Evaluation and sensitivity analysis of diatom DNA metabarcoding for WFD bioassessment of Mediterranean rivers.

Javier Pérez-Burillo[1,2*], Rosa Trobajo[1], Valentin Vasselon[3,4], Frédéric Rimet[5,6], Agnès Bouchez[5,6] & David G. Mann[1,7]

[1] IRTA-Institute for Food and Agricultural Research and Technology, Marine and Continental Waters Programme. Ctra de Poble Nou Km 5.5, E43540, Sant Carles de la Rápita, Catalonia, Spain

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

[3] Pôle R&D « ECLA », France

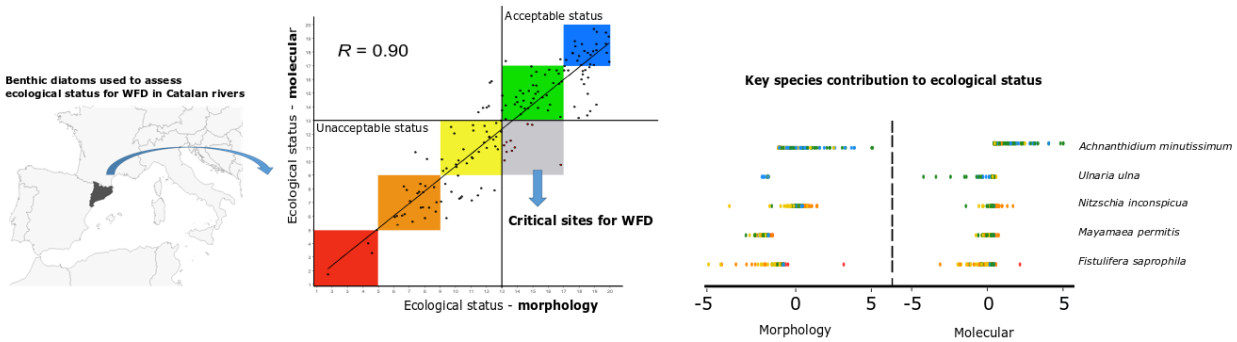
[4]AFB, Site INRA UMR CARRTEL, Thonon-les-Bains, France

[5] INRAE, UMR Carrtel, 75 av. de Corzent, FR-74203 Thonon les Bains cedex, France

[6] University Savoie Mont-Blanc, UMR CARRTEL, FR-73370 Le Bourget du Lac, France

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

*correspondence Javier.perez@irta.cat



Graphical abstract

Abstract

Our study of 164 diatom samples from Catalonia (NE Spain) is the first to evaluate the applicability of DNA metabarcoding, based on high throughput sequencing (HTS) using a 312-bp *rbc*L marker, for biomonitoring Mediterranean rivers. For this, we compared the values of a biotic index (IPS) and the ecological status classes derived from them, between light microscope-based (LM) and HTS methods. Very good correspondence between methods gives encouraging results concerning the applicability of DNA metabarcoding for Catalan rivers for the EU Water Framework Directive (WFD). However, in 10 sites, the ecological status class was downgraded from "Good"/"High" obtained by LM to "Moderate"/"Poor"/"Bad" by HTS; these "critical" sites are especially important, because the WFD requires remedial action by water managers for any river with Moderate or lower status. We investigated the contribution of each species to the IPS using a "leave-one-out" sensitivity analysis, paying special attention to critical sites. Discrepancies in IPS between LM and HTS were mainly due to the misidentification and overlooking in LM of a few species, which were better recovered by HTS. This bias was particularly important in the case of *Fistulifera saprophila*, whose clear underrepresentation in LM was important for explaining 8 out of the 10 critical sites and probably reflected destruction of weakly-silicified frustules during sample preparation. Differences between species in the *rbc*L copy number per cell affected the relative abundance obtained by HTS for *Achnanthidium minutissimum*, *Nitzschia inconspicua* and *Ulnaria ulna*, which were also identified by the sensitivity analysis as important for the WFD. Only minor IPS discrepancies were attributed to the incompleteness of the reference library, as most of the abundant and influential species (to the IPS) were well represented there. Finally, we propose that leave-one-out analysis is a good method for identifying priority species for isolation and barcoding.

Keywords

Environmental DNA, High-throughput Sequencing, *rbc*L, Water Framework Directive, Benthic diatoms, Catalan rivers

Highlights

DNA- and morphology-based diatom assessments of river ecological status are compared

Diatom DNA metabarcoding can be a reliable tool for WFD assessment of Catalan rivers

Sensitivity analysis shows which species drive ecological status assessments

Metabarcoding–morphology ecological status deviations are caused by a few key species

Metabarcoding shows some diatoms are seriously underrecorded in light microscopy

1.Introduction

The key role of diatoms in aquatic systems is well known and is due, amongst other things, to their importance in food webs and biogeochemical cycles and their great contribution to carbon fixation (Armbrust, 2009; Mann, 1999; Smetacek et al., 1999). In addition, their rapid and specific response to environmental changes, great diversity and ubiquitous distribution, and the well-known ecological preferences of many diatom species, have allowed the use of benthic diatoms as biological indicators in biomonitoring programmes, including those for European rivers (Kelly et al., 2008, 2009) demanded by Water Framework Directive (WFD; Directive 2000/60/EC, 2000).

Several diatom indices have been proposed for ecological status assessment, most of them being derived from the formula of Zelinka and Marvan (Zelinka and Marvan, 1961). One of the most commonly used indices for benthic diatoms is the Indice de Polluosensibilité Spécifique (IPS; Cemagref, 1982) which, like other widely used diatom indices, is calculated on the basis of species' relative frequencies, pollution sensitivity values (IPSS) and pollution tolerance values (IPSV). However, the morphological identifications at species level needed for the calculation

of these indices are a time-consuming task and require expert knowledge; furthermore, the taxonomic boundaries are still not well defined in a large number of species and complexes, hampering or even precluding their identification by light microscopy (Mann et al., 2016).

DNA metabarcoding [i.e. the identification of species through a short DNA region, coupled with high-throughput sequencing (HTS)] of environmental samples, has emerged as an alternative method to the classic light microscopical (LM) identifications, due to its speed, reproducibility and cost (Kermarrec et al.,2014; Zimmermann et al., 2015). An increasing number of studies have tested the applicability of this molecular tool for ecological assessment based on benthic diatoms by comparing the ecological index values from DNA metabarcoding with those from LM morphology (Bailet et al., 2019; Kelly et al., 2018; Kermarrec et al., 2014; Mortágua et al., 2019; Vasselon et al., 2017b). Although results have been promising, it has been pointed out that both species composition and relative abundance data obtained by the DNA metarbarcoding may be biased by factors such as the incompleteness of the reference library (Bailet et al., 2019; Rivera et al., 2018a), the DNA extraction method (Vasselon et al., 2017a), the DNA barcode used (Kermarrec et al., 2013), the bioinformatics treatment (Rivera et al., 2020), and the gene copy number per cell (Vasselon et al., 2018). These biases need to be understood, especially their effect on the final IPS score, before the molecular method can be used reliably for routine WFD biomonitoring.

For the management of European rivers covered by the WFD, incongruences between methods become especially important when they cause the perceived ecological status of a water body to change class (five classes are recognized: High, Good, Moderate, Poor and Bad). The most important difference occur when morphological analysis (the current methodological standard) assigns “Good” or “High” ecological status to a particular site but the molecular approach assigns instead a “Moderate”, “Poor” or “Bad” status. This is because the WFD requires action to be taken to improve those aquatic systems that do not reach at least “Good” ecological status and this often has economic implications. We will therefore focus on these

96 “critical sites” in the current paper (i.e. on those Catalan sites whose status alters from
97 Good/High in LM assessments to Moderate/Poor/Bad with DNA metabarcoding), while
98 accepting that a detailed analysis of movements across other status boundaries may also be of
99 interest and relevance to regulators. In particular, we analyse how different biases may
100 contribute to making the IPS score drop below the critical Good to Moderate threshold. There
101 has previously been some analysis of the extent to which particular diatom species contribute
102 to the final ecological status obtained morphologically (Almeida et al., 2014) and to deviations
103 in IPS values between the molecular and morphological methods (Bailet et al.,2019). In both
104 studies, the analyses were based only on relative abundances of species. However, since the
105 IPS value depends not only on the relative abundances of the species present in a sample, but
106 also on their pollution sensitivity values (IPSS) and tolerance values (IPSV), the contribution of
107 each species to the final IPS score for that sample should take all three parameters into
108 account. This will allow the real impact of each species on the final IPS score to be evaluated
109 and thus identify the main species that lead to IPS discrepancies between methods.

110 Therefore, this study of Catalan rivers (NE Spain) aims first to analyse the applicability of DNA
111 metabarcoding as a reliable tool for the WFD biomonitoring of Mediterranean rivers, through
112 the comparison of IPS values obtained from morphological and molecular inventories. The
113 second objective is a sensitivity analysis to quantify the contribution of the different diatom
114 species to the final IPS scores, by either the morphological or molecular method. This will
115 identify which species are driving IPS deviations between the methods, especially in the critical
116 sites that are classified as having unacceptable ecological status (i.e. sites that do not reach
117 Good ecological status) by the DNA metabarcoding approach but are assessed to be
118 acceptable (with Good or High status) using the classical morphological identifications. The
119 third objective is to determine the biases that underlie the differences found between
120 methods in those species identified as important for the WFD according to the sensitivity
121 analysis.

2.Material and methods

2.1. Study site

The study area corresponds to the hydrographic area of Catalonia, which is divided into internal and interregional hydrographic basins (Fig. 1). The former comprises a total of eleven main rivers and extends across 16,423 km^2 (52% of the territory of Catalonia). Among these eleven, the basins of the rivers Llobregat and Ter are the most extensive and occupy approximately half of the total surface covered by the internal basins. The interregional basins are shared with other Spanish regions and cover the Catalan sections of the rivers Ebro, Garona and Xúquer, with a total extent of 15,567 km^2 (48% of the surface area of Catalonia). For this study, 160 out of the total 164 samples were taken from rivers that belong to the internal basins and the remaining 4 samples were collected from the Lower Ebro river (Fig. 1). The rivers sampled are influenced predominantly by Mediterranean climatic factors, though some of them are affected by continental or high mountain climates. This climatic diversity, together with the varied geology and the irregular terrain characteristic of Catalonia, has led to Catalan rivers being classified into 10 different types (ACA, 2010). On the other hand, Catalan rivers are affected by various anthropogenic pressures, such as urban and industrial wastewater discharges, urban and industrial land uses, agriculture, and hydromorphological alterations.

2.2. Diatom sampling

All 164 sites were sampled for epilithon between April and July of 2017 following standard procedures (CEN, 2014a). At each site, diatoms were collected from at least 5 stones by brushing their upper surfaces using a toothbrush. The resulting samples were divided into two aliquots, one of which was preserved with formalin or ethanol and used for morphological analyses as part of the statutory monitoring and control program of the Catalan Water Agency

147 (ACA). The second aliquot was preserved by adding >95% ethanol (to a final concentration of
148 70%) and used for DNA metabarcoding analysis following the recommendations of the
149 technical report of the European Committee for Standardization (CEN, 2018).

150 *2.3. Morphological analyses*

151 Samples were prepared for morphological analyses using light microscopy (LM) according to
152 WFD standards for phytobenthos (CEN, 2014b). Briefly, the organic matter of the samples was
153 removed by chemical oxidation (e.g. by H_2O_2, HNO_3 or H_2SO_4, depending on the consultancy
154 undertaking the analysis for the Catalan Water Agency) and cleaned diatom valves were
155 permanently mounted with Naphrax resin (Brunel microscopes, Chippenham, UK). Finally, at
156 least 400 valves were identified at species level under LM (using a 100× oil immersion
157 objective) and following mainly Krammer and Lange-Bertalot (1986, 1988, 1991a, 1991b) and
158 Lange-Bertalot et al. (2017).

159 *2.4. DNA extraction and PCR amplification*

160 A volume of 2 mL of each benthic sample was centrifuged for 20 min at 4ºC and 12,000 rpm.
161 Ethanol present in the supernatant was removed and total DNA contained in the pellet was
162 extracted using the commercial DNA extraction kit Macheray-Nagel NucleoSpin® Soil kit (MN-
163 Soil). A short *rbc*L region of 312 bp constituted the DNA marker and this was amplified by PCR
164 using an equimolar mix of the modified versions of the primers Diat_rbcL_708F (forward) and
165 R3 (reverse) given by Vasselon et al. (2017b). In order to prepare the HTS library using a 2-step
166 PCR strategy, a part of the P5 (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG) and P7
167 (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA) Illumina adapters were included at the 5' part
168 of the forward and reverse primers, respectively. PCR1 reactions for each DNA sample were
169 performed in triplicate using 1 µL of the extracted DNA in a final volume of 25µL. Conditions
170 and the reaction mix of the PCR1 followed the procedure described in Vasselon et al. (2017b).

171 *2.5. High-throughput sequencing*

For each sample, the three PCR1 replicates were pooled and sent to "Plateforme Génome Transcriptome" (PGTB, Bordeaux, France) where HTS library preparation and sequencing were performed. For the sequencing process, PCR1 products were purified and used as template for a second round of PCR2 with Illumina tailed primers targeting the half of P5 and P7 adapters. The resulting 164 dual-indexed amplicons were pooled for sequencing on an Illumina MiSeq platform using the V2 paired-end sequencing kit (250 bp x 2).

2.6. Bioinformatic analysis

The sequencing facility performed the contig and demultiplexing steps, providing a fastq file for each of the 164 libraries. All the fastq files were then treated together following a bioinformatics process based on Vasselon et al. (2017b), using Mothur software (Schloss et al., 2009). Filtering steps excluded low quality DNA reads that had any of the following properties: reads with lengths <250 bp, Phred quality score < 23 over a moving window of 25 bp, more than 1 mismatch in the primer sequence, homopolymer > 8 bp, or with an ambiguous base. Chimeras were removed using the Uchime algorithm (Edgar et al., 2011). The taxonomic affiliation of the reads was determined using the database adapted for metabarcoding "Rsyst::diatom_rbcl_align_312bp database" (Vasselon et al., 2018b), which is derived from the curated diatom reference library Diat.barcode v7 (Rimet et al., 2019, available at https://www6.inra.fr/carrtel-collection_eng/Barcoding-database and at https://doi.org/10.15454/HYRVUH), and the naïve Bayesian method (Wang et al., 2007) with a confidence score threshold of 60%. Reads not assigned to the Bacillariophyta at the 60% level were excluded from further analyses. A similarity distance matrix based on uncorrected pairwise distances between aligned reads was generated (algorithm proposed by Needleman and Wunsch, 1970) in order to cluster DNA reads into OTUs using the furthest neighbor algorithm as implemented in Mothur; the distance similarity threshold was 95% as previously described for *rbc*L diatom metabarcoding (Vasselon et al., 2017b). Singletons were then filtered and samples represented by less than 3610 reads were removed from the analysis in

198 order to conserve a sufficient sequencing depth to characterize diatom community structure.
199 In order to allow inter-sample comparisons, the remaining samples were then normalized to
200 the same read number using the smallest read abundance among them. Diatom molecular
201 inventories were obtained using the taxonomy of OTUs corresponding to the consensus
202 taxonomy of DNA reads with a consensus confidence threshold over 80%.
203 For brevity, we often use "HTS" to refer to the whole process of deriving ecological status
204 metrics by DNA metabarcoding, when contrasted with the process of obtaining them via light
205 microscopical counts of diatom valves ("LM").

206 *2.7. Ecological status class assignment*

207 Ecological status was determined by applying the IPS (Cemagref, 1982), since it is the diatom
208 index adopted by Spain for the WFD, as well as by many other EU countries. For each site, the
209 IPS was calculated from species inventories (species composition and relative abundances)
210 obtained from both LM and HTS analyses, using OMNIDIA software v5.5 (Lecointe et al., 1993).
211 The WFD ecological status class was assigned by applying the following boundaries based on
212 the Catalan standards (ACA, 2010): High (17 ≤ IPS ≤ 20), Good (13 ≤ IPS < 17), Moderate (9 ≤
213 IPS < 13), Poor (5 ≤ IPS < 9), Bad (1 ≤ IPS < 5). Those sites classified as Good/High by LM but as
214 Moderate/Poor/Bad by HTS are referred to as "critical sites".

215 *2.8. HTS correction factor application*

216 Diatom species sometimes differ in the *rbc*L copy number per cell (depending on the number
217 of gene copies per chloroplast and the number of chloroplasts per cell) and Vasselon et al.
218 (2018) found a strong correlation between *rbc*L copy number per cell and cell biovolume. They
219 therefore suggested that a correction factor (CF) based on cell biovolume should be applied to
220 the proportions of reads before making comparisons with valve counts (morphology).
221 Accordingly, we applied Vasselon et al.'s (2018b) modified CFs (Rivera et al., 2020) to the HTS
222 reads in order to assess their effectiveness in improving the DNA-based ecological status

assessments; the CFs were extracted from Diat.barcode v7 (Rimet et al., 2019). The IPS values and the number of critical sites were compared between the LM inventory and both corrected and uncorrected HTS inventories.

2.9. Evaluation of differences between morphological and molecular approaches and species sensitivity analyses

The percentage of species identified by both methods was determined. The percentage of species identified molecularly that were also identified by the morphological approach, and the percentage of species identified morphologically that were also identified by the molecular approach, were calculated in order to assess the effectiveness of the two methods in identifying taxa. The percentages of the total morphological counts and total molecular reads (of the total 162 samples) contributed by the species identifiable by both methods were also calculated.

To compare IPS outcomes obtained by the two methods (morphology and DNA metabarcoding), the percentage of sites assigned to the same ecological status class was determined and the correlation in IPS values between the methods assessed by Pearson's coefficient. Special attention was paid to the critical sites.

For each of the 162 sites (this was the number of sites remaining after normalizing the data to 3610 reads), a sensitivity analysis to determine the contribution of each species to the IPS value was performed by a "leave-one-out" method. The contribution was calculated as the difference between the IPS value when the entire diatom community observed in a given site was considered and the IPS value for that site once the particular species was left out (i.e. not included in the IPS calculation). Therefore, for each of the species identified in each site, a positive or negative value was obtained, indicating a positive contribution of the species (i.e. the IPS value decreases when the species is omitted during calculation of the IPS) or a negative contribution (i.e. the IPS value increases when the species is not considered), respectively.

248 Calculations of species' IPS contributions were done for both the morphological and the
249 metabarcoding approaches.

250

251 **3. Results**

252 *3.1. Light microscopy*

253 In total, 410 taxa were identified by light microscopy, of which 351 were identified at species
254 level. The number of species identified per sample ranged from 4 to 61, with an average of
255 27.3. The ten most abundant species, in order, were: *Achnanthidium minutissimum, Nitzschia*
256 *inconspicua, Fistulifera saprophila, Amphora pediculus, Planothidium frequentissimum,*
257 *Achnanthidium pyrenaicum, Mayamaea permitis, Cocconeis euglypta, Craticula subminuscula*
258 and *Navicula gregaria* (Supplementary Fig. 1)

259 *3.2. Metabarcoding data*

260 A total of 9,941,912 reads were obtained by MiSeq Illumina sequencing of the 164 samples.
261 After quality filtering steps 3,081,893 reads were retained and clustered into 708 OTUs with an
262 average of 78.2 per sample. The maximum and minimum numbers of OTUs per sample were
263 182 (comprised by a total of 21,654 reads) and 7 (comprised by a total of 14 reads),
264 respectively. To allow inter-sample comparisons, samples were normalized to 3610 reads,
265 representing the minimum number of reads per sample recorded after removal of 2 samples
266 comprising 2033 and 14 reads respectively. The remaining, rarefied data comprised a total of
267 584,820 reads clustered into 615 OTUs, with an average of 61.1 OTUs per sample, the
268 maximum and minimum being 137 and 10. The OTUs were assigned to a total of 148 taxa, of
269 which 138 were species, with an average of 30.9 species per sample and ranging from 5 to 55
270 species per sample (Supplementary data). 18.3% of the reads (corresponding to the 51.4% of
271 the total 615 OTUs) were not successfully classified at species level, the percentage of
272 unclassified reads per sample varying from 0.2% to 71.6%. The ten most abundant species

were *Achnanthidium minutissimum, Fistulifera saprophila, Planothidium victorii, Mayamaea permitis, Cocconeis placentula, Melosira varians, Craticula subminuscula, Gomphonema pumilum* var. *pumilum, Ulnaria ulna,* and *Nitzschia inconspicua* (Supplementary Fig. 1).

3.3. Comparison between molecular and morphological inventories

Taken together, the LM and HTS approaches identified a total of 451 different species, of which 103 (27%) were common to both. Only 29% of the 351 species identified by LM were also identified by HTS, while 75% of the 138 species identified by HTS were also identified by LM. However, when expressed in terms of valve numbers and reads, the agreement between the two approaches was much closer: the species identified by both approaches accounted for 80% of the total valves counted by LM, 72% of the total reads recorded by HTS, and 88% of the total reads recorded by HTS that were successfully assigned to species.

3.4. Ecological status comparison between approaches

IPS values obtained with the morphological inventory varied from 19.9 to 1.7 with an average of 13.9, while IPS values varied from 19.7 to 1.75 with an average of 12.7 in the HTS analysis. IPS values from both approaches were highly correlated (Pearson's R = 0.90) (Fig. 2). 113 sites (69.8%) were assigned to the same ecological status in both approaches and 49 sites (30.2%) showed 1 class of difference (Table 1).

A total of 10 critical sites were identified since they were classified as Good or High (i.e. acceptable ecological status) by the morphological approach but as Moderate, Poor or Bad (i.e. unacceptable ecological status) by HTS (table 1).

When the biovolume CF was applied to the molecular data, IPS values varied from 19.8 to 2.3 with an average of 12.4. The correlation between IPS values obtained from morphology and from CF corrected HTS was 0.92 (Pearson's R), so slightly higher than without applying the CF. However, the number of sites that shared the same ecological status decreased when the CF was applied (106 sites, representing 65.4% of the samples) and the number of sites that

298 showed 1 and 2 classes of differences increased slightly [51 sites (31.5%) and 5 sites (3.1%)
299 respectively]. Furthermore, and importantly, five new critical sites were obtained when CFs
300 were applied, resulting in a total of 15 critical sites.

301 *3.5. Species sensitivity analysis*

302 *3.5.1. All sites*

303 The analyses of species contributions to IPS revealed that in both approaches the species that,
304 on average, most negatively affected the IPS values were *Fistulifera saprophila, Navicula*
305 *veneta* and *Mayamea permitis* (Fig. 3; Supplementary data). *Achnanthidium minutissimum* was
306 the species with the most positive average IPS contribution with both HTS and LM, but the
307 species with the second and third most positive IPS contributions differed between
308 approaches: *A. pyrenaicum* and *Amphora pediculus* were the higher contributors in LM but
309 *Planothidium lanceolatum* and *Cocconeis placentula* in HTS (Fig. 3; Supplementary data)

310 Some other species, such as *Nitzschia inconspicua, N. fonticola, Navicula gregaria,*
311 *Planothidium frequentissimum* and *Melosira varians* sometimes contributed positively to the
312 IPS scores, sometimes negatively (Fig. 3; Supplementary data), depending on the whole diatom
313 assemblage in the sample.

314 A further group of species, *Navicula reichardtiana, Achnanthidium rostropyrenaicum,*
315 *Cocconeis placentula var. lineata, Gomphonema lateripunctatum* and *Cocconeis euglypta,*
316 made zero contribution to the IPS when this was calculated from HTS data due to the lack of
317 sequences of these species in the reference library (Fig. 3; Supplementary data).

318 Overall, the greatest contributions to IPS values were made by the most abundant species.
319 However, lower abundance species (< 5%) also made important contributions if their indicator
320 values were very high or very low. Furthermore, and more importantly perhaps, though it is
321 very easily overlooked, the contribution of these species (i.e. low abundance species, with very
322 high or very low IPSS) was influenced by the IPS score of the whole sample. That is, species

with very low IPSS values made a relatively greater contribution in samples where the overall IPS score was high (and the reverse was also true). An example is given by the sensitivity analysis results for our samples 76 and 138. In sample 76, *Achnanthidium minutissimum* was recorded (HTS) with a relative abundance of 3.66% and the sensitivity analysis showed a contribution of 0.97 towards the overall IPS (HTS) score of 7.76. In contrast, in sample 138, with an overall IPS score of 18.05 and in which *A. minutissimum* was recorded (HTS) in almost the same relative abundance (3.77%) as in sample 76, the sensitivity analysis showed a much lower contribution (0.07) of the species to the overall IPS score (Supplementary data).

3.5.2. Critical sites

Analyses of IPS species contributions (LM vs HTS) are shown in Fig. 4. These show that the species most often responsible for causing sites to become critical was *Fistulifera saprophila*. This species showed a clear discrepancy between its contribution to IPS values calculated from LM valve counts and that from HTS reads. The species was recorded by HTS in all the critical sites (10) and in 8 of them was found to be the first-, second- or third-ranked species (in 4, 2 and 2 sites respectively) for its negative contribution to the IPS (Fig. 4, left). However, with LM, *F. saprophila* was recorded in only 4 of the critical sites and in only 1 of these 4 sites was it ranked as among the four most negative contributors (it was the second).

Mayamaea permitis was also revealed as an important species for some critical sites. It was recorded by HTS in all the 10 critical sites and was the first, second and third species that most negatively contributed to the IPS score in 2, 1 and 3 sites respectively (Fig. 4, left). In the LM analyses, although it was found in 8 of the 10 critical sites, it was the one that contributed most negatively in only 3 sites.

Nitzschia inconspicua is also an important contributor to the low IPS values of critical sites but mainly in the LM based assessments. The species was identified by LM in 8 of the 10 critical

348 sites, and was the first-, second- and fourth-ranked species that most negatively affected the
349 IPS in 2, 1 and 2 sites respectively, while with HTS, although it was identified in 9 of the critical
350 sites, it was never amongst the 3 species that most negatively affected the IPS scores (Fig. 4,
351 left). Hence it cannot be crucial for making sites critical. Discrepancies between methods in the
352 contributions to IPS values in the species *Pleurosira laevis* and *Craticula subminuscula* were
353 relevant in determining 2 critical sites. Both were recorded as the first-ranked species that
354 most negatively contributed to IPS in one site by HTS, while they were never ranked amongst
355 the 3 species that most negatively affected the IPS by LM (Fig. 4, left).

356 *Achnanthidium minutissimum* was the species that contributed most positively to IPS scores
357 throughout, at both critical and non-critical sites. However, despite its important influence on
358 the IPS scores, it doesn't seem that it played a crucial role in making sites critical. In the
359 molecular inventory, the species was ranked first or second in seven critical sites by LM and
360 eight by HTS (Fig. 4, right).

361 *3.5.3. Critical sites resulting when applying CFs*

362 The analysis of species contributions for the extra critical sites resulting when CFs were applied
363 revealed that *F. saprophila* and *M. permitis* were again the main species responsible
364 (Supplementary Fig. 2) as a consequence of the upsurge in their relative abundance after
365 applying CFs (Supplementary Fig. 1).

366 *3.6. LM valve counts vs HTS reads for key species for WFD biomonitoring*

367 Comparing the relative abundance between LM and HTS (without CFs) of some of the most
368 abundant species with major effects on the IPS scores (as identified above), four types of
369 pattern could be identified (Fig. 5);

370 1) A tendency to be underrepresented by HTS. This was shown in *Planothidium*
371 *frequentissimum* and *Nitzschia inconspicua,* which were underrepresented in 97% and 90%
372 respectively of the total samples where the species was identified by both methods.

2) The opposite tendency, overrepresentation by HTS, was shown in *Ulnaria ulna*. Of the total of 162 samples analysed, LM recorded the species in only 36 (22%) samples, while it was identified by HTS in 99 (61%). And in those samples where the species was recorded by both methods, it was overrepresented by HTS in 17 samples (61%).

3) Little or no bias overall in the relative abundances between the methods. This is the pattern shown by *Mayamaea permitis* and *Achnanthidium minutissimum*. For example, in the 108 samples where the species was identified by both methods, *M. permitis* was overrepresented by HTS in 50% and underrepresented in 50%. It is worth highlighting that in 9 of the 10 critical sites, *M. permitis* was overrepresented by HTS or not detected at all in LM. In the case of *A. minutissimum* there was a slight tendency to be underestimated by HTS (in 65% of samples where the species was identified by both methods).

4) The pattern shown by *Fistulifera saprophila*. On the one hand, there was a clear bias towards HTS, the species being recorded by this method in 136 samples (84%) out of the total of 162 analysed but in only 76 samples (46%) by LM. On the other hand, in the samples where both methods recorded this species, the pattern seemed to be of underrepresentation by HTS.

4. Discussion

4.1. DNA based diatom metabarcoding is confirmed as a promising new tool for WFD ecological assessment

Both the strong linear relationship between the ecological status results of both methods (morphology-LM and molecular-HTS), and also the fact that the intercept is close to zero, confirm the high potential of DNA metabarcoding as a new monitoring tool for the WFD assessment of Catalan rivers using benthic diatoms. Recent studies have also demonstrated this same potential for other regions of Europe (rivers in UK, France, Central Portugal and Switzerland: Kelly et al., 2018; Rivera et al., 2020; Mortágua et al., 2019; and Visco et al., 2015)

and elsewhere (Mayotte Island rivers: Vasselon et al., 2017b). However, our study is the first to demonstrate the potential for rivers under a Mediterranean climate regime. Interestingly our study found:

i) A higher percentage of species identified by both methods (i.e. shared species) than recorded previously, viz. 26.7%, which compares with the 13% obtained in the tropical island of Mayotte by Vasselon et al. (2017b; this low percentage could perhaps be expected since the Diat.barcode reference library mainly covers species or isolates from temperate regions), 15.7% in Rivera et al. (2018; though this was not for a river but for lake Bourget) and 21.4% in Rivera et al. (2020; our calculation from their data).

ii) These shared species accounted for a high percentage of total LM counts (80%) and HTS reads (72%).

iii) A high percentage (48.62%) of all the OTUs were successfully assigned at species level compared with those obtained previously in similar studies; for comparison, these were: 50.7% by Rivera et al. (2020; our calculation from their data); 41% by Rivera et al. (2018; for lake Bourget); 35.7% by Vasselon et al. (2017b); 32% by Mortágua et al. (2019) and 30% by Keck et al. (2018).

iv) A very high correlation between the IPS values from both methods and also a high % of samples assigned to the same ecological class. To our knowledge, the highest correlation obtained in IPS values between methods is circa 0.83 (Pearson's R; Rivera et al., 2020) while ours is 0.92 after CFs and 0.90 without CFs (Pearson's R). Likewise, in the present work, the proportion of sites that fall into the same ecological status class regardless of the method used is 69.8%, considerably greater than has been obtained in other similar studies (Bailet et al., 2019; Mortágua et al., 2019; Rivera et al., 2020 and Vasselon et al., 2017b).

In spite of these good results, our analyses revealed differences between the methods that noticeably affected both the IPS values and the ecological status assignations. Some of these differences can be attributed to imperfections in the HTS approach, such as the current incompleteness of the DNA reference database and the lack of a full understanding of the relationship between cell numbers and DNA reads. Others, on the contrary, reflect biases in the LM method that were previously hidden. We discuss both of these below, with special reference to differences that affect the final ecological assessment, changing a site from High or Good status to an unacceptable Moderate, Poor or Bad status, i.e. the differences responsible for creating "critical" sites.

4.2. Key diatom species can be neglected by LM but evident from HTS

Our results suggest that it is the misidentification, overlooking or loss of several species by LM that is the main source of IPS discrepancies between LM and HTS in critical sites. This was clearly evidenced when looking at dissimilarities in both abundance and occurrence in *Fistulifera saprophila* (Fig. 5); this species was not recorded at all in 4 out of the 10 critical sites by LM whereas it was recorded by HTS in all of them.

Fistulifera saprophila is characterized by a low IPS-sensitivity value (IPSS = 2), leading to the species contributing negatively to IPS (Fig. 3), especially in sites where it is abundant. Therefore, overlooking this species by LM leads to a falsely high IPS value, explaining why *F. saprophila* was identified as the most discriminative species for critical sites by the leave-one-out method (Fig .4). Interestingly, Kelly et al. (2018) reported very similar discrepancies in *F. saprophila* between the LM and HTS methods, with many sites registering no valves in LM but moderate to high numbers of HTS reads (up to 50% or more). They attributed the misidentification or absence of the species in LM to its weakly silicified frustules, which are easily dissolved by the oxidising mixtures commonly used to prepare samples (Zgrundo et al., 2013).

448 *Mayamaea permitis* is another small, weakly-silicified diatom that can probably be missed
449 during counting, or lost during the preparation process. Overall, *M. permitis* was not
450 overrepresented by either LM or HTS when considering the whole inventory of samples, but
451 there was a noticeable tendency for it to be overrepresented by HTS in critical sites (Fig. 5),
452 which, by analogy with *F. saprophila*, could be explained if misidentification or loss of cells
453 occurred during LM assessments, hence contributing to misleadingly higher IPS values.

454 Another case of presumed misidentification by LM, this time partly because of taxonomic and
455 nomenclatural changes, was observed in *Planothidium frequentissimum*, which was
456 overrepresented in LM and indeed, scarcely recorded at all by HTS (Fig. 5; Supplementary Fig.
457 1). Our results suggest that *P. victorii* was frequently misidentified as *P. frequentissimum*
458 during LM counting, since the relative abundance distribution of *P. frequentissimum* in LM
459 agreed well with the corresponding distribution obtained for *P. victorii* in HTS (Supplementary
460 Fig. 3). In such cases it can be difficult to determine which method (LM or HTS) is likely to be
461 correct. However, in the present case, the sequences of *P. victorii* (and its taxonomic synonym,
462 *P. caputium*) available in the DNA diatom reference database (Diat.barcode v7; Rimet et al.,
463 2019) come from the same clones used to establish the species (Novis et al., 2012, Jahn et al.,
464 2017) and the sequences of *P. frequentissimum* available in the reference library are also likely
465 to have been reliably identified in the taxonomic revision by Jahn et al. (2017). Furthermore,
466 the genetic diversity of these species is apparently well covered (Jahn et al., 2017). Hence, the
467 IPS discrepancies found between the methods should be attributed, not to HTS identification
468 error, but rather to the difficulties in distinguishing between *P. frequentissimum* and *P. victorii*
469 in LM (due to the lack of easily seen morphological differences between them: Jahn et al.,
470 2017), and/or to the difficulties of keeping up-to-date in routine LM counts with all the
471 taxonomic changes being made (guides are often not affordable; the latest taxonomic changes
472 are not always included, etc.). The importance of correctly identifying *P. frequentissimum* by
473 either method lies in the fact that this species is relevant in determining Moderate ecological

474 status because of its intermediate IPS sensitivity value (IPSS=3.4), which leads to a negative or
475 positive influence on the final IPS value, depending on the other species present (Fig. 3). *P.*
476 *frequentissium* and *victorii* also illustrate another problem that arises when there are two (or
477 more) taxa that are so similar morphologically that it is impossible to distinguish them during
478 routine LM. This automatically means that we cannot use LM to determine whether they do or
479 do not have the same ecological preferences; in fact, it will be only possible to determine the
480 preferences of such cryptic or pseudocryptic taxa through combining HTS surveys with
481 analyses of accompanying environmental data. Unfortunately, the *Planothidium* example is
482 not unique; there are several small but abundant freshwater species that are similarly difficult
483 or impossible to discriminate under LM, e.g. in *Nitzschia* (e.g. *N. inconspicua* and *N. soratensis*,
484 Trobajo et al., 2013), or in *Amphora* (Levkov, 2009). We are currently working on some of
485 these to establish whether the different species/OTUs differ in their ecological preferences.

486 *F. saprophila, M. permitis* and *P. frequentissimum*, therefore, are three examples where HTS
487 offers a more accurate or more complete identification than the traditional morphological
488 identification based on LM characters. These species are especially important for WFD
489 biomonitoring assessments, at least in our area, since they can be abundant and were
490 detected by our leave-one-out analyses as influential in defining different ecological status and
491 critical sites. Identification and counts of these species under LM could in fact lead to rivers
492 being wrongly classified as having acceptable WFD ecological status when their "real" IPS
493 might correspond to one of the unacceptable classes (and thus require remedial action by
494 water managers).

495 *4.3. Pitfalls to be overcome*

496 *4.3.1 Gene copy numbers per cell affect the estimates of abundance of important species for*
497 *WFD*

Variation between species in the average *rbc*L copy number per cell constitutes a major bias that may explain incongruences between methods in the relative abundances of species and therefore differences in IPS scores (Pawlowski et al., 2018; Vasselon et al., 2018). Of the species strongly influencing IPS values in our dataset (Fig. 3) and showing differences in abundance between LM and HTS (Fig. 5), three – *Achnanthidium minutissimum, Nitzschia inconspicua* and *Ulnaria ulna* – are species whose gene copy numbers were estimated directly using qPCR by Vasselon et al. (2018). Our findings are consistent with theirs, in that *A. minutissimum* and even more so *N. inconspicua*, tend to be underrepresented with HTS and have low copy numbers per cell, whereas *U. ulna* has a much higher copy number (10–35× the copy number in the other two species according to Vasselon et al.'s data) and is greatly overrepresented with HTS (Fig. 5). Copy number–related differences in these species are potentially relevant for WFD assessments and *A. minutissimum* and *U. ulna* pose a risk of making sites critical as they mainly affect sites classified by LM within the acceptable ecological status and both will tend to lead to lower IPS values in HTS, *A. minutissimum* by underrepresentation and *U. ulna* by overrepresentation. This is well illustrated by the great IPS differences between methods in those sites where *U. ulna* was clearly overrepresented by HTS (Supplementary data; sites 124, 136, 166 and 188).

As with *Planothidium frequentissimum,* the leave-one-out method revealed that *Nitzschia inconspicua* (IPSS=2.8) showed a IPS contribution that shifted from positive (in sites classified as having Good or High ecological status) to negative (in sites classified with Bad or Poor ecological status), driving the IPS values towards Moderate ecological status (Fig. 3). The importance for biomonitoring of the relative abundance discrepancies in this species (clear underrepresentation of the taxon by HTS: Fig. 5) is that it will exaggerate the corresponding IPS values either negatively or positively, depending on the starting point. In those sites where *N. inconspicua* is abundant, the ecological status will be wrongly determined by HTS (relative to LM) in two ways: a) in those sites classified by LM as having Good or High ecological status, the

IPS will be increased even more (i.e. IPS values overestimated); and, in contrast, b) in those sites classified by LM as having an unacceptable WFD level, the IPS will be lowered making them even worse (i.e IPS values underestimated). Similar conclusions apply to *Craticula subminuscula*, which showed a similar IPS contribution pattern to *N. inconspicua* (Fig. 3) and was especially relevant for explaining one critical site (Fig. 4).

The effects of copy number, exemplified in *A. minutissimum, N. inconspicua* and *U. ulna*, suggest that it could be important to apply biovolume-based correction factors, as recommended by Vasselon et al. (2018), and such factors have been applied in the studies of Vasselon et al. (2018), Mortágua et al. (2019) and Rivera et al. (2020). When we applied the correction factor to our dataset, it led to a slight increase in the Pearson correlation coefficient for the LM vs HTS IPS scores, as found by Rivera et al. (2020) and Mortágua et al. (2019). Interestingly, the greatest reduction in the discrepancies between methods in the relative abundances was observed for the relatively high-volume species, such as *Ulnaria ulna* and *Pleurosira laevis*; the latter species was relevant for one critical site though in most of the samples it had a relative abundance lower than 1% (in both LM and HTS inventories). However, the benefits of CFs are mixed, since use in our dataset increased the number of critical sites from 10 to 15, mainly due to the increase in the relative abundance of *F. saprophila* and, to a lesser extent, *M. permitis* (Supplementary Fig. 3). This is to be expected because application of CFs is based on the assumption that low biovolume species, such as *F. saprophila* and *M. permitis*, generate fewer copies of the *rbc*L marker than larger species and will tend to be underrepresented by HTS.

4.3.2. Gaps in the DNA reference library partly explain IPS discrepancies between methods

The good agreement between LM and HTS methods obtained in this study, in terms of the final IPS score, was likely due in large part to the fact that most of the IPS-determining and abundant benthic diatom species of the Catalan river basin district were represented in the

549 DNA reference database used and could therefore be retrieved when the metabarcoding
550 approach was applied. This reference database, Diat.barcode v7 (Rimet et al., 2019), is
551 becoming widely used in diatom metabarcoding studies (Chonova et al., 2019; Mortágua et al.,
552 2019; Rimet et al., 2018 and Rivera et al., 2020) and is continuously curated by experts from
553 different countries. However, it is far from complete and this could potentially be a source of
554 IPS discrepancies between methods, if the missing species are sufficiently abundant and have a
555 strong indicator value. In our case, the taxa amongst the species recovered by LM with a
556 relative abundance greater than 1% that were not identified by HTS, due to the lack of
557 representative barcodes for them in the reference library, were *Cocconeis euglypta,*
558 *Gomphonema lateripunctatum* and *Cocconeis placentula var. lineata* (Supplementary Fig. 1).
559 Of these, *C. euglypta* was amongst the 10 species that contributed most to IPS (Fig. 3).

560 However, although the reference database includes most of the common and influential
561 species of the Catalan river basin district, it may nevertheless be a cause of differences
562 between LM- and HTS-based IPS scores, because the genetic diversity of some species may be
563 inadequately represented in the reference library, leading to underrepresentation in HTS. This
564 issue was suggested by Kelly et al. (2018) to explain underrepresentation of *A. pediculus* by
565 HTS in a UK rivers dataset; likewise, Vasselon et al. (2019) indicated that the genetic diversity
566 of *Nitzschia inconspicua* was not properly covered until the current version of the reference
567 library (Diat.Barcode v7, Rimet et al., 2019) was released. Hence, the improvements made in
568 successive versions of the reference library may in part explain the better results obtained in
569 our study, relative to previous work, since we used the current version 7 while other previous
570 studies based their bioinformatics treatment on version 6 or lower (Bailet et al., 2019;
571 Mortágua et al., 2019 and Vasselon et al., 2017a).

572 The last point we would make in relation to the reference database is that, at least within a
573 limited geographical area and/or a relatively narrow range of water types, the number of
574 “influential” species that must be included to avoid biases in the ecological status assessment

may often be quite limited, as we demonstrate here (Fig. 3) and as shown also by Kelly et al. (2018, fig. 6.10). Hence, although adding any species and genotypes to the reference database will always be useful, before isolating, culturing and Sanger-sequencing new clones it may be worth carrying out an IPS sensitivity analysis of existing LM-based abundance data, to objectively identify priority species for barcoding and hence avoid unnecessary work that may have negligible benefit for WFD biomonitoring.

4.4. Next priority: reference sites

This work has focused particularly on "critical" sites, due to their importance in the WFD. However, also important for the WFD are the reference sites, i.e. sites little altered by human pressures or lacking any human pressure (European Commission, 2016). Reference sites are a key concept for WFD since the different ecological status classes are determined through quantifying deviations from the biota that would exist in pristine conditions. In this study, only 19 reference sites were sampled and this is not sufficient for comparisons between methods and drawing reliable conclusions. Though none of the 19 reference sites crossed the critical threshold and 11 of them were classified by both methods as having high ecological status, 8 of them were downgraded to good status by HTS. A study of a larger dataset including more reference sites is therefore crucial. With an increased number of samples, a sensitivity analysis, like the one performed in this study, could be undertaken to identify species that tend to be restricted to reference conditions and evaluate possible biases resulting from inaccurate identification or quantification in either LM or HTS. In addition, sensitivity analysis could be used to identify which species from reference sites are not currently included in the reference library and should be considered as priorities for barcoding, due to their high relative abundance and/or contribution to the index in these sites. Examples are *Achnanthidium rostropyrenaicum* and *Gomphonema lateripunctatum*, which seem to be important for our reference sites but are not represented in Diat.barcode and so were only identified by LM in our dataset.

Acknowledgments

We are very grateful to the Catalan Water Agency (ACA, especially to Toni Munné, Carolina Solà and Mònica Flo) for managing and organizing the river survey and providing us with the LM counts. We would like also to thank all the consultancies and people who made this work possible through sampling and morphological analysis: Sorelló, Estudis del Medi Aquàtic: Quim Pou and Roser Ortiz; CERM,Centre d'Estudis dels Rius Mediterranis -Universitat de Vic: Marc Ordeix, Núria Sellarés, Francesc Llach and Núria Flor; GESNA Estudis Ambientals: Rafel Rocaspana, Enric Aparicio, Roger Guillem and Pepita Nolla; Hidrologia i Qualitat de l'Aigua: Romero Roig, Iara Jimènez, Miquel Arrabal and Joan Gomà; and David Mateu and Pep Cabanes, IRTA technicians.

We would also like to thank Nikunj Sharma, Cécile Chardon and Louis Jacas who performed the DNA extraction and DNA library preparations at the molecular laboratory of INRA CARRTEL in Thonon-les-Bains (France) and the Genome Transcriptome Facility of INRA in Bordeaux (France) where the HTS was performed. Thanks also to the three anonymous reviewers for very helpful comments.

The authors also acknowledge support from the CERCA Programme/Generalitat de Catalunya. J. Pérez-Burillo acknowledges IRTA-Universitat Rovira i Virgili for his 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. This article is based upon work from COST Action DNAqua-Net (CA15219), supported by the COST (European Cooperation in Science and Technology) program.

References

625 Agència Catalana de l'Aigua (ACA), 2010. Informe de la validació dels punts de referència
626 segons les directrius de la DMA i dels exercicis d'intercalibració europeus. Departament de
627 Medi Ambient i Habitatge, Generalitat de Catalunya. http://aca-web.gencat.cat/aca

628 Almeida, S.F.P., Elias, C., Ferreira, J., Tornés, E., Puccinelli, C., Delmas, F., Dörflinger, G.,
629 Urbanič, G., Marcheggiani, S., Rosebery, J., Mancini, L., Sabater, S., 2014. Water quality
630 assessment of rivers using diatom metrics across Mediterranean Europe: A methods
631 intercalibration exercise. Sci. Total Environ. 476–477:768–776.
632 https://doi.org/10.1016/j.scitotenv.2013.11.144

633 Armbrust, EV., 2009. The life of diatoms in the world's oceans. Nature 459:185–192.
634 https://doi.org/10.1038/nature08057

635 Bailet, B., Bouchez, A., Franc, A., Frigerio, J.-M., Keck, F., Karjalainen, S.-M., Rimet, F.,
636 Schneider, S., Kahlert, M., 2019. Molecular versus morphological data for benthic diatoms
637 biomonitoring in Northern Europe freshwater and consequences for ecological status.
638 Metabarcoding and Metagenomics 3:21–35. https://doi.org/10.3897/mbmg.3.34002

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

642 CEN, 2014. CEN_EN 13946: Water quality - Guidance for the routine sampling and preparation
643 of benthic diatoms from rivers and lakes, pp. 1–22.

644 CEN, 2014b. CEN_EN 14407: Water quality - Water quality Guidance standard for the
645 identification, enumeration and interpretation of benthic diatom samples from running
646 waters, pp. 1–16.

CEN, 2018. CEN/TR 17245: Water quality –Technical report for the routine sampling of benthic diatoms from rivers and lakes adapted for metabarcoding analysis. CEN/TC 230/WG23 – Aquatic Macrophyte and Algae, 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. 10:1–17. https://doi.org/10.3389/fmicb.2019.00653

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

European Commission, 2016. Introduction to the New EU Water Framework Directive. (Available at: http://ec.europa.eu/environment/water/water-framework/info/intro_en.htm)

Jahn, R., Abarca, N., Gemeinholzer, B., Mora, D., Skibbe, O., Kulikovskiy, M., Gusev, E., Kusber, W.H., Zimmermann, J., 2017. *Planothidium lanceolatum* and *Planothidium frequentissimum* reinvestigated with molecular methods and morphology: four new species and the taxonomic importance of the sinus and cavum. Diatom Res. 32:75–107. https://doi.org/10.1080/0269249X.2017.1312548

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., Juggins, S., Guthrie, R., Pritchard, S., Jamieson, J., Rippey, B., Hirst, H., Yallop, M., 2008. Assessment of ecological status in U.K. rivers using diatoms. Freshw. Biol. 53:403–422. https://doi.org/10.1111/j.1365-2427.2007.01903.x

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., Ní Chatháin, B., Pardo, I., Pfister, P., Picinska-Faltynowicz, J., Rosebery, J., Schranz, C., Schaumburg, J., Van Dam, H., 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. https://doi.org/10.1007/s10750-008-9641-4

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 for Water 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

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. https://doi.org/10.1111/1755-0998.1210

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. Freshw. Sci. 33: 349–363. https://doi.org/10.1086/675079

Krammer, K., Lange-Bertalot, H., 1986. 2/1. Bacillariophyceae. 1. Teil: Naviculaceae, In: Ettl, H., Gerloff, J., Heynig, H., Mollenhauer, D. (Eds.), Sübwasserflora von Mitteleuropa. G. Fischer Verlag, Stuttgart, pp 1–876

Krammer, K., Lange-Bertalot, H., 1986. 2/2. Bacillariophyceae. 2. Teil: Bacillariaceae,Epithemiaceae, Surirellaceae, In: Ettl, H., Gerloff, J., Heynig, H., Mollenhauer, D. (Eds.), Sübwasserflora von Mitteleuropa. G. Fischer Verlag, Stuttgart, pp. 1–596

695 Krammer, K., Lange-Bertalot, H., 1991a. 2/3. Bacillariophyceae. 3. Teil: Centrales,
696 Fragilariaceae, Eunotiaceae, In: Ettl, H., Gerloff, J., Heynig, H., Mollenhauer, D. (Eds.),
697 Sübwasserflora von Mitteleuropa. G. Fischer Verlag, Stuttgart, pp. 1–576

698 Krammer, K., Lange-Bertalot, H., 1991b. 2/4. Bacillariophyceae. 4. Teil: Achnanthaceae
699 Kritische Ergänzungen zu *Navicula* (Lineolatae) und *Gomphonema*, In: Ettl, H., Gerloff, J.,
700 Heynig, H., Mollenhauer, D. (Eds.), Sübwasserflora von Mitteleuropa. G. Fischer Verlag,
701 Stuttgart, pp. 1–437

702 Lange-Bertalot, H., Hofmann, G., Werum, M., Cantonati, M., 2017. Freshwater benthic diatoms
703 of Central Europe: Over 800 common species used in ecological assessment. English edition
704 with updated taxonomy and added species.Koeltz Botanical Books, Schmitten-Oberreifenberg,
705 pp. 1–942.

706 Lecointe, C., Coste, M., Prygiel, J., 1993. OMNIDIA—software for taxonomy, calculation of
707 diatom indexes and inventories management. Hydrobiologia. 269:509–13.
708 https://doi.org/10.1007/BF00028048

709 Levkov, Z., 2009. *Amphora* sensu lato, In: Lange-Bertalot, H. (Ed.), (Vol. 5), Diatoms of Europe.
710 A.R.G. Gantner Verlag K.G., pp. 1–916.

711 Mangot, J.F., Domaizon, I., Taib, N., Marouni, N., Duffaud, E., Bronner, G., Debroas, D., 2013.
712 Short-term dynamics of diversity patterns: Evidence of continual reassembly within lacustrine
713 small eukaryotes. Environ. Microbiol. 15:1745–1758. https://doi.org/10.1111/1462-
714 2920.12065

715 Mann, D.G., 1999. The species concept in diatoms. Phycologia, 38:437–495.
716 https://doi.org/10.2216/i0031-8884-38-6-437.1

717 Mann, D.G., Crawford, R.M., Round, F.E., 2016. Bacillariophyta, In: Archibald, J.M., Simpson,
718 A.G.B., Slamovits, C.H., Margulis, L., Melkonian, M., Chapman, D.J., Corliss, J.O. (Eds.),

719 Handbook of the Protists. Springer, Cham, New York, pp.1–62. https://doi.org/10.1007/978-3-
720 319-32669-6_29-1

721 Mortágua, A., Vasselon, V., Oliveira, R., Elias, C., Chardon, C., Bouchez, A., Rimet, F., João Feio,
722 M., F.P. Almeida, S., 2019. Applicability of DNA metabarcoding approach in the bioassessment
723 of Portuguese rivers using diatoms. Ecol. Indic. 106:105470.
724 https://doi.org/10.1016/j.ecolind.2019.105470

725 Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for
726 similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.
727 https://doi.org/10.1016/0022-2836(70)90057-4

728 Novis, P.M., Braidwood, J., Kilroy, C., 2012. Small diatoms (Bacillariophyta) in cultures from the
729 Styx River, New Zealand, including descriptions of three new species. *Phytotaxa*. 64:11-45.
730 http://doi.org/10.11646/phytotaxa.64.1.3

731 Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A.,
732 Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M.J., Filipe, A.F., Fornaroli, R., Graf, W.,
733 Herder, J., van der Hoorn, B., Jones, J.I., Sagova-Mareckova, M., Moritz, C., Barquín, J., Piggott,
734 J.J., Pinna, M., Rimet, F., Rinkevich, B., Sousa-Santos, C., Specchia, V., Trobajo, R., Vasselon, V.,
735 Vitecek, S., Zimmermann, J., Weigand, A., Leese, F., Kahlert, M., 2018. The future of biotic
736 indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of
737 aquatic ecosystems. Sci. Total Environ. 637–638:1295–1310.
738 https://doi.org/10.1016/j.scitotenv.2018.05.002

739 Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A., Bouchez, A.,
740 2016. R-Syst::diatom: An open-access and curated barcode database for diatoms and
741 freshwater monitoring. Database 2016:1–21. https://doi.org/10.1093/database/baw016

742 Rimet, F., Vasselon, V., A.-Keszte, B., Bouchez, A., 2018. Do we similarly assess diversity with
743 microscopy and high-throughput sequencing? Case of microalgae in lakes. Org. Divers. Evol.
744 18:51–62. https://doi.org/10.1007/s13127-018-0359-5

745 Rimet, F., Gusev, E., Kahlert, M., Kelly, M.G., Kulikovskiy, M., Maltsev, Y., Mann, D.G.,
746 Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., Bouchez, A., 2019. Diat.barcode,
747 an open-access curated barcode library for diatoms. Sci. Rep. 9:1–12.
748 https://doi.org/10.1038/s41598-019-51500-6

749 Rivera, S.F., Vasselon, V., Ballorain, K., Carpentier, A., Wetzel, C.E., Ector, L., Bouchez, A.,
750 Rimet, F., 2018a. DNA metabarcoding and microscopic analyses of sea turtles biofilms:
751 Complementary to understand turtle behavior. PLoS One 13:1–20.
752 https://doi.org/10.1371/journal.pone.019577

753 Rivera, S.F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., Rimet, F., 2018b.
754 Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment.
755 Hydrobiologia 807:37–51. https://doi.org/10.1007/s10750-017-3381-2

756 Rivera, S.F., Vasselon, V., Bouchez, A., Rimet, F., 2020. Diatom metabarcoding applied to large
757 scale monitoring networks: Optimization of bioinformatics strategies using Mothur software.
758 Ecol. Indic. 109:105775. https://doi.org/10.1016/j.ecolind.2019.105775

759 Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski,
760 R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn,
761 D.J., Weber, C.F., 2009. Introducing mothur: Open-source, platform-independent, community-
762 supported software for describing and comparing microbial communities. Appl. Environ.
763 Microbiol. 75:7537–7541. https://doi.org/10.1128/AEM.01541-09

764 Smetacek, V., 1999. Diatoms and the ocean carbon cycle. Protist. 150:25–32.
765 https://doi.org/10.1016/s1434-4610(99)70006-4

766 Trobajo, R., Rovira, L., Ector, L., Wetzel, C.E., Kelly, M., Mann, D.G., 2013. Morphology and
767 identity of some ecologically important small *Nitzschia* species. Diatom Research. 28:37–59.
768 https://doi.org/10.1080/0269249X.2012.734531

769 Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., Bouchez, A., 2017a. Application of high-
770 throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction
771 methods matter? Freshw. Sci. 36:162–177. https://doi.org/10.1086/690649

772 Vasselon, V., Rimet, F., Tapolczai, K., Bouchez, A., 2017b. Assessing ecological status with
773 diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island,
774 France). Ecol. Indic. 82:1–12. https://doi.org/10.1016/j.ecolind.2017.06.024

775 Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., Tapolczai, K.,
776 Domaizon, I., 2018. Avoiding quantification bias in metabarcoding: Application of a cell
777 biovolume correction factor in diatom molecular biomonitoring. Methods Ecol. Evol. 9:1060–
778 1069. https://doi.org/10.1111/2041-210X.12960

779 Vasselon V., Rimet, F., Bouchez, A., 2018b. "Rsyst::diatom_rbcl_align_312bp database: a
780 database adapted to DNA metabarcoding (version v7: 23-02-2018).
781 https://doi.org/10.15454/HYRVUH, Portail Data Inra, V1

782 Vasselon, V., Rimet, F., Domaizon, I., Monnier, O., Reyjol, Y., Bouchez, A., 2019. Assessing
783 pollution of aquatic environments with diatoms' DNA metabarcoding: experience and
784 developments from France water framework directive networks. Metabarcoding and
785 Metagenomics 3: e39646. https://doi.org/10.3897/mbmg.3.39646

786 Visco, J.A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., Pawlowski, J., 2015.
787 Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data.
788 Environ. Sci. Technol. 49:7597–7605. https://doi.org/10.1021/es506158m

789 Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naïve Bayesian classifier for rapid
790 assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol.
791 73:5261–5267. https://doi.org/10.1128/AEM.00062-07

792 Zelinka M., Marvan P., 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit
793 fliessender Gewässer. Archiv fur Hydrobiologie. 57:389–407.

794 Zgrundo, A., Lemke, P., Pniewski, F., Cox, E.J., Latala, A., 2013. Morphological and molecular
795 phylogenetic studies on *Fistulifera saprophila*. Diatom Research 28: 431-443.
796 https://doi.org.10.1080/0269249X.2013.833136

797 Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., Gemeinholzer, B., 2015. Metabarcoding vs.
798 morphological identification to assess diatom diversity in environmental studies. Mol. Ecol.
799 Resour. 15:526–542. https://doi.org/10.1111/1755-0998.12336

800

801

802 **Figure captions**

803 Figure 1 a) Location of Catalonia (Spain) and b) river sites sampled in the internal (white) and
804 interregional (gray) Catalan hydrographic basins.

805 Figure 2. Correlation of IPS values calculated from LM (x axis) and HTS (y axis) inventories
806 considering the total 162 samples. Pearson's coefficient (R) and p-value are given.

807 Figure 3. Species sensitivity analysis (left for LM data, right for the HTS data) calculated by the
808 "leave-one-out" method (see Material and Methods) showing the IPS contributions (X axes) of
809 the 35 most abundant species in the LM counts. Species are ordered according to the average
810 of the IPS contributions for the LM method, from the species with the most positive average
811 (top) to those with the most negative average (bottom). Samples are coloured according to the
812 ecological status class given by the whole diatom assemblage. Note that, for the HTS set, some
813 species (empty symbols) have zero contribution; this is because there are no sequences for
814 these species in the reference database. The IPSS and IPSV values for each species are given
815 after the species name (e.g. 5;1 for *Achanthidium minutissimum* means IPSS=5, IPSV=1).

816 Figure 4: Relative species contributions to the IPS value in the 10 critical sites (those sites
817 whose status alters from Good/High by LM to Moderate/Poor/Bad by HTS). For each species
818 the number of critical sites in which it was ranked the first, second, third or fourth most
819 important contributor to the IPS score (left negatively, right positively), as assessed by the
820 leave-one-out method, is given for both LM and (uncorrected) HTS. X axes: number of critical
821 sites.

822 Figure 5: Relative abundance comparisons between LM valve counts (x axis) and HTS reads (y
823 axis) for methods of selected species. Cross symbol in black correspond to critical sites and
824 circles in grey to non-critical sites. Species represented are the following: a) *Fistulifera*

saprophila b) *Mayamaea permitis* c) *Ulnaria ulna* d) *Achnanthidium minutissimum* e) *Nitzschia inconspicua* e) *Planothidium frequentissimum*

Table captions

Table 1. Comparison between ecological status classes obtained from HTS and LM approaches. Cells in light grey represent the number of sites assigned to the same ecological status class by both methods. Dark grey cell represents the number of sites that cross the critical threshold between acceptable and unacceptable ecological status (i.e. those sites whose status alters from Good/High by LM to Moderate/Poor/Bad by HTS).

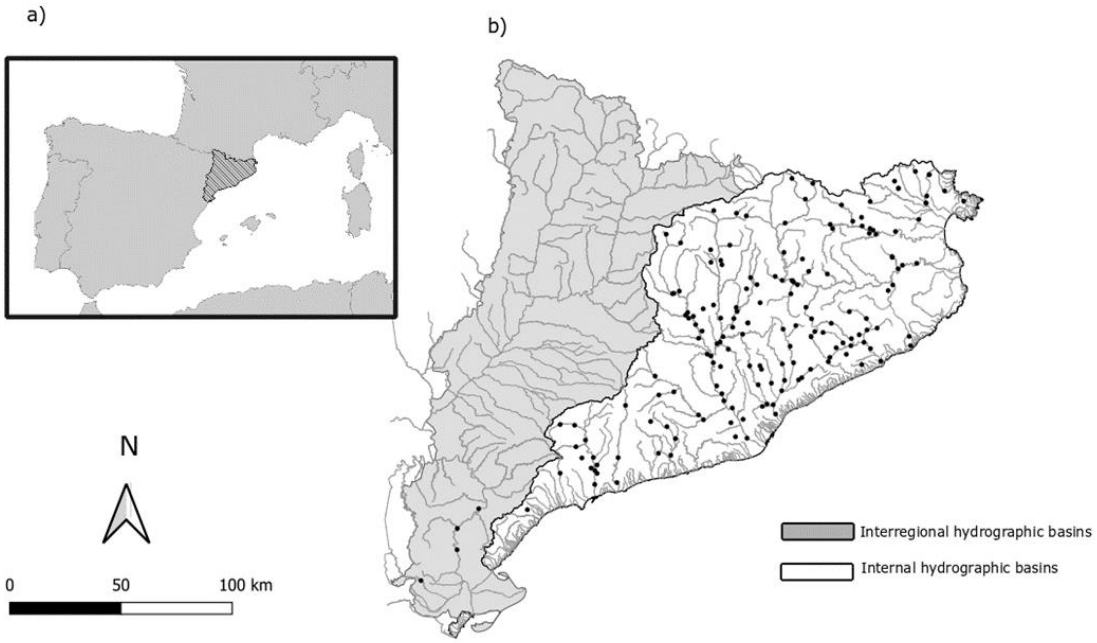
Supplementary information

Supplementary Fig 1. Relative abundance (%) of the most common species (relative abundances > 1%) recorded for the LM and (uncorrected and corrected) HTS inventories. * represents those species not presented in the reference library

Supplementary Fig 2. Graphs comparing the species negative contributions to HTS-calculated IPS scores when CFs are applied (grey) or without CFs (black) in the five extra critical sites resulting when CFs were applied. Only the five species that most negatively contributed to IPS without applying CFs are represented.

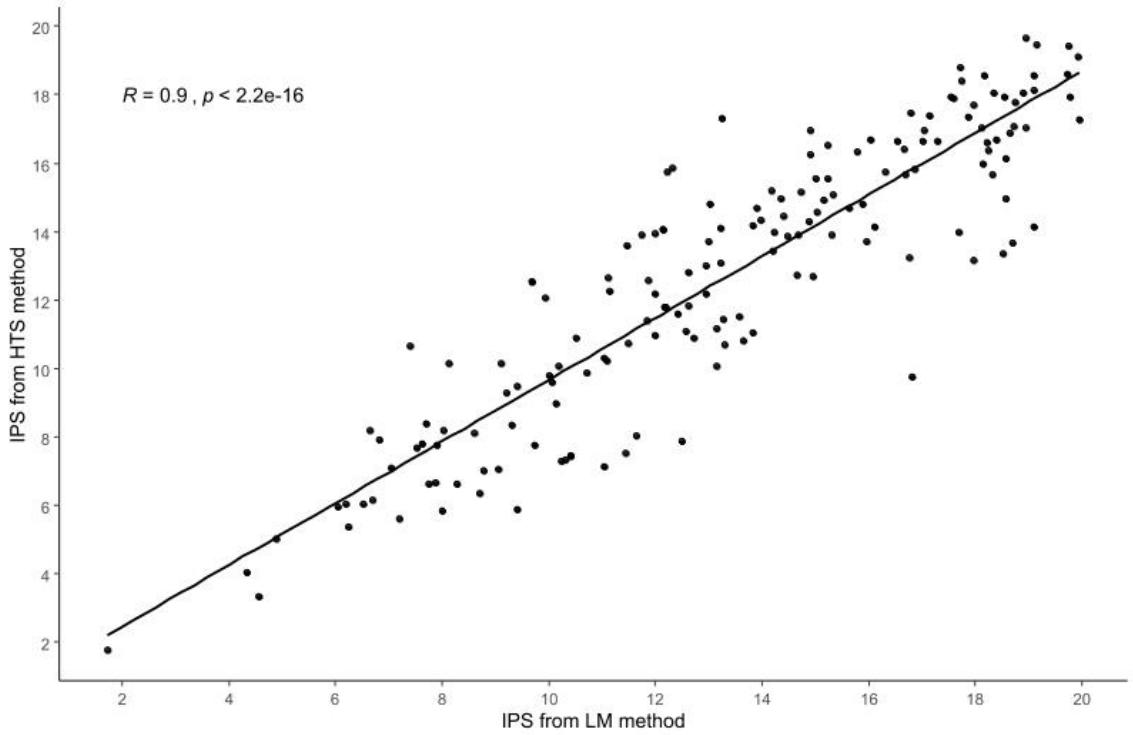
Supplementary Fig 3. Relative abundance (%)of *Planothidium frequentissimum* (only identified with LM) and *P. victorii* (only identified with HTS) throughout the 162 samples examined.

Supplementary Data. Excel file containing HTS reads and IPS contribution of the species obtained by the sensitivity analysis for both methods through the 162 samples analyzed. HTS data was normalized to 3610 reads.



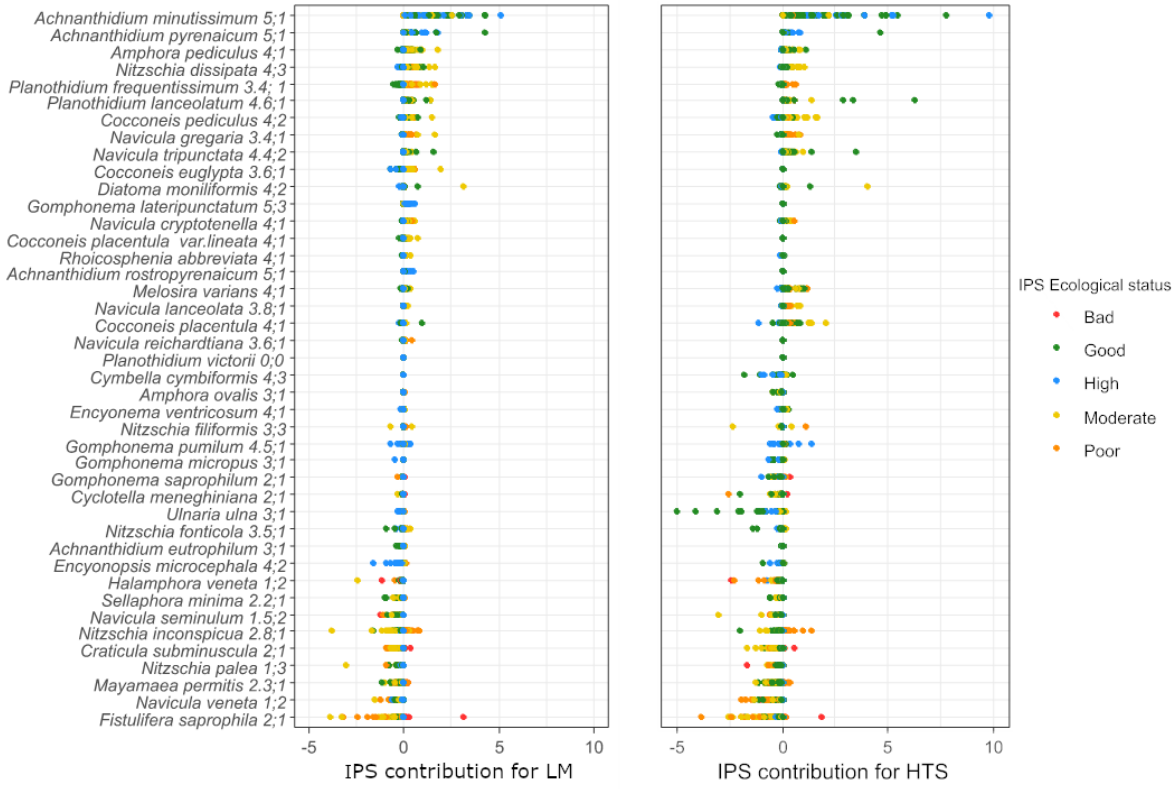
849
850 Fig. 1

851



852
853 Fig. 2

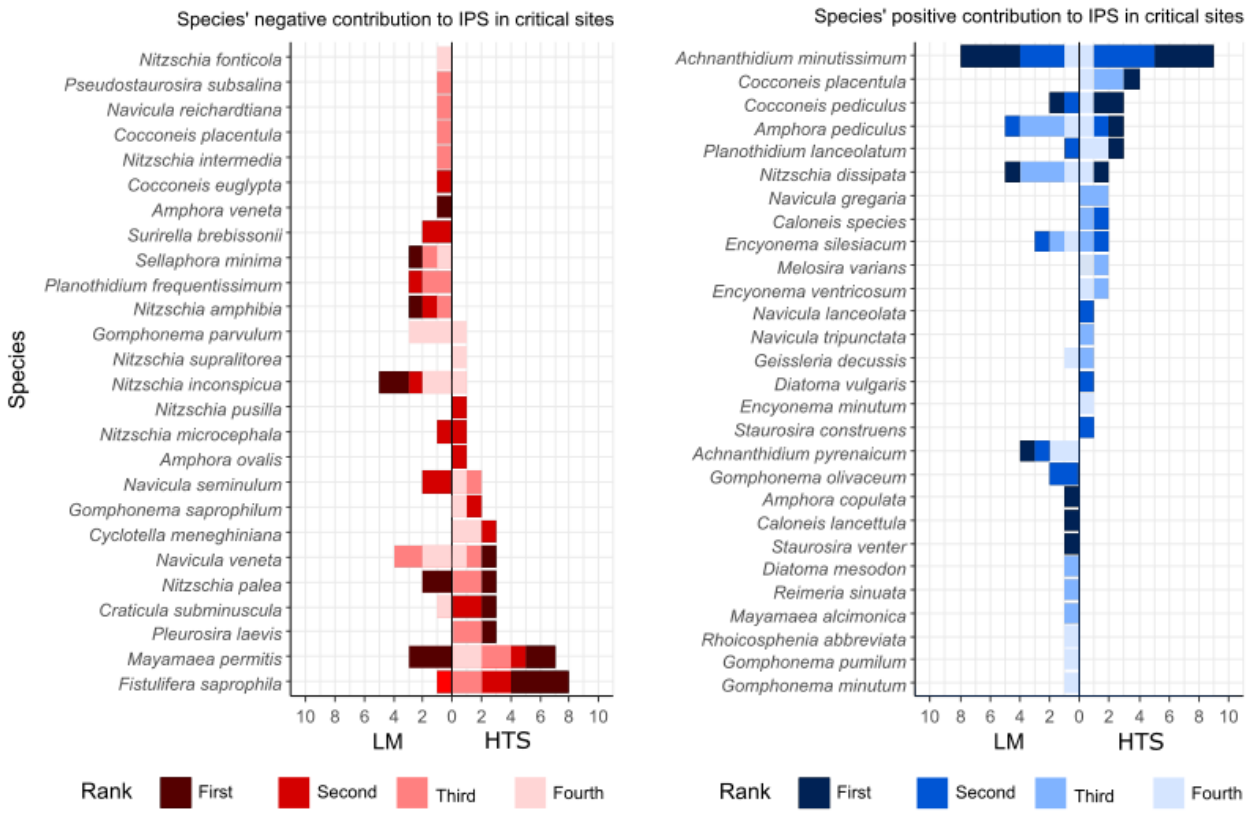
854



855
856 Fig. 3

857

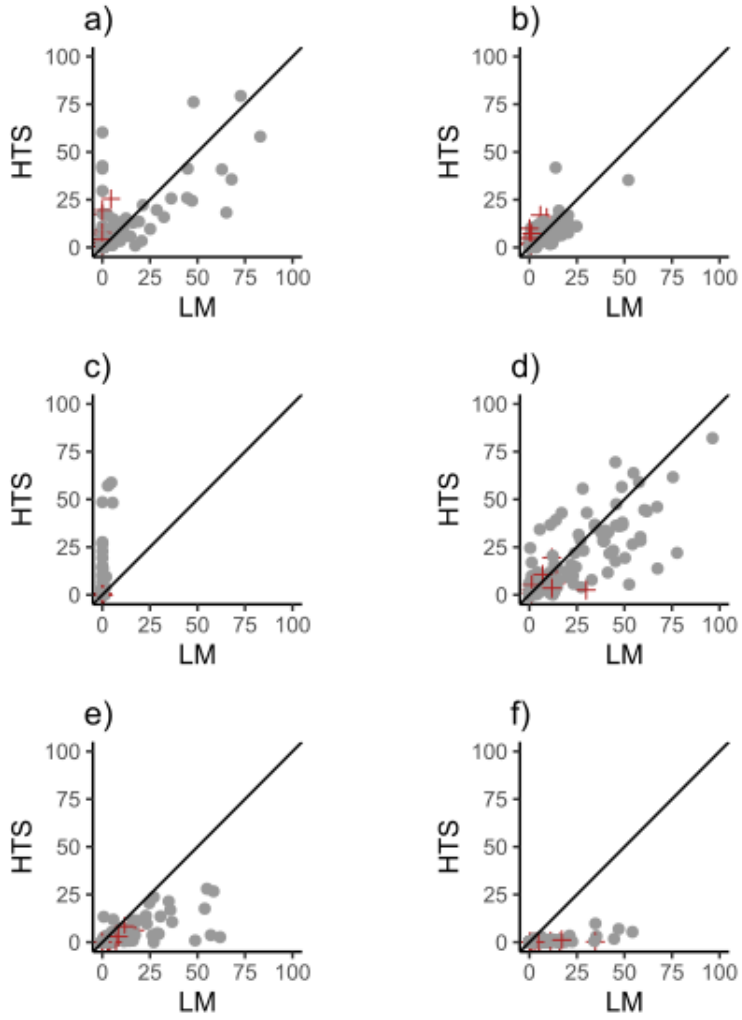
858



859
860 Fig. 4

861

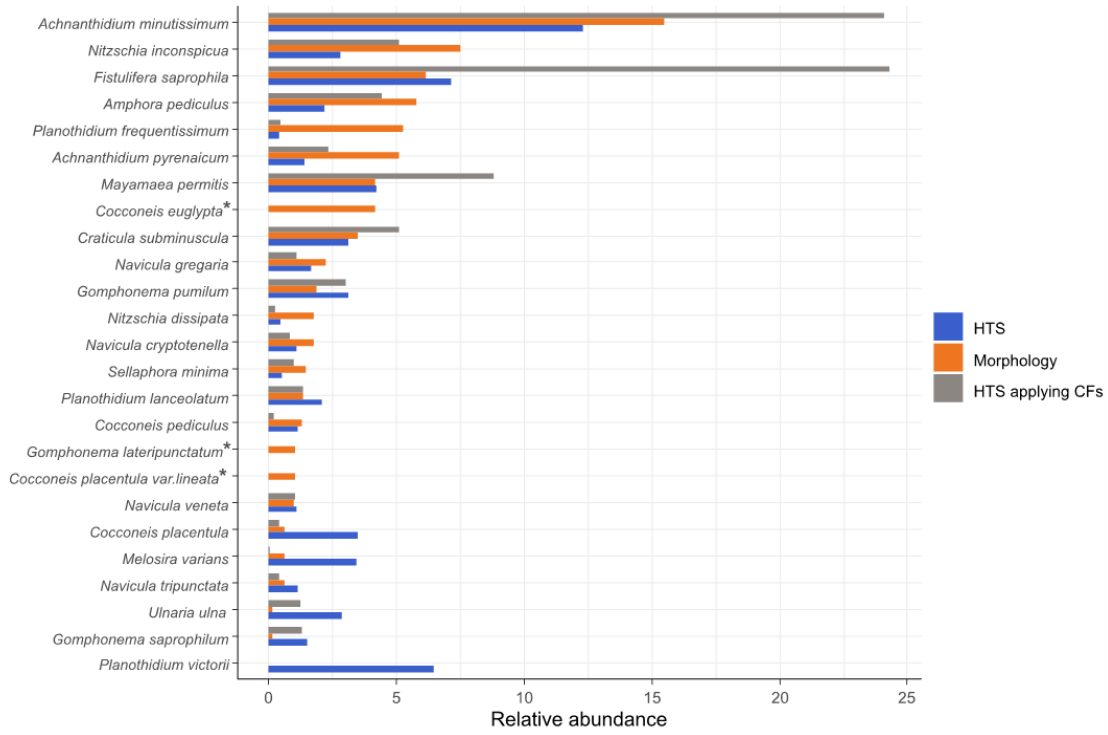
862



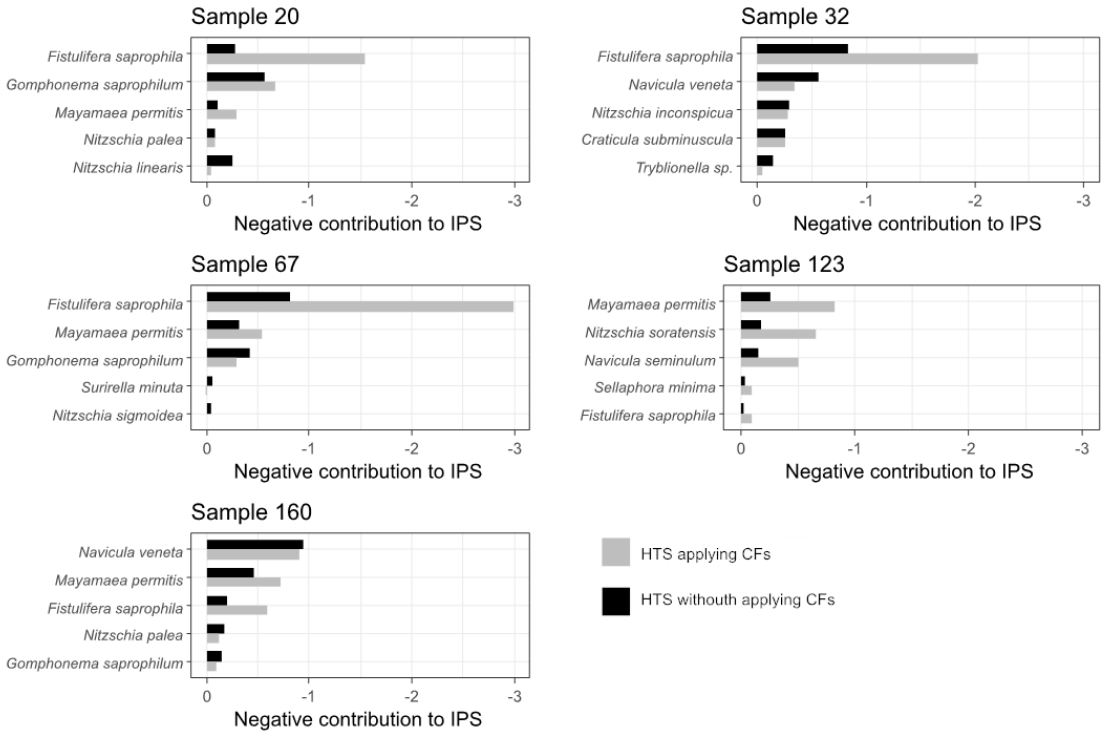
863

864 Fig. 5

865



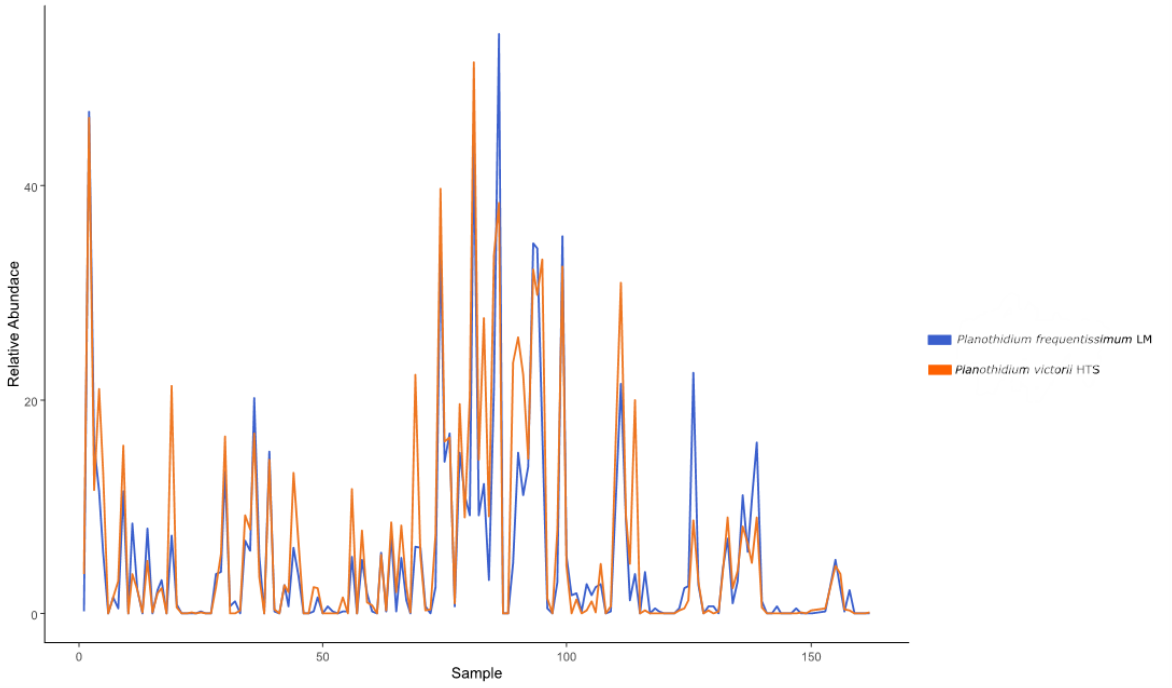
866
867 Supplementary Fig. 1



868

869 Supplementary Fig. 2

870



871

872 Supplementary Fig. 3

873

874

HTS inventory

LM inventory	Bad	Poor	Moderate	Good	High
Bad	4	0	0	0	0
Poor	0	21	2	0	0
Moderate	0	12	28	7	0
Good	0	0	10	36	2
High	0	0	0	16	24

875

876 Table. 1