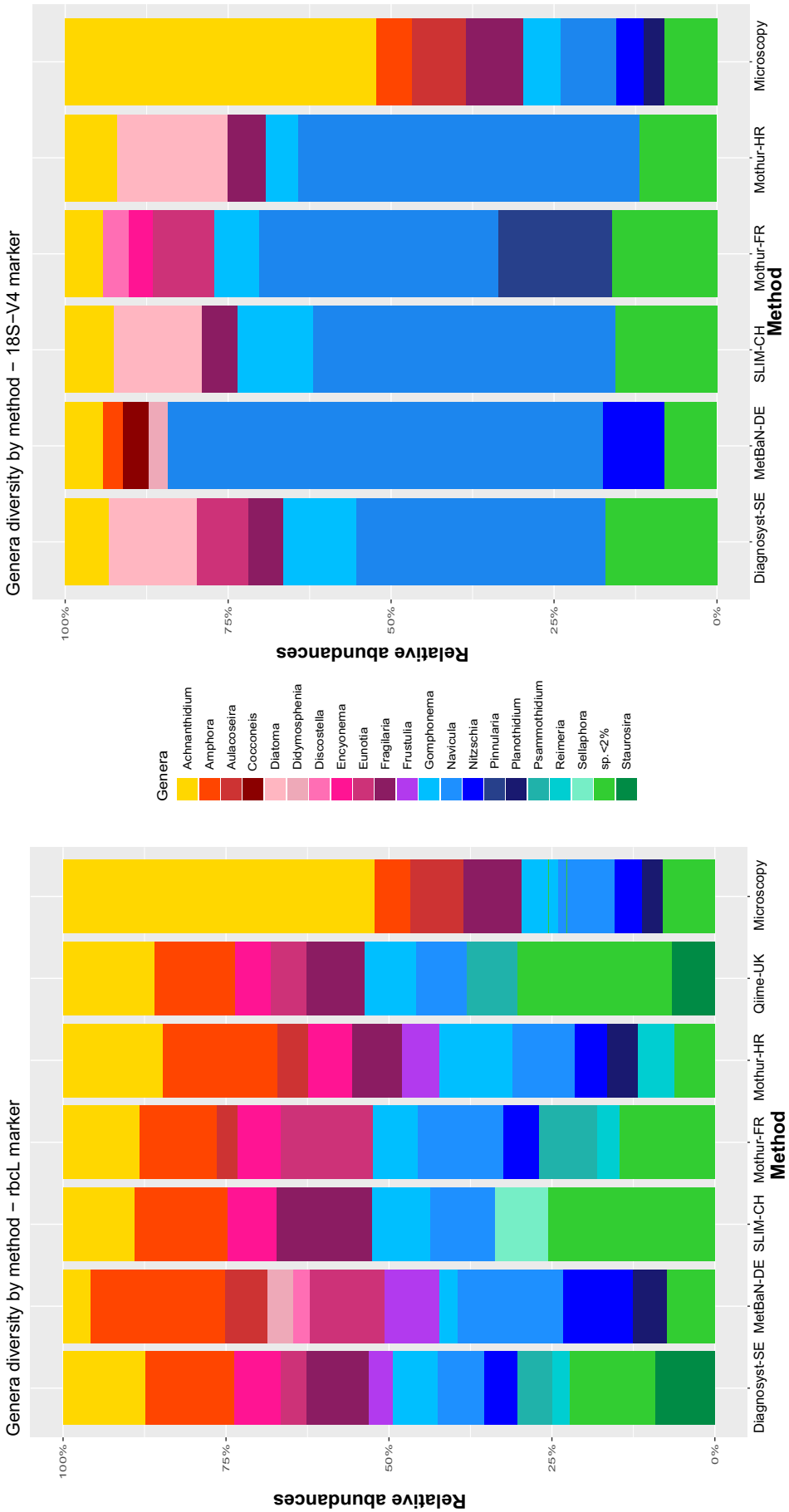


Fig. 2. Generic composition in the molecular and microscopy datasets. Only names for genera with relative abundances over 2% are given. Non dominant genera are represented under "genus <2%". Proportion of unclassified reads is not given.

Table 1
Summary of the output from the 6 bioinformatics pipelines in terms of DNA reads per sample (after filtering), number of OTUs or ISUs across the dataset and number of unassigned reads. The unassigned reads are all the sequences that could not be identified at genus level or below. The Diagnosyst-SE pipeline does not perform clustering and does not give information on DNA reads per sample or unassigned reads number. The Qiime-UK pipeline was run with the rbcL dataset only.

	*rbc*L			18S-V4		
	Reads per sample	Total of OTUs/ISUs	Unassigned reads	Reads per sample	Total of OTUs/ISUs	Unassigned reads
Mothur-FR	5727–343,850	13,321	953,823	2500–290,965	24,707	569,599
Mothur-HR	651–67,548	159	82,964	2–60,878	94	17,944
MetBaN-DE	5680–364,869	13,703	648,318	5931–281,368	20,716	550,719
Qiime-UK	2794–231,394	542,871	1,436,057	NA	NA	NA
SLIM-CH	6142–478,302	11,345	1,203,521	8156–394,475	18,422	1,194,151
Diagnosyst-SE	NA	NA	NA	NA	NA	NA

A Procrustes analysis on the molecular datasets showed that for *rbc*L, the sequence unit assemblages of the Mothur-FR pipeline, the Mothur-HR pipeline and the MetBaN-DE pipeline were well correlated with one another (correlation coefficients of 0.77, 0.85, 0.82 respectively, $p < 0.05$), but were not correlated with the SLIM-CH pipeline or the Qiime-UK pipeline's assemblages (0.40 and 0.20 respectively when compared with the Mothur-FR pipeline, 0.35 and 0.16 respectively when compared with the Mothur-HR pipeline and 0.34 and 0.17 respectively when compared with the MetBaN-DE pipeline). The sequence unit assemblages from the Qiime-UK pipeline and the SLIM-CH pipeline were also not correlated to one another (correlation coefficient 0.21). When using the 18S-V4 marker, a similar pattern appeared with the assemblages from the Mothur-FR, Mothur-HR and MetBaN-DE pipelines significantly correlated with each other (correlation coefficient 0.60, 0.79 and 0.74 respectively) but not to the assemblages produced by the SLIM-CH pipeline (correlation coefficient 0.35, 0.18 and 0.35 respectively).

3.2. Comparison of taxonomic inventories

3.2.1. Taxonomic assignments at genus level

The proportions of the dominant genera identified were quite similar among the pipelines (Fig. 2) when using *rbc*L. However, when using the 18S-V4 marker, MetBaN-DE and Mothur-FR pipelines found a few dominant genera not detected by any other pipelines (*Cocconeis* sp., *Discotella* sp. and *Encyonema* sp.). The proportion of the dominant genera identified by the bioinformatics pipelines was always significantly different from the proportions of dominant genera found by microscopy (Multivariate ANOVA p-value <0.05, Supplementary S4). These differences were mainly driven by the proportions of the genera *Achnanthidium* and *Navicula*.

3.2.2. Taxonomic assignments at the species level

The different bioinformatics pipelines identified different numbers of species. The Mothur-HR pipeline identified fewer species than any other pipeline using both DNA markers (84 species when using the *rbc*L marker and 66 species when using the 18S-V4 marker). The Qiime-UK pipeline identified the highest number of species for *rbc*L (174 species) and the Diagnosyst-SE pipeline identified the highest number of species for 18S-V4 (144 species). Details about the number of species found by each pipeline can be found in the Supplementary S3 (Venn Diagrams).

The proportion of unassigned sequences varied among the bioinformatics pipelines. The lowest amount of unassigned sequences was always obtained with the Mothur-HR pipeline and the highest number with the Qiime-UK pipeline, closely followed by the SLIM-CH pipeline (Table 1). It is not always possible to distinguish between sequences "identified as diatoms (Bacillariophyta) but above genus level" (e. g. identified at family, class, order etc.) and sequences "not identified as diatoms (Bacillariophyta)" (e. g. green algae). Some pipelines (e.g. Mothur), give the taxonomic assignment to the lowest possible level. As an example, we looked into the taxonomic diversity of unclassified reads provided by the Mothur-HR pipeline output (Supplementary S5). With the 18S-V4 marker, the majority of unassigned reads belonged to the Naviculales order. However, with the *rbc*L marker, similar proportions of unassigned reads belonged to the Naviculales and Cymbellales orders as well as to the families Fragilariaceae and Achnanthidiaceae (Supplementary S5).

A Mantel test on the diatom species' assemblages showed that they varied significantly according to the bioinformatics pipeline used (Table 2). The most similar taxonomic assemblages were observed between the SLIM-CH pipeline and the Mothur-FR pipeline when using the 18S-V4 marker. The correlation was weak but significant (Table 2, $r = 0.28$). When using *rbc*L, the highest correlation was found between the MetBaN-DE species assemblage and the Mothur-HR species assemblage (Table 2, $r = 0.59$). The species assemblages obtained by the molecular methods were also always significantly different from the microscopy species assemblage (Table 2). The highest correlation was obtained for the Diagnosyst-SE pipeline with *rbc*L ($r = 0.53$). The highest correlation when using 18S-V4 was obtained with the SLIM-CH pipeline; however, the correlation coefficient was weak (Table 2, $r = 0.24$). The SIMPER analysis on the species data showed that the three most frequently (relative abundance of the whole inventory) found species (cf. Table 3: NTPT, APED and ADMI) contributed to >20% of the overall Bray-Curtis dissimilarity between all groups, i.e. all molecular inventories for both markers (*rbc*L and 18S-V4). Together with six additional species (NRAD, GPAR, FCRS, NCTE, DITE, ELSE, FSBC, PGIB), they accounted for half of the observed dissimilarity between pipelines.

Table 2
Correlation coefficients of Mantel tests between taxa lists obtained by the bioinformatics pipelines and with the taxa lists obtained with microscopy. Statistical significance ($p < 0.05$) is indicated in bold.

		Mothur-FR	Mothur-HR	MetBaN-DE	SLIM-CH	Diagnosyst-SE	Qiime-UK	Microscopy	
18S-V4 marker	Mothur-FR		0.11	0.05	0.06	**0.35**	0.02	**0.41**	*rbc*L marker
	Mothur-HR	0.08		**0.59**	**0.11**	0.14	0.03	**0.42**	
	MetBaN-DE	**0.17**	**0.14**		**0.15**	0.00	−0.02	0.16	
	SLIM-CH	**0.28**	0.04	0.01		0.04	0.01	0.18	
	Diagnosyst-SE	0.01	0.02	0.01	0.02		0.03	**0.53**	
	Qiime-UK							0.09	
	Microscopy	**0.18**	**0.19**	**0.15**	**0.24**	**0.17**			

Table 3
The ten diatom species contributing to more than half of the observed dissimilarity between the taxonomic assemblages according to the SIMPER analysis. The individual contribution and cumulative contribution are given. Species indicated by a star were detected by all six bioinformatics pipelines.

Species	Omnidia code	Contribution to dissimilarities (%)	Cumulative contribution to dissimilarities (%)
Navicula tripunctata	NTPT	9.338	9.338
* *Amphora pediculus*	* APED	7.587	16.92
* *Achnanthidium minutissimum*	* ADMI	7.33	24.25
* *Navicula radiosa*	* NRAD	5.803	30.06
* *Frustulia crassinervia*	* FCRS	5.091	35.15
* *Gomphonema parvulum*	* GPAR	4.877	40.03
* *Navicula cryptotenella*	* NCTE	3.813	43.84
* *Diatoma tenuis*	* DITE	3.026	46.87
* *Encyonema silesiacum*	* ESLE	2.568	49.43
Pinnularia gibba	PGIB	2.523	51.96

Species found in high read abundances were most important in driving the Bray-Curtis differences (Table 3). Out of these 10 species, 8 were found in the outputs from all the pipelines, highlighting relative abundance discrepancies as the origin of observed differences. However, two species (NTPT and PGIB) were not identified by all the pipelines, and in their case, presence/absence discrepancies combined with relative abundance discrepancies lie behind the observed differences.

The IndVal analysis showed that species other than the ones highlighted in the SIMPER analysis were characteristic of the different pipelines and marker datasets. In general, molecular inventories produced with *rbc*L had more species with significant indicator values (IV > 50), showing the distinctiveness of the assemblages produced by different pipelines, compared with those produced using 18S-V4 (Fig. 3). The species list produced using the Qiime-UK pipeline had the highest number of "strong indicator taxa" (Fig. 3). The species list produced from the Diagnosyst-SE pipeline had the most distinct assemblage when using 18S-V4 and the second most distinct when using *rbc*L. Species inventories from the Mothur-HR and MetBaN-DE pipelines did not include any "strong indicator taxa".

To illustrate differences in taxon annotation among pipelines, we selected two genera, *Fragilaria* and *Eunotia*, and show the distribution of the different species names and number of reads allocated to them. The number of reads allocated to the different species names differed considerably among pipelines (Fig. 4a and b). Some species were not detected at all in some cases. For example, *E. implicata* was detected in much higher proportion with the 18S-V4 marker. The reverse situation occurs for *E. glacialis* and *E. pectinalis*, which were detected in much higher proportions with *rbc*L. The Diagnosyt-SE and Mothur-FR pipelines also detected higher proportions of *Eunotia* species than any other pipelines.

3.3. Comparison of IPS scores

The IPS scores obtained from the morphological and the different molecular taxonomic assemblages were compared using Spearman's correlation analyses. When using 18S-V4, the IPS scores were mostly well correlated (Table 4, Rho >0.5) apart from the Mothur-FR dataset which was only well correlated with the MetBaN-DE dataset (Rho = 0.7). The correlation between IPS scores computed using MetBaN-DE and SLIM-CH was also weak and non-significant (Rho = 0.36). The IPS scores calculated on morphological assemblages were also not significantly correlated with the molecular IPS scores (Table 4, Rho <0.5).

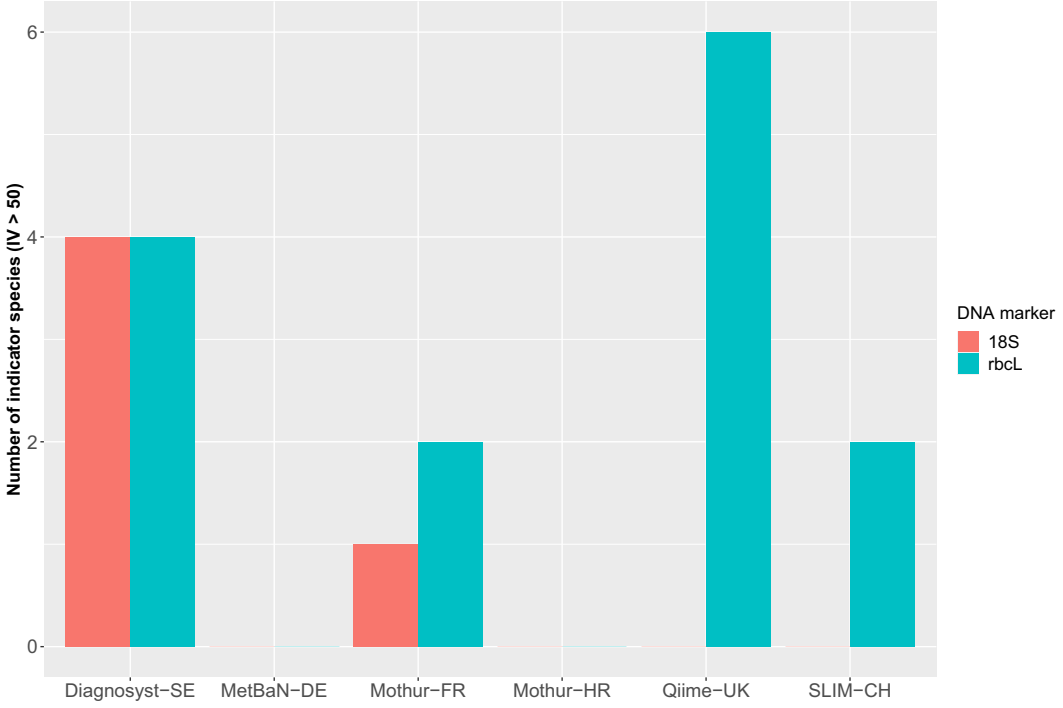


Fig. 3. Number of species with high indicator values in the different molecular inventories. No "high indicator species" are found by the MetBaN-DE and Mothur-HR pipelines on both DNA markers, and by the SLIM-CH pipeline on the 18S-V4 marker.

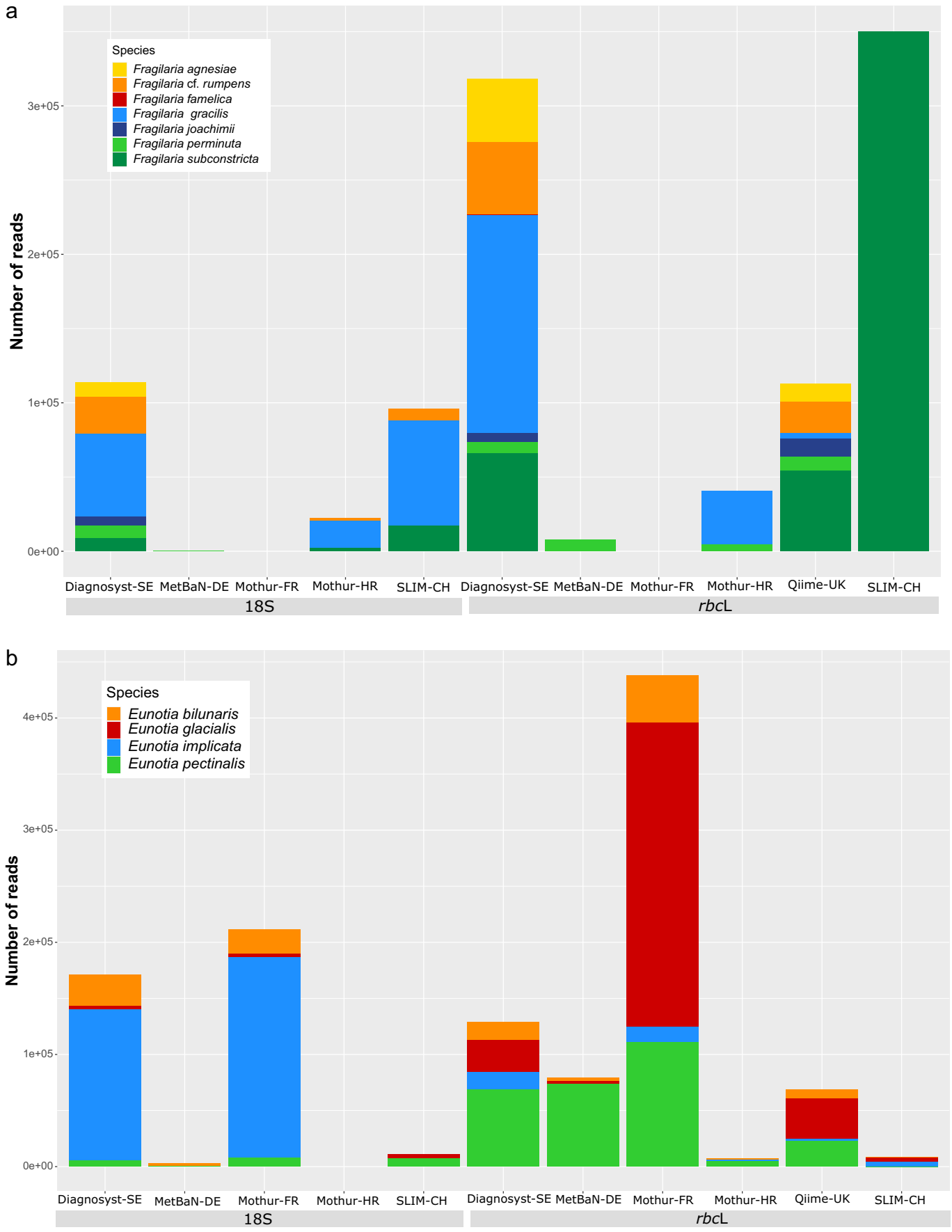


Fig. 4. a. Read numbers annotated to reference database species names for the genus Fragilaria, per DNA marker and bioinformatics pipelines. b. Read numbers annotated to reference database species names for the genus Eunotia, per DNA marker and bioinformatics pipelines.

Table 4
Result of the Spearman (Rho) correlation test comparing the molecular IPS scores to one another and to IPS scores obtained by microscopy. Statistical significance at 0.02 is indicated in bold.

		Mothur-FR	Mothur-HR	MetBaN-DE	SLIM-CH	Diagnosyst-SE	Qiime-UK	Microscopy	
18S-V4 marker	Mothur-FR		**0.89**	**0.84**	**0.63**	**0.92**	**0.95**	**0.66**	*rbc*L marker
	Mothur-HR	0.40		**0.87**	**0.73**	**0.91**	**0.88**	**0.65**	
	MetBaN-DE	**0.70**	**0.63**		**0.66**	**0.88**	**0.88**	**0.65**	
	SLIM-CH	0.09	**0.55**	0.36		**0.72**	**0.69**	**0.58**	
	Diagnosyst-SE	0.34	**0.73**	**0.59**	**0.81**		**0.98**	**0.63**	
	Qiime-UK							**0.63**	
	Microscopy	0.39	0.21	0.39	−0.09	0.20			

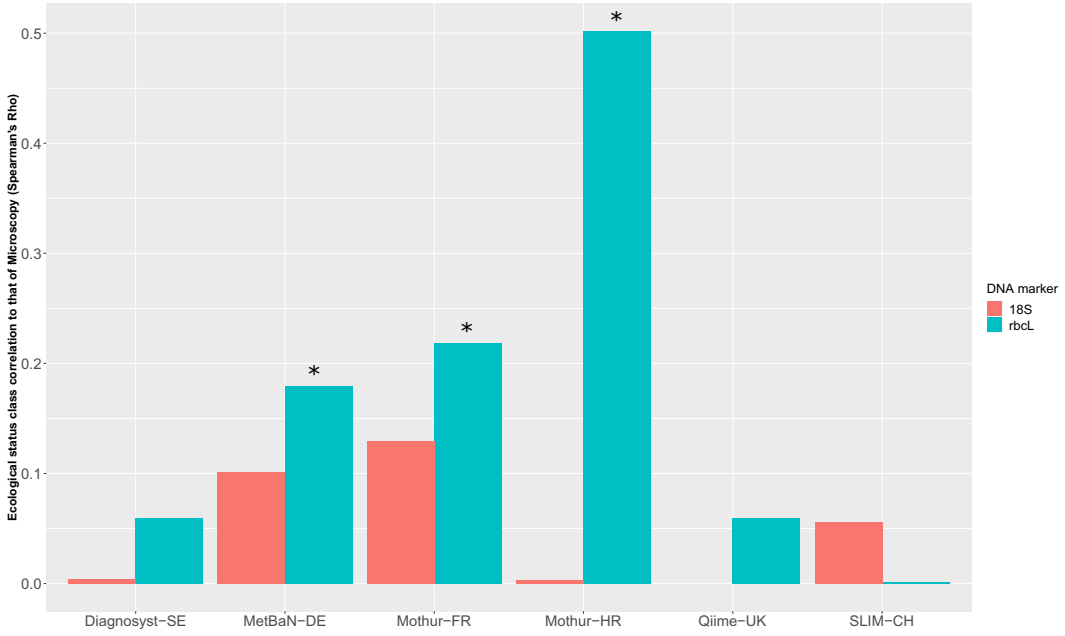


Fig. 5. Spearman's correlation coefficient (Rho) between ecological status classes obtained with the molecular methods and with the microscopy method. Statistical significance ($p = 0.05$) is indicated by black stars.

When using the *rbc*L marker, all the IPS scores calculated on the molecular assemblages were significantly correlated with one another and to the IPS scores computed on the morphology dataset (Table 4, Rho >0.5). The highest correlation with the microscopy IPS scores was obtained with the Mothur-FR pipeline on the *rbc*L marker dataset (Table 4, $R = 0.66$) and the lowest correlation was found with SLIM-CH pipeline when using the 18S-V4 marker dataset (Fig. 4, $R = -0.09$).

3.4. Comparison of ecological status assessment produced by different pipelines

The ecological status classes derived from the ecological index scores differed significantly depending on whether they were generated from molecular or morphological data. Overall, the correlation between morphological and molecular assessments was higher when the *rbc*L marker was used than when using the 18S-V4 marker. With the *rbc*L marker, the bioinformatics pipeline Mothur-HR gave the ecological status classes closest to the microscopy dataset (Fig. 5, $R = 0.50$), followed by the Mothur-FR pipeline (Fig. 5, $R = 0.22$).

The highest proportion of exact agreements with assessments based on microscopy when using the *rbc*L marker was obtained with the Mothur-HR pipeline (Table 5, 64%), closely followed by the Mothur-FR pipeline (Table 5, 50%). When using 18S-V4, the highest proportion of exact estimations and the lowest proportion of underestimations was obtained with the Mothur-FR pipeline (Table 5, respectively 50% and 11%). With both DNA markers, underestimation of ecological status was more frequent than overestimation (Table 5). The Mothur-HR pipeline used with the *rbc*L marker was the only case that avoided any overestimates (Table 5). The highest proportion of underestimates was found when using the MetBaN-DE pipeline for the *rbc*L datasets (Table 5, 50%) and when using the Mothur-HR, Diagnosyst-SE and MetBaN-DE pipelines for 18S-V4 datasets (Table 5, 39%).

Table 5
Percentage of exact agreements, overestimates and underestimates of the ecological status classes, when compared to assessments made with microscopy. The highest percentage of exact assessment is indicated in bold. The ecological assessment is separated into five distinct classes from "high status" (highest) to "bad status" (lowest). In the case of overestimates, the difference ranged from 1 class to 2 classes higher. In the case of underestimates, the difference ranged from 1 class to 4 classes lower.

		Overestimates	Underestimates	Exact
*rbc*L marker	Mothur-FR	14	36	50
	Mothur-HR	0	36	**64**
	Diagnosyst-SE	18	39	43
	MetBaN-DE	11	50	39
	SLIM-CH	29	32	39
	Qiime-UK	19	48	33
18S-V4 marker	Mothur-FR	39	11	**50**
	Mothur-HR	25	39	36
	Diagnosyst-SE	25	39	36
	MetBaN-DE	22	39	39
	SLIM-CH	46	32	22

4. Discussion

The present study kept all variables (samples, DNA extraction, sequencing, reference database, and biotic index) constant except for the bioinformatics pipeline. It aimed to compare the effects of different bioinformatics pipelines, and the implications for ecological assessment, but not to evaluate their effectiveness at characterising the taxonomic composition of the full diatom assemblages found at each study site, nor to reflect the ecological status of those sites. We included the comparison of molecular and microscopical identification of taxa, because the latter is the approach used at present for Water Framework Directive assessments, and we wanted to illustrate the consequences that differences in pipeline and marker performance can have in ecological assessment. However, it is important to remember that morphological analysis includes its own biases and does not necessarily reflect the entire taxa assemblages.

4.1. Major possible sources of discrepancies between the bioinformatics pipelines

One of the main results from our study is that the existing pipelines differ in several aspects, each potentially impacting the outcome. Below, we discuss the potential impact of different bioinformatics parameters. However, for a full understanding, further studies are needed. All the pipelines compared in this study included steps of sequence filtering and chimera removal, which have been shown to be necessary for accurate ecological assessment with molecular data generated from diatoms (Rivera et al., 2020). Across the 6 tested pipelines, a great variety of parameters were used in these two steps.

For sequence filtering, some pipelines used the sequence's quality scores: Mothur-FR used the "phred score", MetBaN-DE used the "alignment score" and Qiime-UK and SLIM-CH used the "mean quality score". Most pipelines also applied some type of length filter, in very different size ranges. For example, when using the *rbc*L marker, sequences retained needed to be longer than 150 bp for the MetBaN-DE pipeline, shorter than 250 bp for the Qiime pipeline, between 253 and 273 bp for the Mothur-FR pipeline and up to 300-315 bp for the Diagnosyst-SE pipeline. Only Mothur-FR and Diagnosyst-SE adapted their length filtering according to the DNA marker used. The Mothur-FR and Mothur-HR pipelines also filtered the sequences according to the length of the overlapping region between the forward and reverse fragment (100 bp for Mothur-FR and 172 bp for Mothur-HR) and the SLIM-CH pipeline did not include any length filtering. Mothur-FR and SLIM-CH also took into account the presence of ambiguous bases. Other parameters included the number of homopolymers (Mothur-FR) and mismatch in the primer sequences (SLIM-CH). Mothur-HR also removed any sequences that did not align perfectly. In terms of sequence filtering, Mothur-FR had the most complex script (five parameters) and Diagnosyst-SE had the simplest one (length filtering only). This means that even at this early stage, all pipelines had removed different sequences and were thus subsequently working on different datasets.

Removal of sequencing errors and chimeras is also a crucial step because these artificial sequences can have an impact on the DNA datasets (clusters made of DNA sequence artefacts), on the species assemblages (false positive in taxonomic assignment), and thus, on subsequent ecological assessment (calculation of index scores). The detection of chimera sequences was performed using different tools, such as VSEARCH (Mothur-FR and SLIM-CH), UCHIME (Mothur-HR) and OBITools (MetBaN-DE). No chimera detection was performed in the Qiime-UK and Diagnosyst-SE pipelines. For both pipelines, chimera detection was deemed unnecessary: the chimeric sequences are excluded by the strict taxonomic alignment algorithm in the Diagnosyst-SE pipeline, and by the Trophic Diatom Index calculations in the original Qiime-UK pipeline. However, in this study we used the IPS biotic index and thus the potential presence of chimera sequences in the Qiime-UK dataset might have had an impact both on the content of the datasets and on subsequent ecological assessment.

Further variations, for example in terms of clustering choices, algorithms and thresholds for taxonomic assignment, also probably contributed to the observed differences in the performance of pipelines. Across the six pipelines tested, three had a clustering step, all using a different algorithm: MetBaN-DE used OBITools at a 99% similarity threshold, Qiime-UK used UCLUST at a 97% similarity threshold and SLIM-CH used VSEARCH at a 97% similarity threshold. The three other pipelines did not include a clustering step: both Mothur scripts used Individual Sequence Units (ISUs), and Diagnosyst-SE used every single query read in the dataset. Creating clusters in the DNA structure of a dataset usually aims at lowering computation time, but at the expense of taxonomic assignment precision.

For the taxonomic assignment, both Mothur scripts used the Wang method but at different confidence scores (60% for Mothur-FR and 85% for Mothur-HR). However, Mothur-HR searched in the reference database for identical sequences during the alignment and hence dismissed query sequences of genotypes unknown to the reference database. This pipeline did not assign a name unless the exact match existed in the reference barcodes, and as a result will always give a lower number of taxa. The Qiime-UK pipeline used BLASTn (95% similarity threshold) and MetBaN used LCS (90% similarity threshold). The SLIM-CH pipeline used VSEARCH and Diagnosyst-SE pipeline has its own supervised clustering algorithm (97% similarity). In addition to the chosen similarity threshold, the parameters chosen for handling taxonomic conflicts can also lead to different final taxonomic assignments.

4.2. Assemblage structure of DNA datasets

We observed OTU clustering scripts producing as many or more taxonomic units in the dataset than ISU scripts. The Mothur-HR had the lowest diversity in terms of DNA assemblages. This pipeline did not perform a clustering step and should have produced a greater number of taxonomic units (in this case ISUs) than other pipelines which are using clustering (and thus producing OTUs). However, the very strict alignment filter resulted in low numbers of ISUs in the final dataset used for taxonomic assignment. Some pipelines also removed singletons (i.e. OTUs represented by only one sequence) but others kept them in their datasets. The latter is the case for the Qiime-UK pipeline (which also did not remove potential chimera sequences), possibly producing by far the greatest amount of observed OTUs. The taxonomic unit assemblages from the Qiime-UK and SLIM-CH pipelines were the most different from all the others, and were both created by clustering at a 97% similarity threshold.

Despite the difference in the number of taxonomic units created, the two Mothur scripts produced a similar structure of DNA assemblages, which may be due to the similarity of their scripts: they both produce ISUs, and use the same tools for cleaning and filtering, albeit with different parameters. The OTUs assemblage of the MetBaN-DE pipeline was also similar to the two ISUs datasets, which can be explained by the high clustering similarity threshold used (99%). A recent study showed that when using the same algorithm (furthest neighbour), a clustering similarity threshold ranging from 95% to 99% should not generate significant differences in the dataset (Tapolczai et al., 2019b). However, the results of Tapolczai et al. (2019b) study cannot be extrapolated to other clustering approaches, and the differences observed in this study seem to originate both from the choice of the clustering algorithm, and from the similarity threshold used.

4.3. Comparison of taxa lists

In this study, 29 samples were pooled together in a single sequencing run (one library for each marker), meaning that we have a good sequencing depth and should be able to detect even taxa with low

abundance and taxa that typically contain a low amount of DNA (e. g. those with a low biovolume). Nevertheless, in this study, the use of a very restricted reference database strongly limited the taxonomic resolution that could be achieved. In order to compare the datasets fairly, we also filtered the light microscopy inventories to keep only taxa that were included in our custom reference database. The results, thus, do not reflect the best performances of either the metabarcoding or microscopy approaches in terms of taxonomic resolution.

The microscopy dataset included 53 taxa once it was filtered. Some pipelines were able to detect most of these 53 taxa; Diagnosyst-SE, for example, detected 47 of them with the 18S-V4 marker and 50 with the *rbc*L marker, while Qiime-UK detected 51 with *rbc*L (Supplementary S3). Even with its strict alignment that dismisses sequences of genotypes unknown to the reference database, Mothur-HR still detected 40 of these taxa with the *rbc*L marker, and 29 with the 18S-V4. Despite using a larger reference database, MetBaN-DE detected the lowest number of taxa recorded by microscopy: 32 when using *rbc*L and 22 when using 18S-V4. This can probably be explained by the intense curation effort, including the adjustment of several species names, which had improved the diat.barcode reference database, the basis for the custom reference database used in this study (Rimet et al., 2019).

All of the pipelines were able to detect the same dominant genera, mostly in similar proportions. However, there were considerable differences in the assignments at species level, and in terms of equivalent quantification of relative abundances compared to microscopy. These limitations of metabarcoding have been highlighted before, and different solutions have been tested, e. g. developing taxonomy-free indices (Apothéloz-Perret-Gentil et al., 2017) or correcting the quantification with a biovolume factor (Vasselon et al., 2018). This biovolume correction factor for diatoms has been developed for *rbc*L marker (for which the copy number will depend on the size of the chloroplast and thus correlate to cell biovolume), but it is currently only available for a few species and ultimately adjust the data back to preconceptions derived from microscopy analyses. In our results, similar patterns appear with the 18S-V4 marker as well (e.g. higher proportions of *Achnanthidium sp*. - mainly *A. minutissimum* - detected with the microscopy) even though the DNA fragment is located in the nucleus. These quantification discrepancies, correlated to cell length and biovolume when using the 18S-V4 marker, have been observed before (Godhe et al., 2008; Mora et al., 2019). A few genera were also detected by only one of the markers, indicating that the discriminating power of the two DNA markers varies between taxa since each marker's reference database contained exactly the same species lists.

At the species level, the Mothur-HR pipeline detected fewer taxa than any other, possibly due to its strict alignment parameter and the low number of ISUs kept in its dataset. In contrast, the Qiime-UK pipeline detected the highest number of taxa overall (75% of which were not included in the microscopy taxa list), probably due to a more relaxed filtering (quality and length filtering only, no removal of singleton or chimeras). The Diagnosyst-SE pipeline detected a large number of taxa as well, despite its very strict taxonomic assignment, also due to a relaxed filtering (only on sequence lengths) that allows it to use almost all the reads in the dataset. On the other hand, the computation power required for this pipeline is much higher than for any of the others. The hardest part of any bioinformatics treatment is to find a balance between data loss and computation time. According to Rivera et al. (2020), working with ISUs instead of OTUs both reduces computation time and improves performance in the calculation of an index for ecological assessment.

The number of unassigned sequences was lowest for the Mothur-HR pipeline, because this pipeline discarded the most reads prior to the taxonomic assignment (strict alignment filtering). By contrast, the SLIM-CH pipeline has the most unassigned reads because it is not optimised for taxonomic assignment (the SLIM-CH pipeline was developed for a taxonomy-free approach) and the assignment performed for this study was done with little optimisation.

Another parameter impacting the number of unassigned reads is the treatment of conflicts in taxonomic assignment (i.e. when a sequence can be matched to several reference barcodes with the same level of similarity): some pipelines, like Diagnosyst-SE will discard such reads altogether, while Mothur-FR and Mothur-HR will provide the assignment at the lowest taxonomic level possible. However, if a pipeline is designed to give the first match it encounters, without accounting for conflicts, the pipeline will be able to provide more species-level identification but with less precision. An advantage of the Mothur software is that it gives information about the lowest taxonomic level it could reach before stopping the assignment, which is useful for detecting shortcomings in the reference database (e. g. missing barcodes in a genus, conflicts between two barcodes). A similar feature is provided in the MetBaN-DE pipeline, which can produce a phylogenetic tree with the unclassified sequences placed among the reference barcodes (and information on the closest species assignments from the sequence placements in the trees). Data lost in bioinformatics analysis significantly affects the quality of the ecological assessment and could be improved by using phylogenetic information in this way (Keck et al., 2018). Qiime-UK and Diagnosyst-SE on the other hand, do not provide any insight into the unassigned reads.

Even with high proportions of unassigned reads, the molecular methods are able to detect many taxa missed by microscopy (Supplementary S3). Because they found the highest number of taxa overall, Qiime-UK and Diagnosyst-SE also had the highest number of taxa detected only by the molecular method (124 and 111 taxa respectively with the *rbc*L, and 97 taxa for Diagnosyst-SE with the 18S-V4). Mothur-HR detected the lowest number of taxa and, consequently the lowest number of taxa detected exclusively by the molecular method (37 with the 18S-V4 marker and 45 with the *rbc*L marker). There are many reasons why a taxon could be missed by microscopy, e.g. being rare, having a small valve, or a fragile valve that could have been destroyed in the preparation process (Zgrundo et al., 2013). Some taxa are also difficult to identify using light microscopy alone, and can easily be misidentified (e. g. small, cryptic or semi-cryptic species: Mann et al., 2010, Kahlert et al., 2019). Of course, metabarcoding will also miss taxa, for example because of a very low amount of DNA, degradation during storage, or damage during DNA extraction, template competition and stochasticity during PCR, loss of amplicons during purification with the magnetic beads, and bias in sequencing (Bálint et al., 2016; Alberdi et al., 2018).

The molecular taxa assemblages closest to the one produced by microscopy were generated by the Diagnosyst-SE and the Mothur pipelines (both the Mothur-FR and Mothur-HR scripts). These three pipelines all handled HTS data in a similar way: none included clustering or gave a taxa name in case of conflict. The strict filtering steps of both Mothur scripts are probably having the same effect as the strict taxonomic assignment in Diagnosyst-SE. The similarity of taxa assemblages is also highlighted by the heatmap (Supplementary S6). For example, the Mothur-FR and Diagnosyst-SE pipelines, which are closest to one another on the heatmap (Euclidean distances), often detect the same taxa and in the same proportions. Taxa assemblages produced by Mothur and Diagnosyst scripts have been compared before, and they were well correlated despite a few important discrepancies (Frigerio et al., 2016). The SIMPER analysis also highlighted the taxa which accounted for most of the dissimilarity when comparing taxa assemblages between pipelines and markers. It was not surprising to find *Achnanthidium minutissimum* and *Navicula tripunctata* among those accounting for >20% of the dataset's dissimilarities: both species are frequently found in Fennoscandia environmental samples, and both have been shown to cause quantification bias before (Vasselon et al., 2018). However, it is possible that the percentage of dissimilarities explained by these two species reflects both the actual biological differences and a stochastic effect (combined effect of frequency across the samples and high abundance). In general, abundant taxa drove differences in the assemblages. For example, *Eunotia* species are abundant and

important indicator taxa in Scandinavian freshwaters, but are underrepresented in barcode reference libraries, reducing opportunities for accurate identification below genus level.

The heatmap also highlights an interesting gap in the detection of species: the Mothur-FR pipeline did not detect any species of Fragilariophyceae (see Supplementary S6 for details), with either the *rbc*L or 18S-V4 markers. This group was, on the other hand, detected without problem by the Diagnosyst-SE and Qiime-UK pipelines, but was relatively scarce using MetBaN-DE or Mothur-HR pipelines. In order to understand what could have caused such a difference in detection of the Fragilariophyceae, we looked into the details of the Mothur-FR script parameters and discovered a syntax inconsistency in the get. lineage command (the "-" was not used to select the Fragilariophyceae group, although it was applied to all other diatom groups). Most Mothur software versions are not sensitive to this inconsistency (e.g. version 1.41.3), however it appeared that the version of Mothur software used in this manuscript (1.43.0) was sensitive. After correcting the syntax in the command, all Fragilariophyceae taxa were detected when using the script, with all other parameters unchanged, leading to results more consistent with those of the Diagnosyst-SE pipeline, on both markers (Supplementary S7).

In this study, we have already shown that different bioinformatics pipelines do generate different results, leading to different interpretations. The error detected in the Mothur-FR script however reminds us that a bioinformatics pipeline is also sensitive to software evolution. This is an important factor to consider when using bioinformatics pipelines and highlights the need for quality control measures when assessing the reliability of a pipeline output (e.g. by including a known mock sample as positive control into the dataset to check taxa detection). Such controls are especially important to enable the use of metabarcoding in routine assessment, as suggested by Elbrecht and Steinke (2019) and Zafeiropoulos et al. (2020). Bioinformatics tools and software are in constant development, and users cannot trust their outputs blindly when the exact same script can work differently, depending on the software version. This could limit the reproducibility of metabarcoding methods and control standards need to be implemented for future development and routine analysis.

4.4. Comparison of ecological assessments

Decisions made in pipeline design have implications for taxa assemblages, which in turn have implications for biotic indices and, ultimately, for ecological assessment.

When using the *rbc*L marker, the derived IPS scores calculated on our molecular datasets were well correlated, not only between pipelines but also when compared to the IPS scores derived from light microscopy. The Mothur-HR pipeline produced the IPS scores second-closest to those produced by microscopy, despite having the lowest number of taxa detected. This shows that the detection of the dominant taxa is sufficient to compute an adequate IPS score, and that the index is robust enough to be used with HTS data. When looking back at the relative abundance of genera detected by each pipeline, the Mothur-HR pipeline calculated relative abundances of *Achnanthidium* that were slightly closer to the relative abundances detected by microscopy. Since the SIMPER analysis identified *Achanthidium minutissimum* as one of the species which accounted for most of the dissimilarity between the molecular datasets, these slight differences in relative abundances could explain in part why the ecological status class produced using this pipeline showed a higher correlation to those of the microscopy approach.

The patterns observed with the index scores were reflected in the derived ecological status classes: the correlation to the microscopy results was overall higher when using *rbc*L. Mothur-HR provided the highest number of assessments matching the results derived from microscopy, and no overestimations compared to microscopy. All pipelines underestimated ecological status classes compared to microscopy. The least underestimation of ecological status was obtained when using the Mothur-FR pipeline with 18S-V4; however, the correlation of these results to the microscopy-based ecological assessment was weak and non-significant. The highest number of underestimations of ecological status occurred when using the MetBaN-DE pipeline with the *rbc*L marker. Incidentally, the proportions of *Achnanthidium* species were strongly underestimated when using this pipeline with the *rbc*L marker (Fig. 2) and the SIMPER analysis showed that *Achnanthidium minutissimum* had a relevant impact on the assemblages' discrepancies (Table 3) and, in turn, on the IPS scores. This discrepancy in *Achnanthidium* proportions could explain, in part, the ecological assessment underestimates when using the MetBaN-DE pipeline. Conversely, *Navicula* species were overrepresented when using the 18S-V4 marker (Fig. 2) and the SIMPER analysis highlighted three *Navicula* species, typically representative of relatively good ecological quality, which most likely impacted the IPS scores (Table 3). This could explain the higher rate of overestimation of ecological assessment when using the 18S-V4 marker compared to the *rbc*L marker.

We now need to disentangle a complex array of differences in taxa absence, presence or abundance between pipelines, and then understand the effect of those differences on index scores and, in turn, on ecological status classes. For example, while the three pipelines showing the closest correlations (with the *rbc*L marker) between their IPS scores and those produced by microscopy were Mothur-FR, Mothur-HR and MetBaN-DE, only the first two also had a relatively high correlation with the taxa list from microscopical analysis. Rather, it was the Diagnosyst-SE pipeline which produced the taxa assemblages with the highest correlation to those obtained from microscopy, but the IPS scores were not well correlated. On the other hand, the good correlation of IPS scores was not reflected by a good correlation of taxa detection in the MetBaN-DE pipeline. This means that a good taxa detection was not a prerequisite to achieve IPS values comparable to those obtained from microscopical analysis. The reason is probably a combination of factors, different for each pipeline: The result of the IPS calculation is dependent on both taxa presence, but also on the abundance of taxa, and on their IPS sensitivity values. However, the differences between the pipelines are mixing those factors in many combinations, and disentanglement of these, along with analysis of their importance will need further study. As an example, abundance differences were shown to be important, but the three pipelines with the closest correlation to the microscopical values had quite different abundances of the most dominant taxa, some of them with higher IPS sensitivity value than others. One plausible explanation of generally underestimated IPS scores across all pipelines, and in turn, ecological status classes, is the general underrepresentation of *A. minutissimum* across all pipelines and both markers, when compared to microscopy. This taxon was often dominant in microscopical analyses and has the highest IPS sensitivity value. It can therefore only be replaced in relative abundance by taxa with similar or lower IPS value, in turn leading to an overall underestimation of IPS scores. However, the pipelines and markers differed in which taxa were replacing *A. minutissimum* in relative abundances, and while both abundance and taxa detection differences were important, there was, however, no unifying pattern across all samples and pipelines. A detailed analysis of which taxa exactly were over/underrepresented in the different pipelines and thus contributed to a higher or lower IPS score of output dataset would raise the complexity of this comparison to a level that would be hard to evaluate (especially considering the high variation of sizes/biovolumes of the different taxa). A first attempt has been performed by Pérez-Burillo et al. (2020) for the Mothur pipeline, where, again, *A. minutissimum* was found as one of the most important taxa impacting the IPS scores.

It is important to point out that because our custom reference database limits the diversity of taxa that could be detected by the metabarcoding approach and because we filtered the microscopy inventories to match it, derived IPS scores and ecological quality status obtained in this study are not an accurate reflection of real environmental conditions, but rather of the performance of one method compared to another.

Two of the pipelines included in this study are also usually used with another index: The TDI for the Qiime-UK pipeline and the DICH index for the SLIM-CH pipelines. The latter was developed based on molecular data only, and is not dependent on taxonomic assignment. This approach, however, needs to be tested and calibrated on a wider training dataset to enable its application to different ecoregions (Apothéloz-Perret-Gentil et al., 2017) and may not be consistent with current wording of the WFD. Overall, the potential of indices not based on relative abundances of taxa but on other metrics adapted to HTS data should be investigated in the future. For example, with the development of new metrics optimised to the properties of HTS data, rather than using HTS data with metrics that were developed and optimised for morphology-based assessment (Kelly et al., 2018b).

In conclusion, it is impossible to recommend one pipeline over another in this study although a few key lessons can be learned regarding their use for environmental assessment. Differences in clustering methods were a great source of disparity in the results, and computation time could be lowered with strict filtering and the use of ISUs units. Relaxed filtering combined with a strict assignment algorithm allowed for a better taxonomic assignment with fewer conflicts but at the expense of increased computation time. The balance between sequence filtering (data loss), computation time and accuracy of the results was optimal in the Mothur-HR script for ecological assessment with the IPS, but did not catch the whole taxa diversity. The Diagnosyst-SE pipeline would probably perform best in cases where the taxonomic diversity is the focus. The choice of parameters within a pipeline should thus be motivated by the objective of the bioinformatics analyses: taxa diversity, with presence/absence and relative abundance data, ecological index etc. There is also a strong case for achiving raw metabarcoding outputs to enable data to be reworked as understanding of bioinformatics improves.

Quantification discrepancies of taxa are a common problem of the current approaches used for diatoms metabarcoding, and one of the main challenges of its use for monitoring. Ji et al. (2020) recently developed a pipeline to enable the accurate quantification of eukaryotic taxa by combining shotgun sequencing, DNA mapping to mitogenomes and the combination of three correction factors to remove false positives and to correct the stochasticity effect during sequencing and across runs. This pipeline opens new potential for taxa quantification, but seems rather complex for an application to mass routine monitoring. It becomes more and more apparent that the use of relative abundance may not be the gold standard in metabarcoding, and the potential of other metrics such as taxa presence/absence (Buchner et al., 2019) and life forms need to be investigated for the development of future approaches.

4.5. Comparison of the two DNA markers

4.5.1. Preferential identification of taxa

Our results show that different numbers of reads were assigned to a specific taxon depending on which marker was used. The results could hint at greater variability of the 18S-V4 DNA fragment in diatoms in our dataset as compared to the *rbc*L fragment. The 18S-V4 DNA fragment is a well-known hyper-variable region, which makes it interesting for barcoding and metabarcoding, but it is not always easily aligned when using automatic alignment tools (e.g. MAFFT, MUSCLE, Clustal). It is possible that some insertions/deletions might be mistaken as chimeras (Boyer et al., 2016). Frequent introns can also make the sequences much longer than the lengths set for sequence filtering in the pipeline's parameters, therefore those sequences might be discarded (Gaonkar et al., 2018). Alternatively, indels and Taq substitution errors that could be actual artefacts might not be detected and removed during the filtering, therefore increasing the number of OTUs (Brown et al., 2015; Bálint et al., 2016; Tedersoo et al., 2018). In comparison, the *rbc*L fragment aligns more easily and allows for a more straightforward detection of error/chimeras.

Some taxa were detected by both markers, and others were only detected by one of them. For example, *Navicula tripunctata* is often an abundant species in biofilm samples, and was detected by all pipelines and both markers, albeit with great variation in the number of reads among datasets, especially when using the 18S-V4 marker. The discrepancies were even more marked for *Eunotia* and *Fragilaria* species (Fig. 4a and b). One possible reason is that while our custom reference database included reference barcodes for the same taxa with both markers, it did not account for the strain origin or for the number of reference sequences included. For example, there are two barcodes for *Amphora pediculus* in the 18S-V4 reference database, while there are 22 barcodes in the *rbc*L reference database. Thus, the *rbc*L reference database allows for more variations in the gene marker and may, in turn, detect a wider diversity of query sequences of *A. pediculus*.

However, in the case of the *Eunotia* species, *E. bilunaris* was represented by 11 barcodes in the *rbc*L reference database and only one barcode in the 18S-V4 reference database. Nonetheless, the species was detected in similar proportions by both markers. And while both reference databases had only one barcode each for *E. pectinalis*, *E. glacialis* and *E. implicata* respectively, these three species were found in very different proportions depending on which marker or pipeline was used. In this case, the reference database was not necessarily the origin of these discrepancies, and they probably result from other bias in the metabarcoding approach such as template competition and stochasticity during PCR and sequencing (Bálint et al., 2016; Alberdi et al., 2018). Another explanation might be that some taxa are better separated when using one or the other DNA fragment, or that Scandinavian diatoms might be slightly genetically different from the available reference barcodes of Central European strains.

4.5.2. Better pipeline performance with the rbcL reference database and datasets

Overall, we observed fewer discrepancies, both in taxa assemblages and in ecological assessment (IPS scores and ecological status), when using the *rbc*L marker than when using the 18S-V4 marker. The *rbc*L datasets included more taxa diversity, and generated assemblages that were better correlated to each other and to the microscopy dataset. The index scores calculated and the status classes derived from these were also better correlated to microscopy-based ecological assessment when using the *rbc*L marker. Both DNA markers perform well for diatom identification for monitoring purposes (while acknowledging that the use of short DNA fragment limits taxa discrimination) and all of the pipelines selected in this study have been optimised for a use with one of these markers. MetBaN-DE and SLIM-CH were developed and optimised on HTS data of the 18S-V4 DNA marker, but these pipelines still performed better with the *rbcL* marker in the present study.

We used three pairs of degenerated primers for the *rbc*L fragment, raising its detection and discrimination power, but only one pair of primers for the 18S-V4 fragment. Moreover, a recent extensive curation effort on the diat.barcode reference database (Rimet et al., 2019) has improved the quality of taxonomic assignment with the *rbc*L marker. These curation efforts have been focused on the *rbc*L barcodes only, and no similar curation has been done on the 18S-V4 barcodes yet. Even with our "partial" reference database, this could explain why the *rbc*L marker performed better than the 18S-V4 marker in our study.

We recommend using a well-curated reference database rather than a non-curated one, however, extensive, for quality purposes, because accurate taxonomic assignments, even with low numbers of taxa, proved efficient for ecological assessment.

5. Conclusion

This exercise provides an overview and deeper insight into the approaches currently developed for environmental assessment based on diatom metabarcoding. All the pipelines tested in this study were heavily optimised by their working group to produce the best possible representation of the HTS data. However, the different antecedents of the working groups and different training datasets used in the development of the

pipelines (different ecoregion, sequencing technologies, reference database, ecological index…) resulted in relatively low reproducibility of the pipelines performances when confronted with a different "type" of data. Discrepancies in the detection and quantification of the relative abundance of taxa, both between the pipeline's outcomes, and between HTS and microscopy data, remained a dominant problem. However, the large variation in bioinformatics parameters between pipelines made it difficult to disentangle their effects on the taxa list outcome. Any future intercalibration exercises comparing molecular methods thus need to make sure that a standardized pipeline is used for data analysis. We also found a direct error in one of the scripts leading to taxa differences, highlighting the need for quality control in routine analysis using positive controls with calibrated and mock data (Siegwald et al., 2017, Elbrecht and Steinke, 2019). Such positive controls would also ensure that newly developed metrics based on molecular data are broadly applicable and not specific to particular pipelines. Our study clearly shows that there is a need to encapsulate best practice for bioinformatics pipelines in a European technical standard to ensure that datasets are compatible and reflect the entire natural diatom assemblage (Kelly et al., 2019). We would also like to stress the need for good representation of the assemblages (bearing in mind that microscopy, itself, gives a biased picture), to allow for future development of metrics based not just on the existing approaches but also on the analysis of life-forms or guilds, which still need underlying species data. The future of ecological assessment with diatom HTS data probably lies in the development of new indices using, for example, the information that molecular data can give us on (semi)cryptic taxa and the unravelling of species complexes. However, to make the best use of such data, we need large-scale attempts to better understand diatom ecology and distribution. Our critical comparison of molecular and microscopical data was necessary to understand how taxonomic information is affected by bioinformatics. However, the longer-term goal should be to break free from the preconceptions we have brought with us from careers based around light microscopy and to recognise HTS data as distinct. A critical comparison of HTS and microscopy is a necessary starting point, but may eventually become a constraint.

CRediT authorship contribution statement

Bonnie Bailet: Conceptualization, Formal analysis, Methodology, Investigation, Validation, Visualization, Writing - original draft, Writing - review & editing. **Laure Apothéloz-Perret-Gentil:** Data curation, Resources, Writing - review & editing. **Ana Baričević:** Data curation, Writing - review & editing. **Teofana Chonova:** Data curation, Writing - review & editing. **Alain Franc:** Data curation, Resources, Writing - review & editing. **Jean-Marc Frigerio:** Data curation, Resources. **Martyn Kelly:** Investigation, Resources, Writing - review & editing. **Demetrio Mora:** Visualization, Writing - review & editing. **Martin Pfannkuchen:** Data curation, Funding acquisition, Resources, Writing - review & editing. **Sebastian Proft:** Data curation, Resources, Writing - review & editing. **Mathieu Ramon:** Data curation, Resources. **Valentin Vasselon:** Conceptualization, Methodology, Data curation, Resources, Writing - review & editing. **Jonas Zimmermann:** Funding acquisition, Resources, Writing - review & editing. **Maria Kahlert:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project was partly funded by Stiftelsen Oscar och Lili Lamms Minne (http://www.stiftelsenlamm.a.se/) and by the Swedish Agency for Marine and Water Management. We would like to thank the Norwegian Institute for Water Research (NIVA), which provided the water chemistry data for all Norwegian samples. We also would like to thank Jukka Aroviita and the MaaMet project financed by the Ministry of Agriculture and Forestry in Finland, as well as Satu-Maaria Karjalainen for providing the Finnish samples and water chemistry data. The development of the MetBaN pipeline was supported by the Federal Ministry of Education and Research (German Barcode of Life 2 Diatoms (GBOL2), grant number 01LI1501E, website: www.bolgermany.de). This work was also partially supported by the Croatian Science Foundation project: Life strategies of phytoplankton in the northern Adriatic (UIP-2014-09-6563) and by the European COST-Action DNAqua Net (CA15219).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.140948.

References

Alberdi, A., Aizpurua, O., Gilbert, M.T.P., Bohmann, K., 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods Ecol. Evol. 9, 134–147.

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. Mol. Ecol. Resour. 17, 1231–1242.

Bailet, B., Bouchez, A., Franc, A., Frigerio, J.-M., Keck, F., Karjalainen, S.-M., Rimet, F., Schneider, S., Kahlert, M., 2019. Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe freshwater and consequences for ecological status. Metabarcoding and Metagenomics 3.

Bálint, M., Bahram, M., Eren, A.M., Faust, K., Fuhrman, J.A., Lindahl, B., O'Hara, R.B., Öpik, M., Sogin, M.L., Unterseher, M., Tedersoo, L., 2016. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiol. Rev. 40, 686–700.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., Coissac, E., 2016. obitools: a unix-inspired software package for DNA metabarcoding. Mol. Ecol. Resour. 16, 176–182.

Brown, E.A., Chain, F.J.J., Crease, T.J., MacIsaac, H.J., Cristescu, M.E., 2015. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecology and Evolution 5, 2234–2251.

Buchner, D., Beermann, A.J., Laini, A., Rolauffs, P., Vitecek, S., Hering, D., Leese, F., 2019. Analysis of 13,312 benthic invertebrate samples from German streams reveals minor deviations in ecological status class between abundance and presence/absence data. PLoS One 14, e0226547.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336.

Cemagref, 1982. Etudes des methodes biologiques quantitative d'appreciation de la qualite des eaux.

CEN, 2014. Water Quality - CEN/TC 230 - Guidance for the Routine Sampling and Preparation of Benthic Diatoms From Rivers and Lakes. CEN Stand.

CEN, 2018a. Water Quality - CEN/TR 17244 - Technical Report for the Management of Diatom Barcodes. CEN Stand.

CEN, 2018b. In: C. Stand (Ed.), Water Quality - CEN/TR 17245 - Technical Report for the Routine Sampling of Benthic Diatoms From Rivers and Lakes Adapted for Metabarcoding Analyses, pp. 1–8.

Chonova, T., Kurmayer, R., Rimet, F., Labanowski, J., Vasselon, V., Keck, F., Illmer, P., Bouchez, A., 2019. Benthic diatom communities in an Alpine river impacted by waste water treatment effluents as revealed using DNA metabarcoding. Front. Microbiol. 10, 653.

Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18, 117–143.

Commission, E, 2000. In: O. J. o. E. Communities (Ed.), Directive 2000/60/EC of the European Parliament and of the Council of 23 Rd October 2000 Establishing a Framework for Community Action in the Field of Water Policy, pp. 1–72.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F., Notredame, C., 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 39, W13–W17.

Dufrene, M., Legendre, P., 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. 67, 345–366.

Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L.A., Pawlowski, J., Cordier, T., 2019. SLIM: a flexible web application for the reproducible processing of environmental DNA metabarcoding data. BMC Bioinformatics 20, 88.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.

Elbrecht, V., Steinke, D., 2019. Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. Freshw. Biol. 64, 380–387. https://doi.org/10.1111/fwb.13220.

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

Frigerio, J.-M., Rimet, F., Bouchez, A., Chancerel, E., Chaumeil, P., Salin, F., Thérond, S., Kahlert, M., Franc, A., 2016. diagno-syst: A Tool for Accurate Inventories in Metabarcoding. arXiv:1611.09410 q-bio.QM.

Gaonkar, C.C., Piredda, R., Minucci, C., Mann, D.G., Montresor, M., Sarno, D., Kooistra, W.H.C.F., 2018. Annotated 18S and 28S rDNA reference sequences of taxa in the planktonic diatom family Chaetocerotaceae. PLoS One 13, e0208929.

Godhe, A., Asplund, M.E., Härnström, K., Saravanan, V., Tyagi, A., Karunasagar, I., 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. Appl. Environ. Microbiol. 74, 7174–7182.

Hammer, Ø., Harper, D.A.T., Ryan, P.D., 2001. PAST: paleontological statistics software package for education and data analysis. Palaeontol. Electron. 4–9.

Jarlman, A., Kahlert, M., Sundberg, I., Herlitz, E., 2016. Påväxt i sjöar och vattendrag – kiselalgsanalys. Version 3:2: 2016-01-20. Handledning för miljöövervakning Undersökningstyp. Havs- och Vattenmyndigheten, Göteborg.

Ji, Y., Huotari, T., Roslin, T., Schmidt, N.M., Wang, J., Yu, D.W., Ovaskainen, O., 2020. SPIKEPIPE: a metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Mol. Ecol. Resour. 20, 256–267.

Joshi, N., Fass, J., 2011. Sickle: A Sliding-window, Adaptive, Quality-based Trimming Tool for FastQ Files (Version 1.21) [Software].

Kahlert, M., Andrén, C., HJarlman, A., 2007. Bakgrundsrapport för revideringen 2007 av bedömningsgrunder för Påväxt – kiselalger i vattendrag [Background report for revision 2007 of assessment criterua for periphyton - diatoms in watercourses]. Report 2007. Swedish University of Agricultural Sciences (SLU), Department of Environmental Assessment, p. 23.

Kahlert, M., Albert, R.-L., Anttila, E.-L., Bengtsson, R., Bigler, C., Eskola, T., Gälman, V., Gottschalk, S., Herlitz, E., Jarlman, A., Kasperovičienė, J., Kokocinski, M., Luup, H., Miettinen, J., Paunksnyte, I., Piirsoo, K., Quintana, I., Raunio, J., Sandell, B., Weckström, J., 2008. Harmonization is more important than experience—results of the first Nordic–Baltic diatom intercalibration exercise 2007 (stream monitoring). J. Appl. Phycol. 21, 471–482.

Kahlert, M., Kelly, M.G., Mann, D.G., Rimet, F., Sato, S., Bouchez, A., Keck, F., 2019. Connecting the morphological and molecular species concepts to facilitate species identification within the genus *Fragilaria* (Bacillariophyta). J. Phycol. 55, 948–970.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Keck, F., Vasselon, V., Rimet, F., Bouchez, A., Kahlert, M., 2018. Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. Mol. Ecol. Resour. 18, 1299–1309.

Kelly, M., Bennett, C., Coste, M., Delgado, C., Delmas, F., Denys, L., Ector, L., Fauville, C., Ferréol, M., Golub, M., Jarlman, A., Kahlert, M., Lucey, J., White, B., Pardo, I., Pfister, P., Picinska-Faltynowicz, J., Tison-Rosebery, J., Schranz, C., Vilbaste, S., 2009. A comparison of national approaches to setting ecological status boundaries in phytobenthos assessment for the European Water Framework Directive: results of an intercalibration exercise. Hydrobiologia 621, 169–182.

Kelly, M., Urbanic, G., Acs, E., Bennion, H., Bertrin, V., Burgess, A., Denys, L., Gottschalk, S., Kahlert, M., Karjalainen, S.M., Kennedy, B., Kosi, G., Marchetto, A., Morin, S., Picinska-Fałtynowicz, J., Poikane, S., Rosebery, J., Schoenfelder, I., Schoenfelder, J., Varbiro, G., 2014. Comparing aspirations: intercalibration of ecological status concepts across European lakes for littoral diatoms. Hydrobiologia 734, 125–141.

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann, D., Pass, D., Sapp, M., Sato, S., Glover, R., 2018a. A DNA Based Diatom Metabarcoding Approach for Water Framework Directive Classification of Rivers.

Kelly, M., Juggins, S., Phillips, G., Willby, N., 2018b. Evaluation of Benthic Diatom Classification in UK Rivers Using LM and NGS Methods. Report Number E18–56 (PO 4057194).

Kelly, M.G., Chiriac, G., Soare-Minea, A., Hamchevici, C., Juggins, S., 2019. Use of phytobenthos to evaluate ecological status in lowland Romanian lakes. Limnologica 77, 125682.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.F., Bouchez, A., 2013. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Mol. Ecol. Resour. 13, 607–619.

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, 349–363.

Lecointe, C., Coste, M., Prygiel, J., 1993. "Omnidia": software for taxonomy, calculation of diatom indices and inventories management. Hydrobiologia 269-270, 509–513.

Mann, D., Sato, S., Trobajo, R., Vanormelingen, P., Souffreau, C., 2010. DNA barcoding for species identification and discovery in diatoms. Cryptogam. Algol. 31, 557–577.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 17, 3.

Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., Neufeld, J.D., 2012. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 13, 31.

McCune, B., Mefford, M.J., 2011. PC-ORD. Multivariate Analysis of Ecological Data. Version 6. MjM Software, Gleneden Beach, Oregon.

Mora, D., Abarca, N., Proft, S., Grau, J., Enke, N., Jiménez, J., Skibbe, O., Jahn, R., Zimmermann, J., 2019. Morphology and metabarcoding: a test with stream diatoms from Mexico highlights the complementarity of identification methods. Freshwater Science 38, 448–464.

Mortágua, A., Vasselon, V., Oliveira, R., Elias, C., Chardon, C., Bouchez, A., Rimet, F., João Feio, M., Almeida, S.F.P., 2019. Applicability of DNA metabarcoding approach in the bioassessment of Portuguese rivers using diatoms. Ecol. Indic. 106, 105470.

Pérez-Burillo, J., Trobajo, R., Vasselon, V., Rimet, F., Bouchez, A., Mann, D.G., 2020. Evaluation and sensitivity analysis of diatom DNA metabarcoding for WFD bioassessment of Mediterranean rivers. Sci. Total Environ. 727, 138445. https://doi.org/10.1016/j.scitotenv.2020.138445.

Proft, S., Grau, J., Caswara, C., Mazzoni, C., Mora, D., Zimmermann, J., 2017. MetBaN Automated Pipeline for Metabarcoding Data Using Taxonomical/Phylogenetical Classification of Organisms.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rimet, F., Abarca, N., Bouchez, A., Kusber, W.-H., Jahn, R., Kahlert, M., Keck, F., Kelly, M., Mann, D., 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 18, 37–54.

Rimet, F., Gusev, E., Kahlert, M., Kelly, M.G., Kulikovskiy, M., Maltsev, Y., Mann, D.G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., Bouchez, A., 2019. Diat.barcode, an open-access curated barcode library for diatoms. Sci. Rep. 9, 15116.

Rivera, S.F., Vasselon, V., Bouchez, A., Rimet, F., 2020. Diatom metabarcoding applied to large scale monitoring networks: optimization of bioinformatics strategies using Mothur software. Ecol. Indic. 109, 105775.

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

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., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C., Caboche, S., 2017. Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. PLoS One 12, e0169563. https://doi.org/10.1371/journal.pone.0169563.

SIS - Standardiseringens stöd, 2014a. Water Quality - Guidance for the Identification and Enumeration of Benthic Diatom Samples From Rivers and Lakes. p. 24.

SIS - Standardiseringens stöd, 2014b. Water Quality - Guidxtance for the Routine Sampling and Preparation of Benthic Diatoms from Rivers and Lakes. p. 28.

Smol, J.P., Stoermer, E.F., 2010. The Diatoms: Applications for the Environmental and Earth Sciences. Second edition. Cambridge University Press.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Stein, E.D., White, B.P., Mazor, R.D., Miller, P.E., Pilgrim, E.M., 2013. Evaluating ethanol-based sample preservation to facilitate use of DNA barcoding in routine freshwater biomonitoring programs using benthic macroinvertebrates. PLoS One 8, e51273.

Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., Vasselon, V., 2019a. Diatom DNA metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. Front. Ecol. Evol. 7, 409.

Tapolczai, K., Vasselon, V., Bouchez, A., Stenger-Kovács, C., Padisák, J., Rimet, F., 2019b. 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, 166–179.

Tedersoo, L., Tooming-Klunderud, A., Anslan, S., 2018. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol. 217, 1370–1385.

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? Freshwater Science 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.

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.

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, 7597–7605.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267.

Weigand, H., Beermann, A.J., Čiampor, F., Costa, F.O., Csabai, Z., Duarte, S., Geiger, M.F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A.M., Willassen, E., Wyler, S.A., Bouchez, A., Borja, A., Čiamporová-Zat'ovičová, Z., Ferreira, S., Dijkstra, K.-D.B., Eisendle, U., Freyhof, J., Gadawski, P., Graf, W., Haegerbaeumer, A., van der Hoorn, B.B., Japoshvili, B., Keresztes, L., Keskin, E., Leese, F., Macher, J.N., Mamos, T., Paz, G., Pešić, V., Pfannkuchen, D.M., Pfannkuchen, M.A., Price, B.W., Rinkevich, B., Teixeira, M.A.L., Várbíró, G., Ekrem, T., 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: gap-analysis and recommendations for future work. Sci. Total Environ. 678, 499–524.

Zafeiropoulos, H., Viet, H.Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C., Pafilis, E., 2020. PEMA: a flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. GigaScience 9. https://doi.org/10.1093/gigascience/giaa022.

Zgrundo, A., Lemke, P., Pniewski, F., Cox, E.J., Łatała, A., 2013. Morphological and molecular phylogenetic studies on Fistulifera saprophila. Diatom Research 28, 431–443.

Zhang, J., Kobert, K., Flouri, T., Stamatakis, A., 2013. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614–620.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., Gemeinholzer, B., 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Mol. Ecol. Resour. 15, 526–542.