Evaluation of two short overlapping *rbcL* markers for diatom metabarcoding of environmental samples: effects on biomonitoring assessment and species resolution.

Javier Pérez-Burillo[1,2*], David G. Mann[1,3] & Rosa Trobajo[1]

[1]IRTA-Institute for Food and Agricultural Research and Technology, Marine and Continental Waters Programme. Ctra de Poble Nou Km 5.5, E43540, LaRàpita, Tarragona, Spain.

[2]Departament de Geografia, Universitat Rovira i Virgili, C/ Joanot Martorell 15, E43500, Vila-seca, Tarragona, Spain

[3]Royal Botanic Garden Edinburgh, Edinburgh, EH3 5LR, Scotland, UK

*Corresponding author

E-mail addresses: jperezburillo@gmail.com (J, Pérez-Burillo), dmann@rbge.org.uk (D.G, Mann), rosa.trobajo@irta.cat (R, Trobajo),

Abstract

Two short diatom *rbcL* barcodes, 331-bp and 263-bp in length, have frequently been used in diatom metabarcoding studies. They overlap in a common 263-bp region but differ in the presence or absence of a 68-bp tail at the 5′ end. Though the effectiveness of both has been demonstrated in separate biomonitoring and diversity studies, the impact of the 68-bp non-shared region has not been evaluated. Here we compare the two barcodes in terms of the values of a biotic index (IPS) and the ecological status classes derived from their application to an extensive metabarcoding dataset from United Kingdom rivers; this comprised 1703 samples and was produced using the 331-bp primers. In addition, we assess the effectiveness of each barcode for discrimination of genetic variants around and below the species level. The strong correlation found in IPS values between barcodes (Pearson's R = 0.98) indicates that the choice of the barcode does not have major

implications for current WFD ecological assessments, although a very few sites (55: 3.23% of those analysed) were downgraded from an acceptable WFD class ("good") to an unacceptable one ("moderate"). Analyses of the taxonomic resolution of the two barcodes indicate that for many ASVs, the use of either marker – 263-bp and 331-bp – gives unambiguous assignations at species level though with differences in bootstrap confidence values. Such differences are caused by the stochasticity involved in the naïve Bayesian classifier used and by the fact that genetic distance, regarding closely related species, is increased when using the 331-bp barcode. However, in three cases, species differentiation fails with the shorter marker, leading to underestimates of species diversity. Finally, two ASVs from *Nitzschia* species evidenced that the use of the shorter marker can sometimes lead to false positives when the extent and nature of infraspecific variation are poorly known.

Key words: Water Framework Directive, ecological assessment, infraspecific variation, High-throughput sequencing, species discrimination,

1. Introduction

Diatom DNA metabarcoding of environmental samples has proved to be an efficient method for biomonitoring purposes and the study of species diversity (e.g. Bailet et al., 2019; De Luca et al., 2021; Kelly et al., 2020; Mortágua et al., 2019; Pérez-Burillo et al., 2020; Stoof-Leichsenring et al., 2020; Vasselon et al., 2017). This method (metabarcoding of environmental samples) is based on high-throughput sequencing (HTS) of a particular barcode of interest that must offer good resolution at species level. The reduced cost and the availability of MiSeq sequencing technology have made it the most often used HTS technology nowadays, superceding previous technologies (e.g. 454 GS-FLX with achievable read-lengths of 900-bp, Ion Torrent). However, MiSeq platforms provide high quality reads for a short region of only around 400-bp and therefore the barcodes used for metabarcoding with this technology must be correspondingly short. The two main markers used for diatom metabarcoding studies are the V4 region of the nuclear 18S rRNA gene and a region within the plastid *rbcL* gene, both regions being circa 300-400 bp long (including primers). The *rbcL* marker is more often used, partly because it was designed specifically for diatoms, and because it is better covered by Diat.barcode (Rimet et al., 2019), which is the most complete and curated reference library available for diatom metabarcoding to date. Furthermore, overall *rbcL* gives better discrimination between closely related species than 18S rDNA (e.g. Evans et al. 2007, p. 357; Urbánková & Veselá, 2013). Consequently, better and more confident taxonomic resolution can be achieved when using *rbcL* compared to 18S rDNA (Apothéloz-Perret-Gentil et al. 2021; Bailet et al., 2020).

In this context, two similar barcodes of the *rbcL* gene have been developed independently by different research groups for diatom metabarcoding. One of these barcodes covers a region of 263-bp and is amplified by the primer pair Diat_rbcL_708F (Stoof-Leichsenring et al., 2012) and R3 (Bruder and Medlin 2007). These primers were further degenerated by Vasselon et al. (2017), in order to cover a wider diversity of diatoms, resulting in three forward primers (Diat_rbcL_708F1,

Diat_rbcL_708F2 and Diat_rbcL_708F3) and two reverse primers (R3_1 and R3_2). The second barcode includes the same 263-bp region as the previous one but has an extra tail of 68-bp located at the 5′ end. This latter, developed by Kelly et al. (2018, 2020), therefore comprises 331 bp and is amplified by the primer pair rbcL-646F and rbcL-998R. Thus, although both barcodes overlap in the shared region of 263-bp, they could potentially differ in their ability to discriminate between species, which would be relevant for biodiversity analyses but also for the monitoring and management of freshwater rivers covered by the Water Framework Directive (WFD), since the diatom indices computed for such purposes, such as the Indice de Polluosensibilité Spécifique (IPS; Cemagref, 1982), rely on species composition and relative abundance. Both barcodes (hereafter referred to as the 263- and 331-bp markers) have been demonstrated to be effective for biomonitoring and diversity analyses (e.g. Kang et al., 2021; Kelly et al., 2018, 2020; Rimet et al., 2018b; Rivera et al., 2020). Nevertheless, we might hypothesize that the 68-bp tail might confer an advantage for species assignment in two ways. On the one hand, it might be possible that related species are identical in the 263-bp shared region but differ at variable sites in the extra 5′ tail. On the other hand, the accuracy of some automated methods commonly applied for classifying metabarcoding data increases as the length of the query sequence increases (Porter et al., 2014; Karim & Abid, 2021). In this regard, it might be expected that use of the longer (331-bp) barcode could increase the effectiveness of the Naïve Bayesian classifier (Wang et al. 2007), a Kmer-based method that is one of the most commonly implemented classifiers for assigning reads to named taxa in metabarcoding studies.

These two aspects have not yet, to our knowledge, been explored for the two similar diatom *rbcL* markers. Therefore, this study aimed to (1) compare the effect of choosing one or the other marker on WFD ecological assessments through the comparison of IPS scores: is there any significant advantage in using the longer marker? (2) assess the effectiveness of the two markers for discriminating genetic variants at or below the species level. For achieving these aims, we used a

large dataset of environmental samples collected during several biomonitoring campaigns in UK rivers (Kelly et al. 2018, 2020).

2. Material and Methods

2.1 Dataset and bioinformatics analyses

The dataset used in this study comprised 1703 benthic diatom samples that were originally taken as part of routine WFD biomonitoring programmes of UK rivers held in 2014, 2016 and 2017 (Kelly et al., 2018, 2020). High-throughput sequencing (HTS) of these samples was based on the 331-bp *rbcL* marker amplified by the rbcL-646F and rbcL-998R primers, and we were supplied with the fastq files from MiSeq output. Further details about the preparation of samples for HTS are described in Kelly et al. (2018, 2020). We conducted bioinformatics analyses on the forward (R1) and reverse (R2) reads to generate the Amplicon Sequence Variants (ASVs) that constituted the fundamental units on which further examinations were carried out. ASVs were generated using the R package *DADA2* (Callahan et al., 2016) and the different runs (a total of 10) were analysed separately. The rbcL-646F and rbcL-998R primers were removed from R1 and R2 reads using cutadapt (Martin, 2011). Then, the R1 and R2 reads were truncated to 220–240 and 160–180 nucleotides respectively, based on their quality profiles (median quality score < 30), and those reads with ambiguities or showing an expected error (maxEE) higher than 2 were removed. The DADA2 denoising algorithm was then applied to determine an error rates model in order to infer amplicon sequence variants (ASVs). Finally, ASVs detected as chimeras were discarded using the DADA2 function "removeBimeraDenovo". Since the ASVs generated were based on the 331-bp *rbcL* marker, they also contained the 263-bp region targeted by the three forward primers Diat_rbcL_708F1, Diat_rbcL_708F2 and Diat_rbcL_708F3 and the two reverse primers R3_1 and R3_2. To avoid any incongruence during the comparative analyses of the two markers, the only ASVs selected for further analyses were those in which the forward primers Diat_rbcL_708F1,

Diat_rbcL_708F2 or Diat_rbcL_708F3 were also identified. For this, cutadapt was applied again, this time on the 331-bp ASVs already generated, to unambiguously identify and remove these primers specifically designed for the 263-bp marker. Thus, two datasets with the same number of ASVs were finally generated, one containing ASVs with a total length of 331-bp (i.e. those based on the rbcL-646F and rbcL-998R primers) and a second one including the same ASVs but truncated to a length of 263-bp.

We emphasize here that this was not a study based on laboratory application of the two sets of primers to the same samples. This would be interesting and, as far as we know, has never been undertaken, but it would introduce extra variables whose effects we did not set out to determine. The first is clearly that the forward primers of the two markers are very unlikely to be exactly equivalent in their selectivity. For example, judging by the spread of ochrophyte, rhodophyte and chlorophyte taxa represented in 331-bp and 263-bp datasets (the UK dataset analysed here and the French–Catalan datasets of Rivera et al. 2020 and Pérez-Burillo et al. 2021), the 331-bp primers are less specific for diatoms than the 263-bp primers (our unpublished data). Furthermore, although the region amplified by the two markers have the same 3′ terminus, the reverse primers also differ: the R3_1/R3_2 and rbcL-998R primers differ in length (R3_1/R3_2 = 22bp; rbcL-998R = 27bp) and in the degree of degeneration (R3_1 and R3_2 both include one more degenerate base than rbcL-998R). It is therefore quite possible that there would be different primer biases during amplification from the same pool of diatoms. Our study was only to investigate the extent to which the extra 5′ tail provides extra taxonomic resolution for biodiversity assessment and has any implications for the WFD assessments.

2.2 Reference library preparation and taxonomic assignment.

A custom-made reference library composed of 331-bp sequences was used for performing the taxonomic assignment of the ASVs generated. By controlling the reference sequence length (rather than using reference sequences that have not been trimmed to the same length), it is easier to

evaluate how the different marker lengths are affecting the taxonomic assignment. The custom-made library consisted of all the sequences from the curated diatom reference library Diat.barcode v10 (Rimet et al., 2019) that cover the full 331-bp *rbcL* marker. It was created by extracting a subset of diatom *rbcL* sequences (a total of 2807 sequences) from Diat.barcode v10 that covered the 331-bp marker, aligning them (using MUSCLE: Edgar, 2004), and truncating them to the target 331-bp region using MegaX (Kumar et al., 2018). Then, all the remaining *rbcL* diatom sequences included in Diat.barcode v10 were extracted and aligned against the aligned subset using the *align.seqs* function implemented in Mothur software (Schloss et al., 2009), with default parameters. The resulting alignment of 331-bp diatom sequences was further filtered with Mothur (using the *screen. seqs* function*)* to keep only sequences without ambiguities. The taxonomic assignment of 263-bp and 331-bp ASVs was performed using two methods: 1) the naïve Bayesian classifier method (Wang et al., 2007) using the "assignTaxonomy" function from *DADA2* and 2) the Basic Local Alignment Search Tool (BLAST). Prior to the next analyses, and in order to remove non-diatom variants that likely occurred in our dataset, only ASVs classified into Bacillariophyta and receiving 100% bootstrap support (i.e. the percentage of times that an ASV is assigned by the classifier to the same taxon) by the Bayesian classifier were kept for downstream analyses. As a result, a total of 2933 ASVs were used in this study.

2.3 Comparative analyses between the 331-bp and 263-bp markers

The effect of marker choice on taxonomic assignment of ASVs was assessed by comparing the number of 263-bp and 331-bp ASVs that had an identical match (considered here as a pairwise-alignment with 100% similarity, no gaps and mismatches, and a full cover of the query sequence) with reference sequences from Diat.barcode v10. Out of the ASVs with identical matches, we determined the number of fully identified species to which each ASV was identical. In addition, the number of 263-bp and 331-bp ASVs assigned at species level by the naïve Bayesian classifier was compared through different bootstrap support values (i.e. above 60%, above 85% and above 99%)

The ecological status of each sample was determined by applying the IPS diatom index, since this is adopted in many EU countries for WFD bioassessment of rivers. For each sample, the IPS was calculated twice, one using the species inventory derived from the 263-bp ASVs, and the other using the inventory from the 331-bp ASVs. IPSS and IPSV values for each species were extracted from OMNIDIA software v5.5 (Lecointe et al., 1993). Comparisons of the IPS values were performed using ASVs that had a species assignment bootstrap value ≥ 85%, since thresholds from 80% to 85% are commonly applied for diatom biomonitoring assessments (e.g. Rivera et al., 2020; Mortágua et al., 2019; Vasselon et al., 2017). The WFD ecological status class for each sample was assigned by applying the following boundaries (Afnor, 2007): High ($17 \leq IPS \leq 20$), Good ($13 \leq IPS < 17$), Moderate ($9 \leq IPS < 13$), Poor ($5 \leq IPS < 9$), Bad ($1 \leq IPS < 5$).

2.4 In-depth analyses on species discrepancies

Samples that differed in absolute IPS values regarding the type of marker were further evaluated in order to elucidate the causes that led to these dissimilarities in the index. For this, we examined the species showing the greatest dissimilarities in relative abundance between marker datasets. To do this, we compared the taxonomic assignments and bootstrap support values provided by the naïve Bayesian classifier, as well as the most similar sequences and species determined by BLAST. In order to guarantee that the most similar sequences to each ASV were not excluded during any of the steps involved in the building of the custom reference library, BLAST analyses were also executed comparing ASVs against all the sequences included in Diat.barcode v10. Haplotype networks based on the TCS algorithm (Clement et al. 2002) were constructed in the most important cases where the taxonomic assignment of ASVs varied according to the choice of marker. The ASVs included in the network analyses were those that were recorded with at least 10 reads and occurred in more than 1 sample. A quick check for residual errors was made by examining the ASV alignment for stop codons: only one was found (ASV3000), occurring in 2 samples with 300 reads. Haplotype networks were performed and visualized using PopART software (Leigh and Bryant, 2015).

188 2.5 Shannon entropy comparisons between 331-bp and 263-bp markers

189 In order to compare and illustrate the nucleotide and amino-acid variability of the extra 68-bp
190 region provided by the 331-bp marker, Shannon's entropy values were calculated from both the
191 reference sequences from the 331-bp custom reference library and the 331-bp ASVs obtained.
192 Before calculating Shannon entropy values on ASVs, several filter steps were applied in order to
193 remove likely artefacts. For this, only ASVs with 331-bp length were kept and those showing an
194 abundance lower than 10 reads and/or occurring in only 1 sample were also removed. The resulting
195 ASVs were aligned against the custom 331-bp reference library and those with gaps and/or stop
196 codons were further discarded. In addition, duplicated sequences from the custom reference library
197 (i.e. sharing the 331-bp marker) were removed. Shannon entropy was thus calculated on a total of
198 2617 ASVs and 1886 reference sequences. Entropy values were computed using the
199 "MolecularEntropy" function implemented in the R package *HDMD* (McFerrin, 2013) and the
200 values were standardized to 4 and 20 for nucleotides and amino acids respectively, as these figures
201 represent the number of possible states in a DNA or protein sequence.

202 **3. Results**

203 3.1 Effects of the marker on taxonomic assignment

204 The number of ASVs assigned at the species level by the naïve bayesian classifier was always
205 higher when using the longer marker, regardless of the bootstrap confidence threshold applied
206 (Table 1). On the other hand, BLAST analyses indicated that for the 263-bp marker, a total of 536
207 different ASVs (18.3%) had at least one identical match (identical matches considered only when
208 query ASV sequences were fully covered) with reference sequences included in Diat.barcode while
209 this number was reduced to 426 ASVs (14.5%) when considering the full 331-bp marker. In
210 addition, 29 ASVs based on the 331-bp marker were identical to reference sequences from more
211 than 1 species and these ambiguous assignments corresponded to a total of 62 different species but
212 to a total of 74 species when considering only the 263-bp marker (Supplementary Table 1). These

ambiguous assignations at the species level were exemplified, among others, in some ASVs classified into the genera *Fragilaria* (ASVs 59, 131 and 346; Fig. 4), *Iconella* (ASVs 270 and 361), *Surirella* (ASV 26; Fig. 3) and *Gomphonema* (ASVs 6, 148, 216, 274 and 610) (Supplementary Table 1).

3.2 Effects of the marker choice on ecological status assessment

IPS values calculated from both markers were very similar and strongly correlated (Pearson's R = 0.98) (Fig. 1). 1621 sites (95.2%) shared the same ecological status class with both markers and only 82 (4.8%) showed 1 class of difference; none of the sites showed more than 1 class of difference. Out of the 82 sites with 1 class of difference, 57 corresponded to absolute differences in the IPS scores that were < 1 and 25 to absolute differences in IPS scores > 1. The total numbers of sites classified into "Moderate", "Poor" or "Bad" status (i.e. unacceptable classes for WFD) were 388 (22.82%) and 371 (21.79%) for the 263-bp and 331-bp markers respectively. In addition, a total of 55 sites (3.23% of the 1703 sites analysed) were downgraded from "Good" ecological status when using one marker to "Moderate" status when using the other.

3.3 Effects of the marker choice on species abundance and taxonomic resolution

The species showing the greatest dissimilarities in relative abundance between markers are listed in Fig. 2. Examination of bootstrap support values and BLAST outputs for both 263-bp and 331-bp ASVs of these species revealed there are three main reasons for the abundance dissimilarities:

i) False negatives: Some ASVs were classified into the same species by both the 263-bp and 331-bp markers but the identifications could be rejected for one or other marker because bootstrap support values did not reach the confidence threshold (i.e. bootstrap values ≥ 85), ultimately causing differences between markers in species' relative

abundance. Some false negatives arose when the assignments of 263-bp ASVs received much lower bootstrap support values than their 331-bp counterparts. This occurred when the genetic distance between ASVs and closely related reference sequences (as measured by the number of base-pair mismatches between ASVs and reference sequences reported by BLAST analyses) decreased when using the shorter marker compared to the longer one. In this regard, the most important cases were detected in ASVs from the *Achnanthidium minutissimum* complex (observed in ASVs closely related to *A. jackii* and *A. pyrenaicum*, such as ASV909, ASV1420, ASV7083), *Nitzchia perminuta* (detected in ASVs assigned to this species but similar also to *N. acidoclinata*, for instance, ASV2288), *Encyonema ventricosum* (ASVs also similar to *E. minutum*, such as ASV929), *Diatoma moniliformis* (ASVs also similar to *D. tenuis*, e.g. ASV73, ASV403 and ASV1159) or *Navicula rostellata* (ASV200 and ASV721, two ASVs similar to reference sequences classified as *Navicula* sp. and *Haslea howeana*) (Supplementary Data 1 & 2). By contrast, other false negatives were detected with no increase in genetic distance between ASVs and closely related reference sequences. This was particularly evident in ASV33 and ASV136, two abundant ASVs belonging to *Cocconeis euglypta* and *Gomphonema affine* respectively (Supplementary Data 1 & 2)

ii) Some ASVs were unambiguously classified at the species level based on the 331-bp marker, but not based on the 263-bp marker. This was seen in ASVs in *Surirella* (ASV17), *Fragilaria* (ASV140) and *Halamphora* (ASV1784). Within *Surirella*, ASV17 had identical matches with reference sequences from *Surirella brebissonii* (including *S. brebissonii* var. *kuetzingii*) when the ASV was based on the 331-bp marker and could therefore be identified unambiguously. The effect of reducing the barcode marker to the 263-bp region was to make ASV17 identical to reference sequences belonging to 10 different taxa (i.e. *Surirella angusta, Surirella sp., S. cf. pinnata, S. brightwellii, S. ovalis var. apiculata, S. cf. minuta, S. minuta,* and *S. lacrimula*, as well as the two that

263 are identical over the whole of the 331-bp marker, *Surirella brebissonii* and *Surirella*
264 *brebissonii var. kuetzingii*). A haplotype network for these and other *Surirella* species
265 and related ASVs is given in Fig. 3 and shows the changes in assignment and
266 relationships when the marker length is reduced from 331 bp (Fig. 3a) to 263 bp (Fig.
267 3b). In the case of *Fragilaria* species, ASV140 matched only one species (*F. agnesiae*)
268 based on the 331-bp marker (Fig. 4a), but was identical to three species, *Fragilaria*
269 *agnesiae*, *Fragilaria* sp. and *Fragilaria* cf. *nanoides*, with the 263-bp marker (Fig. 4b).
270 A third case (not graphed) was ASV1784, which shared the full 263-bp marker with
271 reference sequences from *Halamphora montana* and *Halamphora banzuensis* species
272 but differed from the latter by two mutations located at the 30th and 34th positions of the
273 331-bp marker.
274 iii) A third group comprised ASVs that could not be identified to species with either marker:
275 they were identical to reference sequences from more than one taxon for both the 263-
276 and the 331-bp marker. In these cases, differences in species' relative abundance
277 between markers occurred when the taxonomic classification provided by one marker
278 did not reach the selected confident threshold (i.e. bootstrap values ≥ 85) but this
279 threshold was reached when using the other marker. This pattern is likely associated
280 with the random component of the naïve Bayesian classifier and it was observed in
281 ASVs classified into the genera and *Achnanthidium* (ASV12) and *Iconella* (ASV 361)
282 (Supplementary Data 3).

283

284 A more complex and particularly instructive case illustrating the potential complexities of
285 interpretating the metabarcoding data, is given by *Nitzschia* ASVs 1690 and 3022. These two
286 haplotypes shared the full 263-bp marker with reference sequences from *Nitzschia dissipata* var.
287 *media* and *N. heufleriana*, respectively, and therefore seemed securely identified, ASV 3022 as *N.*

dissipata var. media and ASV 1690 as *N. heufleriana* (Fig. 5b). However, when considering the full 331-bp marker these ASVs were not identical to the same two reference sequences and had no exact match in the reference dataset. Instead, each of them differed by 1 nucleotide from both *N. dissipata* var. *media* and *N. heufleriana*, making identification impossible at species level (Fig 5a).

3.4 Nucleotide and amino-acid variability.

In order to provide context for the differences in species discrimination between the 311- and 263-bp markers, we calculated Shannon entropy values at each site within the marker region (there were no indels: as far as we know, all river diatom taxa sequenced so far have the same length *rbcL*). The average Shannon entropy values for nucleotides and amino acids indicated that the maximum variability of the barcode markers takes place in the 263-bp shared region, although overall the average entropy values for the extra 68 bp at the 5′ end region of the 311-bp marker were very similar to those in the shared 263-bp region (Fig. 6; Table 2). The average entropy values of the full 331-bp marker for both nucleotides and amino acids were slightly higher in ASVs than in the reference sequences (Table 2).

Discussion

4.1.1 The choice of *rbcL* marker does not have major implications for diatom-based WFD ecological assessment of rivers

The extra length of the 331 bp marker means that it inevitably provides more information on genetic diversity, given the variability of the extra 68-bp tail (Fig. 6). Our results indicate, however, that the choice of the 263-bp or 331-bp *rbcL* marker has no important effects on WFD ecological status assessments, since IPS scores derived from both markers were very highly correlated (i.e. Pearson's R = 0.98 and intercept close to 0) and the vast majority of sites were classified into the same ecological status class regardless of the marker used (i.e. 95.2%). In addition, out of the sites that differed in the ecological status assignment, most of them correspond to absolute deviations in the

IPS scores of < 1. However, the overall number of sites classified into "Moderate", "Poor" and "Bad" status differed with the marker chosen, and this number was higher when using the 263-bp one. As a consequence, some particular sites were assigned to the "Good" ecological status when using one marker, but they were assigned instead to the "Moderate" status when using the other (observed in a total of 55 out of 1703 samples studied). Though the proportion of such samples is very low, they should not be overlooked since the WFD demands remedial actions for those aquatic systems that fail to reach at least "good" ecological status.

At first, it might be interpreted that the discrepancies in IPS values for those sites that alter their ecological status from acceptable (i.e. "Good") to unacceptable ("Moderate") classes are brought about by differences in species' relative abundances caused by the higher taxonomic resolution of the 331-bp marker (i.e. the 331-bp marker can unambiguously classify some ASVs at the species level that 263-bp marker cannot). However, our results indicated that the choice of the marker was decisive for discriminating taxa at species level in only three ASVs (discussed further in section 4.2) and more importantly, these ASVs were scarcely represented in most of the samples: only ASV17 (*Surirella brebissonii*) contributed at least 10% of reads' relative abundance in 7 samples (supplementary Data 4). Thus, most of the discrepancies observed between markers in species' relative abundance, and hence in WFD ecological status assignations, cannot be attributed to differences in taxonomic resolution between markers. Instead they are likely due to other factors such as the stochasticity involved in the Bayesian classifier (Wang et al., 2007) and false negatives. In this regard, our results showed that the use of the extra 68-bp region can reduce the number of false negatives by increasing the genetic distance between ASVs and closely related taxa and therefore if initiating a new metabarcoding study, the 331-bp marker could be preferable.

4.2. In a few cases the choice of marker is decisive for discriminating certain taxa at species level

For some freshwater diatom species the choice of the marker is crucial for discriminating at the species level and hence may materially alter conclusions when the focus is on aspects of

biodiversity, such as species distributions and ecology, rather than on biomonitoring. In our dataset, this was observed in three ASVs from the species *S. brebissonii* (ASV17), *H. montana* (ASV1784) and *F. agnesiae* (ASV140). Because of its relatively high abundance and occurrence, ASV17 is the most important example. It was successfully classified at the species level when using the full 331-bp marker (an identical match to *S. brebissonii*) whereas the 263-bp shared region of this ASV was also identical to several other *Surirella* species from the Pinnatae group. Species of the Pinnatae group are characterized by close phylogenetic relationships reflected in small interspecific genetic differences, not only in *rbcL* but also in other molecular markers (Ruck et al., 2016), and morphological separation of *S. brebissonii* from other species of this group is difficult (morphometric characteristics overlap between species: English & Potapova, 2012; Krammer & Lange-Bertalot, 1987). In this case, differentiating species could even be relevant for biomonitoring, because *S. brebissonii* can dominate diatom assemblages (for instance, in some German rivers: Lange-Bertalot et al., 2017) and differs in IPSS and IPSV values from some other species of the Pinnatae group, (*S. brebissonii* and *S. lacrimula* have IPSS=3 and IPSV=2, whereas all *S. angusta* and *S. ovalis* var. *apiculata* have IPSS=4 and IPSV=1, and *S. brightwellii* has IPSS=2 and IPSV=3).

Other cases where the 331 bp marker is decisive for species identification include *Halamphora montana* vs *H. banzuensis* (ASV1784), two species with very different habitat requirements. *H. montana* occurs in intermittently wet terrestrial microhabitats and eutrophic freshwaters (Lange-Bertalot et al., 2017) and is characterized by intermediate IPS sensitivity values (IPSS=2.9). In contrast, *H. banzuensis* is a marine species (recently described by Stepanek, & Kociolek, 2018) and hence has no associated IPS indicator values. The little variation found between both 263-bp and 331-bp *rbcL* markers for these species is not exceptional within *Halamphora*, as other examples of close phylogenetic relationships between freshwater and marine species can be found within the genus (Stepanek & Kociolek, 2019). Similarly, *F. agnesiae* (ASV140) cannot be identified using

361 the 263-bp marker, but in this case the effects are unclear: *F. agnesiae* is a recently described
362 species without a full ecological characterization (Kahlert et al., 2019).

363 In all these cases, therefore, there is a clear benefit in using the longer marker and this will no doubt
364 also be true in many other diatoms where there are currently few or no reference sequences (for a
365 number of genera, such as *Brachysira*, and more generally for oligotrophic freshwater taxa and
366 marine littoral diatoms, there is especially poor coverage in the reference database).

367 4.3. A small proportion even of the 331-bp *rbcL* variants cannot be unambiguously classified at the
368 species level

369 We identified a total of 29 ASVs for which the full 331-bp marker was identical to reference
370 sequences from more than one species and therefore neither of the two barcode markers would
371 assign the haplotype unambiguously at the species level. These cases reflect the lack of a barcode
372 gap even for the full 331-bp *rbcL* marker and indicate that, without a complete reference database, it
373 is impossible to determine in many cases whether the diversity of ASVs represents intraspecific
374 diversity or the presence of separate but currently undescribed species. Thus, as noted in the
375 previous section, for studying aspects related to the diversity, ecology and biogeography of certain
376 species, as opposed to practical WFD biomonitoring, current *rbcL* metabarcoding has clear
377 limitations.

378 Overall, the 331-bp marker is superior in that the diversity that can be detected is greater and the
379 proportion of ambiguous identifications is lower. Sometimes too, an apparently straightforward
380 identification with the shorter marker is deceptive. Particularly instructive in this regard is the
381 example of *Nitzschia* ASVs 1690 and 3022, which seem to be identifiable confidently and indeed
382 unambiguously with the 263-bp marker (100% matches with *N. dissipata* var. *media* and *N.*
383 *heufleriana* reference sequences, respectively) but not with the 331-bp marker: the two ASVs
384 cannot be identified from the 331-bp versions since they are not identical to either of the reference
385 sequences that are available but separated from each of them by the same genetic distance. In this

386 case, to interpret the metabarcoding datasets fully in terms of nominal species and varieties, much
387 more information would be needed about the correspondence between *rbcL* variation and
388 morphology.

389 To conclude, some species cannot be assigned at the species level even when using the longer
390 marker and it is unrealistic to expect that the reference library will be able to cover all the existing
391 genetic variants in the near future. This is because the process of obtaining new Sanger sequences
392 and curating barcodes (Rimet et al., 2019) is laborious and expensive, and determining which ASVs
393 belong to which species from the metabarcoding dataset alone can be done only in special
394 circumstances (e.g. when a species is particularly abundant in samples for which matching DNA
395 and microscopical data are available: Rimet et al., 2018a). Nevertheless, the far greater number of
396 ASVs in the UK dataset, relative to microscopically separable species, and the low proportion of
397 ambiguous assignments made in our study of a very extensive dataset (i.e. 29 ASVs out of 2933 in a
398 total of 1703 benthic samples) shows that DNA metabarcoding of short *rbcL* markers is a very
399 effective method for surveying diatom biodiversity at the species level in aquatic systems. The
400 arrival of long-read sequencing platforms (e.g. Pacific Bioscience or Oxford Nanopore
401 Technologies), with reliable sequencing lengths far above 1200–1500 bp (the lengths of 'full'
402 diatom *rbcL* sequences in GenBank) will further improve resolution.

403

404 4.4. Both markers capture high genetic diversity within and between nominal diatom species, which
405 can be important for ecological understanding

406 Most of the genetic variants examined were not represented in the reference library: out of the 2933
407 ASVs separated by the 331-bp marker, identical matches with reference sequences were found for
408 only 426 (14.5%) and 536 ASVs (18.3%) respectively for the 331- and 263-bp markers. To some
409 extent, this is because of the lack of reference sequences for many nominal species, but it also
410 reflects the high intraspecific diversity that characterizes diatom species, at least as these are

411 currently circumscribed (e.g. Amato et al., 2007; Perez-Burillo et al. 2021; Pinseel et al., 2017;
412 Souffreau et al., 2013). The question that arises is whether the intraspecific diversity detected by the
413 two *rbcL* markers is only 'genetic noise', or whether it contains information on ecological or
414 biogeographical differentiation and therefore needs to be recorded and analysed. First indications
415 are that, while closely related species often share a similar ecology (Keck et al., 2018), closely
416 related ASVs can differ in ecological preferences and distribution (Pérez-Burillo et al., 2021).
417 Therefore, while it will always be important to relate the ASVs of metabarcoding datasets to formal
418 morphology-based taxonomy – e.g. to ensure continuity with previous studies and allow cross-talk
419 with fields where DNA-based approaches are limited in their application (e.g. stratigraphical or
420 palaeoecological studies) – degrading analysis to the level of nominal species is suboptimal. For
421 example, from a biomonitoring perspective it will mean that diatom indexes are being computed
422 using only a part of the information from the total captured, especially when strict confidence
423 thresholds are applied. In particular, we found that around 70% of the ASVs were not assigned to a
424 species by the naïve Bayesian classifier when the confidence threshold was ≥ 99%. Hence an
425 attractive alternative to the present approach, if environmental data are available for an extensive set
426 of metabarcoded samples, is a direct calibration of the environmental preferences of ASVs or
427 OTUs, as suggested by other studies (e.g. Apothéloz-Perret-Gentil et al., 2017; Feio et al., 2020;
428 Smucker et al., 2020; Tapolczai et al., 2019). Microscopy-based approaches remain important,
429 however, since they give opportunities to study traits that are not or only partially taxon-related,
430 such as life-history stage and teratological forms (Falasco et al. 2021) or, in the case of some marine
431 and freshwater diatoms, existence as endosymbionts (Pérez-Burillo et al., 2022; Takano et al.,
432 2007).

433

434

435

Conclusions

The main goal of this study was to analyse the effect of using two similar and short *rbcL* diatom markers for biomonitoring programmes. Our results show that the choice of marker does not have major implications for WFD ecological assessments. Our second objective was to study the effect of marker choice on species resolution. We found that for some taxa, the use of the larger 331-bp marker allows resolution at species level or leads to a reduction in the number of ambiguous assignations (i.e. ASVs identical to reference sequences from more than one species), compared to the shorter 263-bp *rbcL* marker, reflecting the fact that the extra 5′ tail of the 331-bp marker is quite variable (approximately as much so as the average of the 263-bp marker). The higher resolution of the longer marker may therefore be preferable in ecological or biogeographical studies, especially with increasing demonstrations that closely related lineages, previously included within the same (morpho-)species can differ in their distributions and ecological preferences.

Acknowledgements

We especially thank Dr Kerry Walsh (UK Environment Agency) for making the UK metabarcoding datasets available to us and for her encouragement to use them. J. Pérez-Burillo acknowledges IRTA and Universitat Rovira i Virgili for his Martí Franqués PhD grant (2018PMF-PIPF-22). The Royal Botanic Garden Edinburgh is supported by the Scottish Government's Rural and Environment Science and Analytical Services Division. We also acknowledge support from the CERCA Programme/Generalitat de Catalunya. We thank the three anonymous reviewers for their very constructive comments which helped to improve the paper.

References

Afnor, N. F.,2007. T90-354. Qualite de l'eau. Determination de l'Indice Biologique Diatomees (IBD). Afnor, 1-79.

Amato, A., Kooistra, W.H.C.F., Levialdi Ghiron, J.H., Mann, D.G., Pröschold, T., Montresor, M., 2007. Reproductive isolation among sympatric cryptic species in marine diatoms. Protist. 158, 193 207. https://doi.org/10.1016/j.protis.2006.10.001

Apothéloz-Perret-Gentil, L., Bouchez, A., Cordier, T., Cordonier, A,. Guéguen, J., Rimet, F., Vasselon, V., Pawlowski, J., 2021. Monitoring the ecological status of rivers with diatom eDNA metabarcoding: A comparison of taxonomic markers and analytical approaches for the inference of a molecular diatom index. Mol Ecol. 30, 2959-2968. https://doi.org/10.1111/mec.15646,

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. https://doi.org/10.1111/1755-0998.12668.

Bailet, B., Apothéloz-Perret-Gentil, L., Baricevic, A., Chonova, T., Franc, A., Frigerio, J.-M., Kelly, M., Mora, D., Pfannkuchen, M., Proft, S., Ramon, M., Vasselon, V., Zimmermann, J., Kahlert, M., 2020. Diatom DNA metabarcoding for ecological assessment: comparison among bioinformatics pipelines used in six European countries reveals the need for standardization. Sci. Total Environ. 745, 140948. https://doi.org/10.1016/j.scitotenv.2020.140948.

Bruder, K., Medlin, L.K., 2007. Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms (Bacillariophyta). I. The genus *Placoneis*. Nova Hedwigia. 85, 331-352

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high resolution sample inference from illumina amplicon data. Nat. Methods. 13, 581-583. https://doi.org/10.1038/nmeth.3869.

Cemagref, A., 1982. Étude des méthodes biologiques quantitative d'appréciation de la qualité des eaux. Bassin Rhône-Méditerranée-Corse. Centre National du Machinisme Agricole, du Génie rural, des Eaux et des Forêts, Lyon, France.

Clement, M., Snell, Q., Walker, P., Posada, D., Crandall, K., 2002. TCS: estimating gene genealogies. In Proceedings of the 16th International Parallel and Distributed Processing Symposium, p.184.

De Luca, D., Piredda, R., Sarno, D., Kooistra,W.H.C.F., 2021. Resolving cryptic species complexes in marine protists: phylogenetic haplotype networks meet global DNA metabarcoding datasets. ISME J. 15, 1931-1942. https://doi.org/10.1038/s41396-021-00895-0.

Edgar R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids Res. 32, 1792-1797. https://doi.org/10.1093/nar/gkh340

English, J. D., Potapova, M. G., 2012. Ontogenetic and interspecific valve shape variation in the Pinnatae group of the genus *Surirella* and the description of *S. lacrimula* sp. nov. Diatom Res. 27, 9-27. https://doi.org/10.1080/0269249X.2011.642950

Evans, K. M., Wortley, A. H., Mann, D. G., 2007. An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). Protist. 158, 349-364. https://doi.org/10.1016/j.protis.2007.04.001.

Falasco, E., Ector, L., Wetzel, C.E., Badino, G. & Bona, F. 2021. Looking back, looking forward: a review of the new literature on diatom teratological forms (2010–2020). Hydrobiologia 848, 1675–1753. https://doi.org/10.1007/s10750-021-04540-x

Feio, M. J., Serra, S. R., Mortágua, A., Bouchez, A., Rimet, F., Vasselon, V., Almeida, S. F., 2020. A taxonomy-free approach based on machine learning to assess the quality of rivers with diatoms. Sci. Total Environ. 722, 137900. https://doi.org/10.1016/j.scitotenv.2020.137900.

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. https://doi.org/10.1111/jpy.12886

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverría Galindo, P., Vences, M., Wang, J., Schwalba, A., 2021. Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam Co, Tibet: the effect of sample-size and bioinformatics on the identified communities. Ecol. Indic. 121, 107070. https://doi.org/10.1016/j.ecolind.2020.107070.

Karim, M., Abid, R., 2021. Efficacy and accuracy responses of DNA mini-barcodes in species identification under a supervised machine learning approach. 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). 1-9. 10.1109/CIBCB49929.2021.9562838

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. https://doi.org/10.1111/1755-0998.12919.

Kelly, M., Boonham, N., Juggins, S., Kille, P., Mann, D.G., Pass, D., Sapp, M., Sato, S., Glover, R., 2018. A DNA Based Diatom Metabarcoding Approach forWater Framework Directive Classification of Rivers. Environment Agency. https://assets.publishing.service.-gov.uk/government/uploads/system/uploads/attachment_data/file/684493/A_DNA_based_metabarcoding_approach_to_assess_diatom_communities_in_rivers_-_report.pdf.

Kelly, M.G., Juggins, S., Mann, D.G., Sato, S., Glover, R., Boonham, N., Sapp, M., Lewis, E., Hany, U., Kille, P., Jones, T., Walsh, K., 2020. Development of a novel metric for evaluating diatom assemblages in rivers using DNA metabarcoding. Ecol. Indic. 118, 106725. https://doi.org/10.1016/j.ecolind.2020.106725.

Krammer K., Lange-Bertalot H., 1987. Morphology and taxonomy of *Surirella ovalis* and related taxa. Diatom Res. 2, 77-95. https://doi.org/10.1080/0269249X.1987.9704986

Kumar, S., Stecher, G., Li,M., Knyaz, C., Tamura, K., 2018. MEGA X:molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35, 1547-1549. https://doi.org/10.1093/molbev/msy096.

Lange-Bertalot, H., Hofmann, G., Werum, M., Cantonati, M., 2017. Freshwater Benthic Diatoms of Central Europe: Over 800 Common Species Used in Ecological Assessment. English Edition with Updated Taxonomy and Added Species. Koeltz Botanical Books, Schmitten-Oberreifenberg, pp. 1-942.

Lecointe, C., Coste, M., Prygiel, J., 1993. OMNIDIA—software for taxonomy, calculation of diatom indexes and inventories management. Hydrobiologia. 269, 509-513. https://doi.org/10.1007/BF00028048.

Leigh, J.W., Bryant, D., 2015. POPART: full-feature software for haplotype network construction. Methods Ecol Evol. 6, 1110-1106. https://doi.org/10.1111/2041-210X.12410

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17, 10-12. https://doi.org/10.14806/ej.17.1.200

McFerrin, L., 2013. HDMD: Statistical Analysis Tools for High Dimension Molecular Data (HDMD). R package version 1.2. https://CRAN.R-project.org/package=HDMD

Mortágua, A., Vasselon, V., Oliveira, R., Elias, C., Chardon, C., Bouchez, A., Rimet, F., João Feio, M., Almeida, S.F., 2019. Applicability of DNA metabarcoding approach in the bioassessment of Portuguese rivers using diatoms. Ecol. Indic. 106, https://doi.org/10.1016/j.ecolind.2019.105470.

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

589 Mediterranean rivers. Sci. Total Environ. 727, 138445
590 https://doi.org/10.1016/j.scitotenv.2020.138445.
591
592 Pérez-Burillo, J., Trobajo, R., Leira, M., Keck, F., Rimet, F., Sigró, J., Mann, D.G., 2021. DNA
593 metabarcoding reveals differences in distribution patterns and ecological preferences among
594 genetic variants within some key freshwater diatom species. Sci. Total Environ. 728, 149029
595 https://doi.org/10.1016/j.scitotenv.2021.149029
596
597 Pérez-Burillo, J., Valoti, G., Witkowski, A., Prado, P., Mann, D. G., Trobajo, R., 2022. Assessment
598 of marine benthic diatom communities: insights from a combined morphological–
599 metabarcoding approach in Mediterranean shallow coastal waters. Mar. Pollut. Bull. 174,
600 113183. https://doi.org/10.1016/j.marpolbul.2021.113183.
601
602 Pinseel, E., Vanormelingen, P., Hamilton, P.B., Vyverman, W., Van de Vijver, B., Kopalova, K.,
603 2017. Molecular and morphological characterization of the *Achnanthidium minutissimum*
604 complex (Bacillariophyta) in Petuniabukta (Spitsbergen, high Arctic) including the
605 description of *A. digitatum* sp. nov. Eur. J. Phycol. 52, 264-280.
606 https://doi.org/10.1080/09670262.2017.1283540.
607
608 Porter, T.M, Gibson, J.F., Shokralla, S., Baird, D.J., Golding, G.B., Hajibabaei, M., 2014. Rapid
609 and accurate taxonomic classification of insect (class Insecta) cytochrome coxidase subunit 1
610 (COI) DNA barcode sequences using a naïve Bayesian classifier. Mol. Ecol. Resour. 14,929-
611 942. https://doi.org/10.1111/1755-0998.12240.
612
613 Rimet, F., Gusev, E., Kahlert, M., Kelly, M.G., Kulikovskiy, M., Maltsev, Y., Mann, D.G.,
614 Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., Bouchez, A., 2019.

Diat.barcode, an open-access curated barcode library for diatoms. Sci.Rep. 9, 1-12. https://doi.org/10.1038/s41598-019-51500-6.

Rimet, F., Abarca, N., Bouchez, A., Kusber, W., Jahn, R., Kahlert, M., Keck, F., Kelly, M.G., Mann, D.G., Piuz, A., Trobajo, R., Tapolczai, K., Vasselon, V. AND Zimmermann, J., 2018a. The potential of High-Throughput Sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. Fottea 18, 37-54. doi: 10.5507/fot.2017.013

Rimet, F., Vasselon, V., A.-Keszte, B., Bouchez, A., 2018b. Do we similarly assess diversity with microscopy and high-throughput sequencing? Case of microalgae in lakes. Org. Divers. Evol. 18, 51-62. https://doi.org/10.1007/s13127-018-0359-5.

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. https://doi.org/10.1016/j.ecolind.2019.105775.

Ruck, E,C,, Nakov, T., Alverson, A.J., Theriot, E,C., 2016. Phylogeny, ecology,morphological evolution, and reclassification of the diatom orders Surirellalesand Rhopalodiales. Mol. Phylogenet. Evol. 103, 155-171. https://doi.org/10.1016/j.ympev.2016.07.023.

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, platformindependent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537-7541. https://doi.org/10.1128/AEM.01541-09.

Smucker, N. J., Pilgrim, E. M., Nietch, C. T., Darling, J. A., Johnson, B. R., 2020. DNA metabarcoding effectively quantifies diatom responses to nutrients in streams. Ecol Appl. 30, e02205. https://doi.org/10.1002/eap.2205.

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

Stepanek, J.G., Kociolek, J.P., 2018. *Amphora* and *Halamphora* from coastal and inland waters of the United States and Japan, with the description of 33 new species. Biblioth. Diatomol. 66,1-260

Stepanek, J.G., Kociolek, J.P., 2019. Molecular phylogeny of the diatom genera *Amphora* and *Halamphora* (Bacillariophyta) with a focus on morphological and ecological evolution. J. Phycol. 55, 442-456. https://doi.org/10.1111/jpy.12836.

Stoof-Leichsenring, K.R., Pestryakova, L.A., Epp, L.S., Herzschuh, U., 2020. Phylogenetic diversity and environment form assembly rules for Arctic diatom genera—a study on recent and ancient sedimentary DNA. J. Biogeogr. 47, 1166-1179. https://doi.org/10.1111/jbi.13786.

Stoof-Leichsenring, K.R., L.A., Epp, L.S., Tiedemann, R., 2012. Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. Mol. Ecol. 21, 1918-1930. https://doi.org/10.1111/j.1365-294X.2011.05412.x.

667 Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., Vasselon, V., 2019. Diatom DNA
668 metabarcoding for biomonitoring: strategies to avoid major taxonomical and bioinformatical
669 biases limiting molecular indices capacities. Front. Ecol. Evol. 7, 407
670 https://doi.org/10.3389/fevo.2019.00409.
671
672 Takano, Y., Hansen, G., Fujita, D., Horiguchi, T., 2007. Serial replacement of diatom
673 endosymbionts in two freshwater dinoflagellates, *Peridiniopsis* spp. (Peridiniales,
674 Dinophyceae). Phycologia, 47, 41-53. https://doi.org/10.2216/07-36.1.
675
676 Urbánková, P., Veselá, J., 2013. DNA-barcoding: A case study in the diatom genus *Frustulia*
677 (Bacillariophyceae). Nova Hedwigia. 142, 147-162.
678
679 Vasselon, V., Rimet, F., Tapolczai, K., Bouchez, A., 2017. Assessing ecological status with diatoms
680 DNA metabarcoding: scaling-up on a WFD monitoring network (Mayotte island, France).
681 Ecol. Indic. 82, 1-12. https://doi.org/10.1016/j.ecolind.2017.06.024
682
683 Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naïve bayesian classifier for rapid
684 assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol.
685 73, 5261-5267. https://doi.org/10.1128/AEM.00062-07.
686

Tables

Bootstrap support	≥60	≥70	≥80	≥90	≥99
263-bp marker	1937	1719	1489	1220	744
331-bp marker	2023	1786	1584	1316	888

Table 1. Comparison between the 263-bp and 331-bp markers in the number of ASVs assigned at the species level by the naïve Bayesian classifier through different bootstrapping support values (from 60 to 99).

Region	Shannon Entropy - Nucleotides		Shannon Entropy - Amino acids	
	Reference sequences	ASVs	Reference sequences	ASVs
5’ end 68-bp	0 – 0.62 (0.13±0.18)	0 – 0.58 (0.14±0.17)	0 – 0.24 (0.05±0.08)	0 – 0.26 (0.06±0.08)
Shared 263-bp	0 – 0.92 (0.17±0.22)	0 – 0.94 (0.17±0.22)	0 – 0.56 (0.07±0.11)	0 – 0.54 (0.08±0.11)
Full 331-bp	0 – 0.92 (0.16±0.21)	0 – 0.94 (0.17±0.22)	0 – 0.56 (0.06±0.10)	0 – 0.54 (0.07±0.10)

Table 2. Range, average and standard deviation of Shannon entropy values calculated on ASVs and Reference sequences in the different regions of the two *rbcL* markers surveyed; the 68-bp region located at the 5′ end of the 331-bp marker, the 263-bp region shared by both markers and the full 331-bp region.

Figures caption

Fig. 1 Correlation of IPS values derived from 263-bp and 331-bp markers considering the total 1703 samples analyzed. Pearson's coefficient (R) and p-value are given. Coloured squares represent boundaries for the different WFD ecological status classes: blue=high ($17 \leq IPS \leq 20$); green=good ($13 \leq IPS < 17$); yellow= moderate ($9 \leq IPS < 13$); orange= poor ($5 \leq IPS < 9$); red=bad ($1 \leq IPS < 5$).

Fig. 2. Top 15 species showing the greatest differences in relative abundance between 263-bp and 331-bp markers considering the total 1703 samples analyzed. Bars in red and blue represent species for which the greatest relative abundance was provided by the 263-bp and 331-bp respectively.

Fig. 3. TCS haplotype networks of *Surirella* species and closely related ASVs based on 331-bp (figure a) and 263-bp (figure b) rbcL markers. ASVs represented (as white circles) are those recorded with at least 10 reads in more than 1 sample, lack stop codons in their amino-acids composition and share at least 95% of similarity with reference sequences from the included *Surirella* species. Black circles represent hypothetical variants automatically inferred. Nodes represented by reference sequences for which identical ASVs were not found are indicated by an asterisk. Circles with dashed borders represent ASVs that differ in the 331-bp region but are identical in the 263-bp. Note that ASVs 17 and 26 have been represented in bold red and in a larger font to facilitate their visual identification in the network

Fig. 4. TCS haplotype networks of several *Fragilaria* species and closely related ASVs based on 331-bp (figure a) and 263-bp (figure b) rbcL markers. ASVs represented (as white circles) are those recorded with at least 10 reads in more than 1 sample, lack stop codons in their amino-acids composition and share at least 95% of similarity with reference sequences from the included *Fragilaria* species. Black circles represent hypothetical variants automatically inferred. Nodes represented by reference sequences for which identical ASVs were not found are indicated by an asterisk. Circles with dashed borders represent ASVs that differ in the 331-bp region but are identical in the 263-bp. Note that ASVs 59, 131, 140 and 346 have been represented in bold red and in a larger font to facilitate their visual identification in the network

Fig. 5. TCS haplotype networks of several *Nitzchia* species and closely related ASVs based on 331-bp (figure a) and 263-bp (figure b) rbcL markers. ASVs represented (as white circles) are those recorded with at least 10 reads in more than 1 sample, lack stop codons in their amino-acids composition and share at least 95% of similarity with reference sequences from the included *Nitzchia* species. Note that some *Nitzschia* ASVs met these criteria, but were removed for easier visualization of the networks. Black circles represent hypothetical variants automatically inferred. Circles with dashed borders represent ASVs that differ in the 331-bp region but are identical in the 263-bp. Note that ASVs 1690 and 3022 have been represented in bold red and in a larger font to facilitate their visual identification in the network.

Fig. 6 Shannon's entropy per nucleotide (figure a) and amino-acid (figure b) position obtained for 1886 reference sequences of 331-bp from Diat.barcode v10 (represented by a red line) and a total of 2617 ASVs obtained in this study (represented by a blue dashed line). ASVs included for computing entropy values were those that were recorded with at least 10 reads in more than 1 sample and did not show stop codons in their amino-acid composition. Entropy values have been standardized to 4 and 20 for nucleotides and amino acids respectively.

Figures

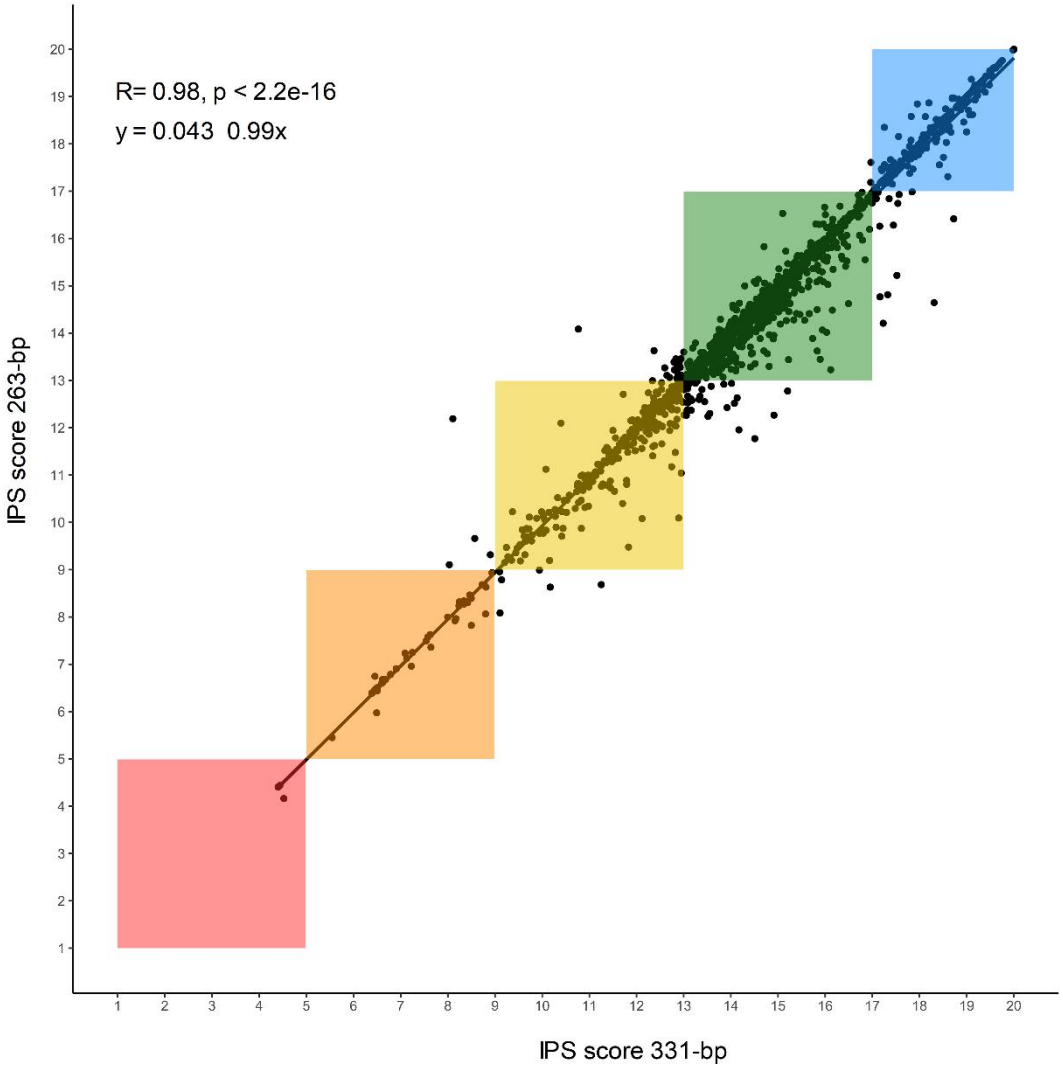
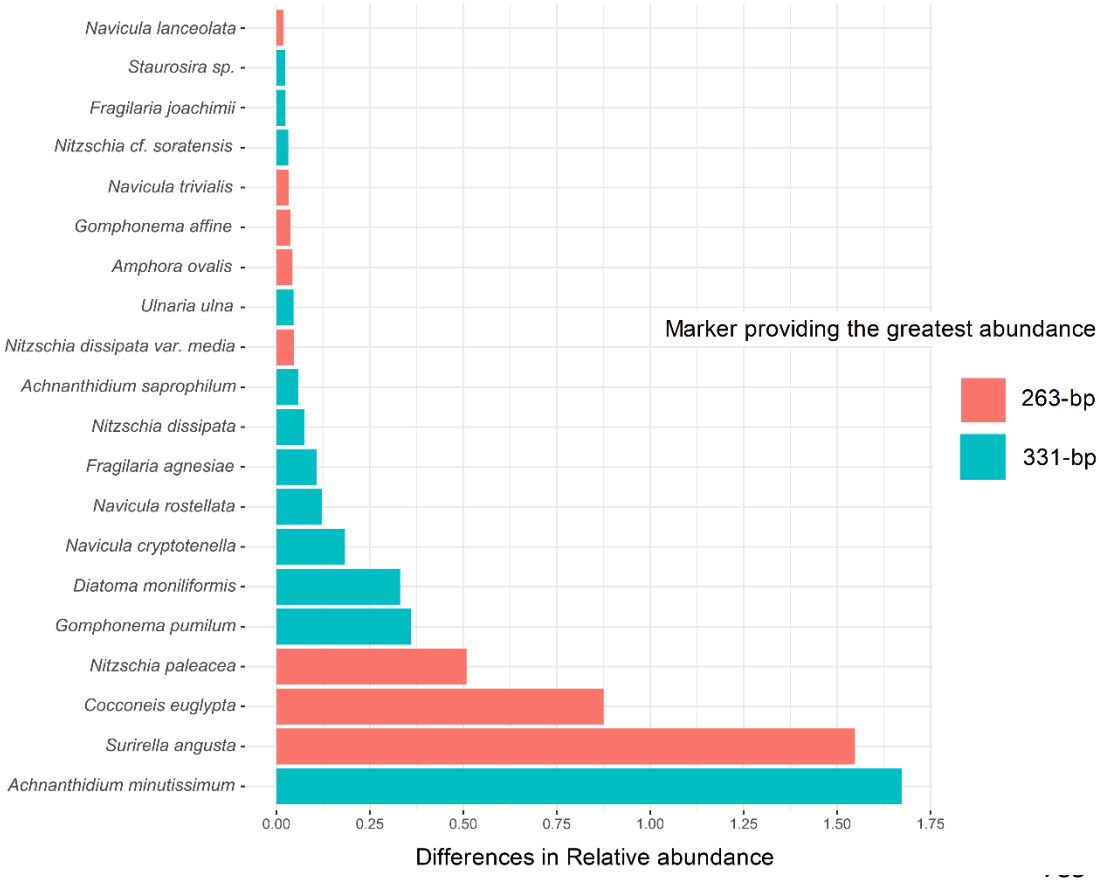
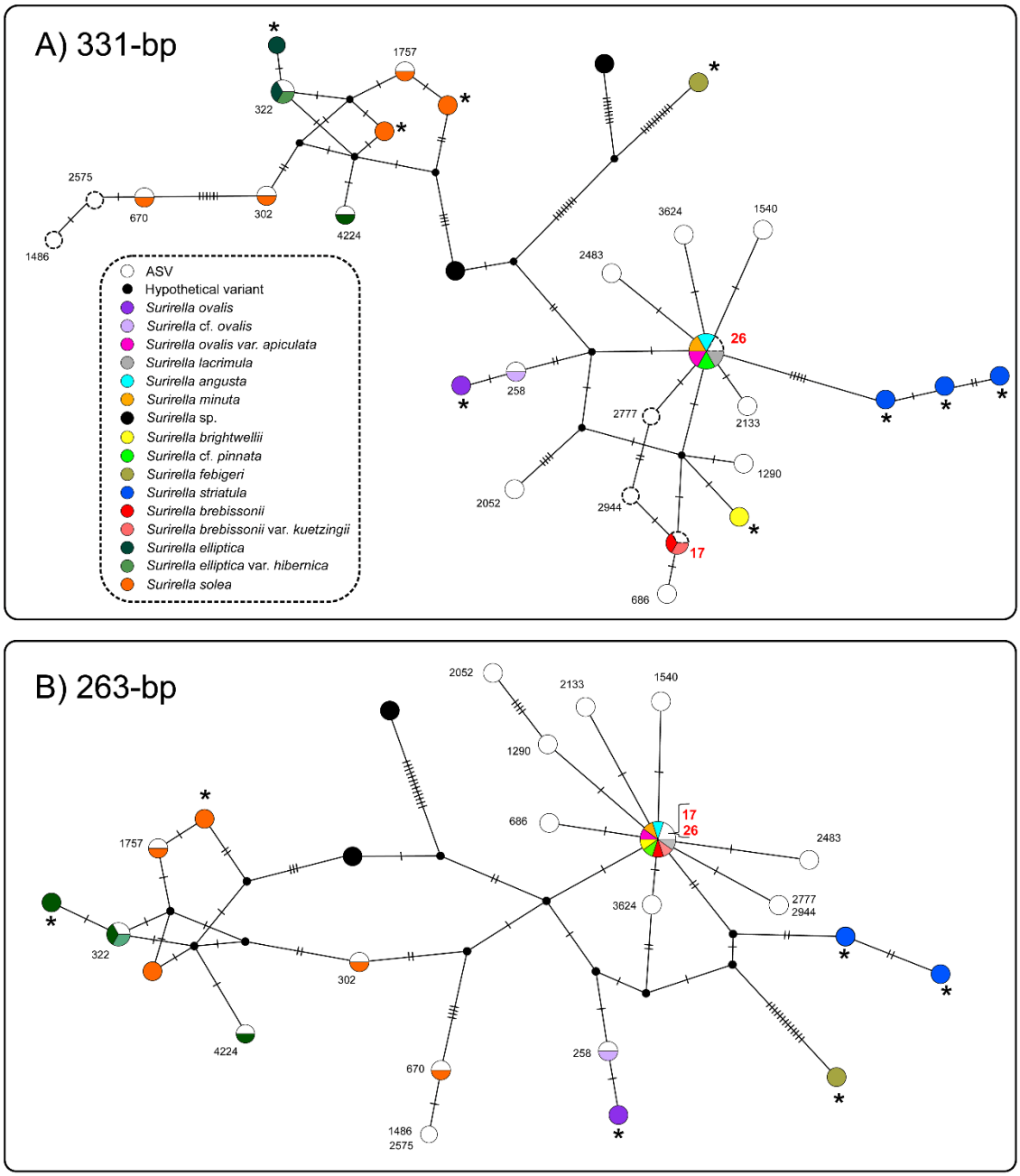


Fig. 1

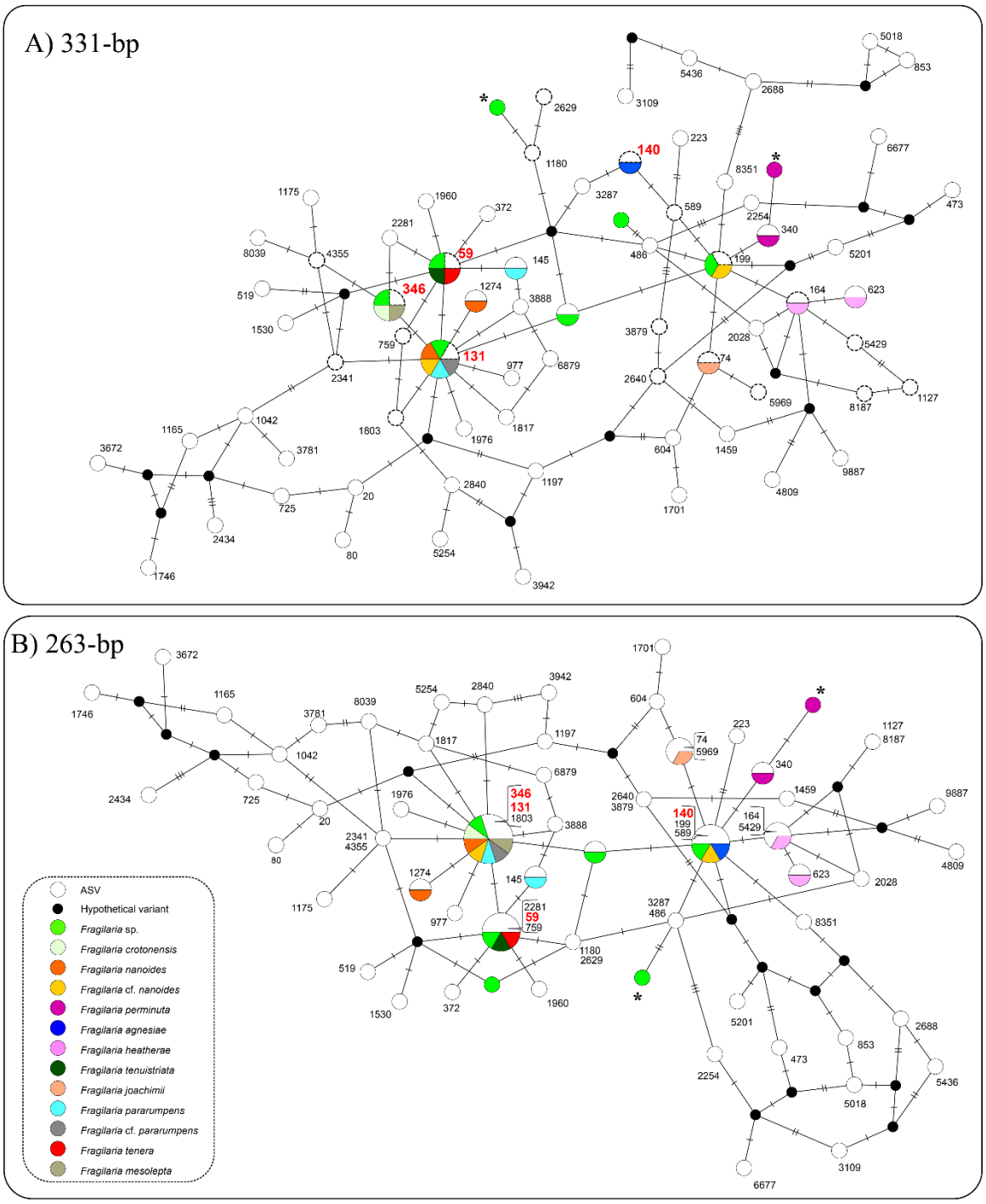

Navicula lanceolata
Staurosira sp.
Fragilaria joachimii
Nitzschia cf. soratensis
Navicula trivialis
Gomphonema affine
Amphora ovalis
Ulnaria ulna
Nitzschia dissipata var. media
Achnanthidium saprophilum
Nitzschia dissipata
Fragilaria agnesiae
Navicula rostellata
Navicula cryptotenella
Diatoma moniliformis
Gomphonema pumilum
Nitzschia paleacea
Cocconeis euglypta
Surirella angusta
Achnanthidium minutissimum
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Differences in Relative abundance
Marker providing the greatest abundance
263-bp
331-bp

786 Fig. 2


A) 331-bp
1757
322
2575
670
302
4224
1486
3624
1540
2483
26
258
2777
2133
1290
2052
2944
17
686
ASV
Hypothetical variant
Surirella ovalis
Surirella cf. ovalis
Surirella ovalis var. apiculata
Surirella lacrimula
Surirella angusta
Surirella minuta
Surirella sp.
Surirella brightwellii
Surirella cf. pinnata
Surirella febigeri
Surirella striatula
Surirella brebissonii
Surirella brebissonii var. kuetzingii
Surirella elliptica
Surirella elliptica var. hibernica
Surirella solea
B) 263-bp
2052
2133
1540
1290
686
17
26
2483
1757
3624
2777
2944
322
302
4224
258
670
1486
2575

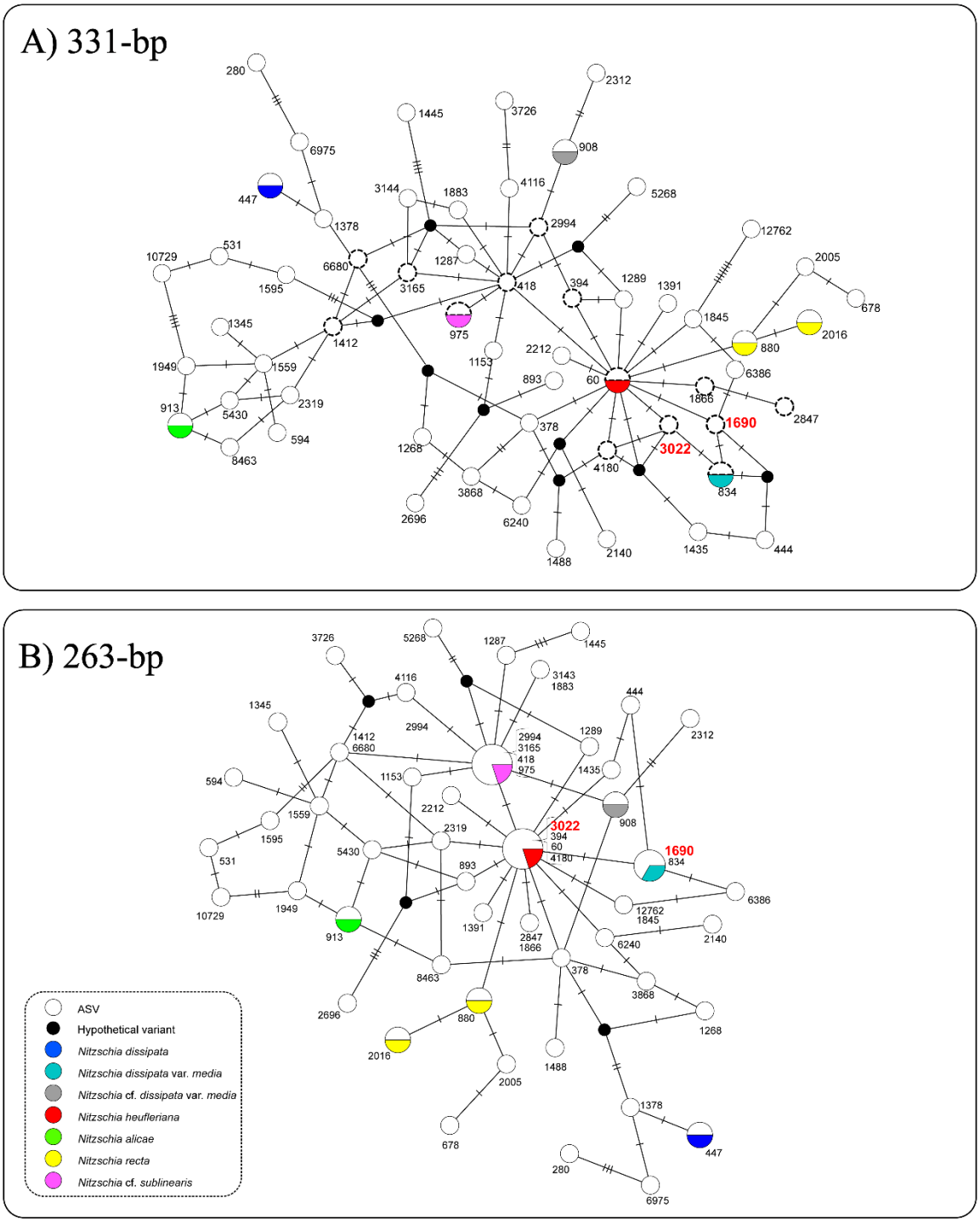
810

811 Fig. 3

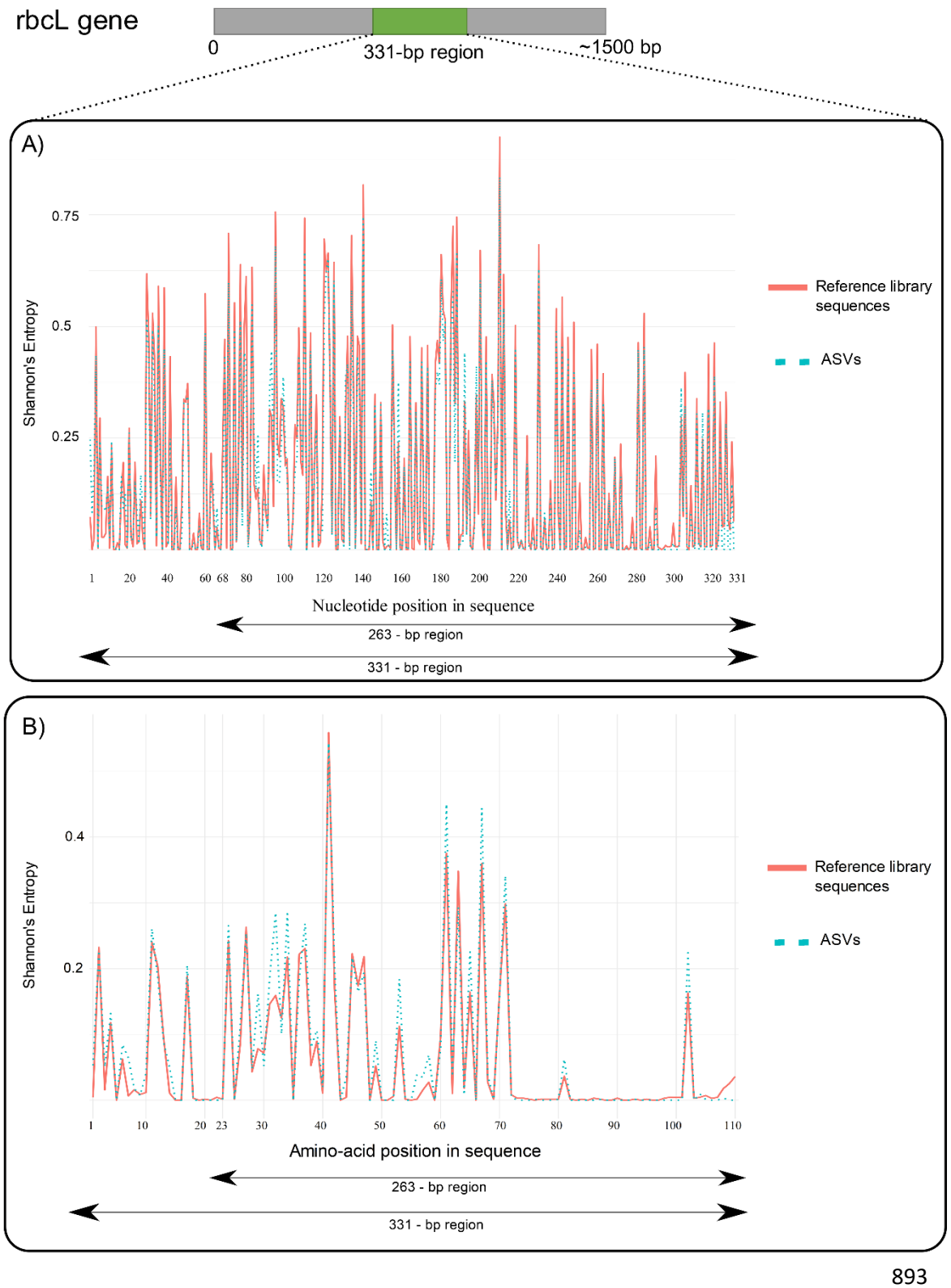

A) 331-bp
B) 263-bp
ASV
Hypothetical variant
Fragilaria sp.
Fragilaria crotonensis
Fragilaria nanoides
Fragilaria cf. nanoides
Fragilaria perminuta
Fragilaria agnesiae
Fragilaria heatherae
Fragilaria tenuistriata
Fragilaria joachimii
Fragilaria pararumpens
Fragilaria cf. pararumpens
Fragilaria tenera
Fragilaria mesolepta

837

838 Fig. 4


A) 331-bp
280
6975
447
1445
3726
2312
908
3144
1883
4116
5268
1378
2994
12762
10729
531
6680
1287
1595
3165
418
394
1289
1391
2005
678
1845
1345
975
1412
2016
880
1949
1559
1153
2212
6386
893
60
1866
913
5430
2319
2847
1690
378
1268
3022
4180
594
8463
3868
834
2696
6240
1435
444
1488
2140
B) 263-bp
3726
5268
1287
1445
4116
3143
1883
444
1345
2994
1289
2312
1412
6680
2994
3165
418
975
1435
1153
594
2212
1559
908
2319
3022
1595
394
60
1690
5430
4180
834
531
893
6386
10729
1949
12762
1845
913
1391
2847
1866
6240
2140
8463
378
3868
2696
880
1268
2016
1488
2005
1378
678
447
280
6975
ASV
Hypothetical variant
Nitzschia dissipata
Nitzschia dissipata var. media
Nitzschia cf. dissipata var. media
Nitzschia heufleriana
Nitzschia alicae
Nitzschia recta
Nitzschia cf. sublinearis

865 Fig. 5


rbcL gene
0
331-bp region
~1500 bp
A)
Shannon's Entropy
0.75
0.5
0.25
1 20 40 60 68 80 100 120 140 160 180 200 220 240 260 280 300 320 331
Nucleotide position in sequence
263 - bp region
331 - bp region
Reference library sequences
ASVs
B)
Shannon's Entropy
0.4
0.2
1 10 20 23 30 40 50 60 70 80 90 100 110
Amino-acid position in sequence
263 - bp region
331 - bp region
Reference library sequences
ASVs

893

894 Fig. 6

Description of Supplemental material

Supplementary Table 1. List of the 29 ASVs that shared the full 331-bp region with reference sequences from more than 1 species. Correspond species of the reference sequences matching the 263-bp and 331-bp markers are shown. Note that identical matches between ASVs and reference sequences were considered as a pairwise-alignment with 100% similarity, no gaps and mismatches, and a full cover of the query ASV by the reference sequence.

Supplementary Data 1. Outputs from BLAST analyses executed on the pairwise comparison between ASVs based on the 331-bp marker (query sequences) and references sequences from Diat.barcode v10 (subject sequences). Columns indicate the percentage of identical matches; alignment length; number of mismatches; number of gap openings; BLAST bit score and BLAST E-value. Note that ASVs shown are those that share at least 95% of sequence identity with reference sequences.

Supplementary Data 2. Outputs from BLAST analyses executed on the pairwise comparison between ASVs based on the 263-bp marker (query sequences) and references sequences from Diat.barcode v10 (subject sequences). Columns indicate the percentage of identical matches; alignment length; number of mismatches; number of gap openings; BLAST bit score and BLAST E-value. Note that ASVs shown are those that share at least 95% of sequence identity with reference sequences.

Supplementary Data 3. Taxonomic classification provided by the Naïve Bayesian classifier for 331-bp and 263-bp ASVs. Bootstrap support values provided at the species level are given.

Supplementary Data 4. Abundance distribution (given as sequencing reads) of ASVs throughout the 1703 samples analyzed.