DNA metabarcoding reveals differences in distribution patterns and ecological preferences among genetic variants within some key freshwater diatom species

Javier Pérez-Burillo[1,2], Rosa Trobajo[1*], Manel Leira[3,4], François Keck[5,6], Frédéric Rimet[7,8], Javier Sigró[2] & David G. Mann[1,9]

[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 Ràpita, Tarragona, Spain

[2]Center for Climate Change (C3), Departament de Geografia, Universitat Rovira i Virgili, C/ Joanot Martorell 15, E43500, Vila-seca, Tarragona, Spain

[3] BioCost Research Group, Facultade de Ciencias and Centro de Investigacións Científicas Avanzadas (CICA), Universidade de A Coruña, 15071, A Coruña, Spain

[4] Biodiversity and Applied Botany Research Group, Departmento de Botánica, Facultade de Bioloxía, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain;

[5]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

[6]Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland

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

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

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

*corresponding author: rosa.trobajo@irta.cat

Abstract

Our study evaluates differences in the distribution and ecology of genetic variants within several ecologically important diatom species that are also key for Water Framework Directive monitoring of European rivers: *Fistulifera saprophila* (FSAP), *Achnanthidium minutissimum* (ADMI), *Nitzschia inconspicua* (NINC) and *Nitzschia soratensis* (NSTS). We used DADA2 to infer amplicon sequence variants (ASVs) of a short *rbcL* barcode in 531 environmental samples from biomonitoring campaigns in Catalonia and France. ASVs within each species showed different distribution patterns. Threshold Indicator Taxa ANalysis revealed three ecological groupings of ASVs in both ADMI and FSAP. Two of these in each species were separated by opposite responses to calcium and conductivity. Boosted regression trees additionally showed that both variables greatly influenced the occurrence of these groupings. A third grouping in FSAP was characterized by a negative response to total organic carbon and hence was better represented in waters with higher ecological status than the other FSAP ASVs, contrasting with what is generally assumed for the species. In the two *Nitzschia* species, our analyses confirmed earlier studies: NINC preferred higher levels of calcium and conductivity. Our findings suggest that the broad ecological tolerance of some diatom species results from overlapping preferences among genetic variants, which individually show much more restricted preferences and distributions. This work shows the importance of studying the ecological preferences of genetic variants within species complexes, now possible with DNA metabarcoding. The results will help reveal and understand biogeographical distributions and facilitate the development of more accurate biological indexes for biomonitoring programmes.

Keywords. ASV, environmental DNA, Water Framework Directive, rbcL, ecological preferences, species distribution

1. Introduction

Diatoms play a crucial role in aquatic systems due, amongst other things, to their importance in food webs and biogeochemical cycling and their great contribution to carbon fixation (Armbrust, 2009; Mann, 1999; Smetacek, 1999). They are also widely used as ecological indicators in palaeoenvironmental studies and biomonitoring programmes. For example, in European rivers, it is compulsory (within the European Union) to monitor benthic diatom communities (Water Framework Directive [WFD], Directive 2000/60/EC, 2000) because of their rapid and specific response to environmental changes, great diversity, and ubiquitous distribution, and the availability of information on the ecological preferences of many species. However, it has become evident in the last two decades that many of these species are complexes of genetic variants (e.g. Pinseel et al. 2017; Souffreau et al. 2013). These often show scarcely discernible or no morphological differences (they are “cryptic”) and therefore it is difficult or impossible to determine their geographical distributions and ecological preferences using traditional methods based on microscopical identifications. Therefore, the significance of this intraspecific variation is still not clear: although It is suggested that closely related diatoms often share a similar ecology (Keck et al., 2016a, b, 2018b), it is also evident that they can differ (Pinseel et al., 2017; Poulíčková et al. 2008, 2017; Rynearson et al., 2006).

DNA metabarcoding has recently been developed for biomonitoring the ecological status of rivers (e.g. Kelly et al. 2020; Mortágua et al., 2019; Pérez-Burillo et al. 2020; Rivera et al., 2020; Vasselon et al., 2017b) and it has proved as well to be a reliable and efficient method for surveying species diversity from environmental samples (Deiner et al., 2017; Malviya et al. 2016; Piredda et al., 2017). DNA metabarcoding also offers a way to study the significance of genetic variants within species, especially following the development of bioinformatic pipelines such as DADA2 (Callahan et al., 2017), which use a denoising algorithm to remove sequencing artifacts and generate ‘amplicon sequence variants’ (ASVs); these are believed to be real DNA sequences that were present in the original environmental samples. A recent

example of using an ASV approach in diatoms was by Tapolczai et al. (2021), where they assessed the responses of river diatom communities to agricultural land use in Hungary; in some cases, they reported different ecological preferences among ASVs from the same species. However, despite the clear potential of ASV-based metabarcoding approaches, there do not appear to have been any studies to date that have used a large dataset to examine the ecology and distribution of genetic variants and hence to elucidate their significance.

The aim of this work was therefore to study the distribution and ecological preferences of different ASVs within selected species complexes of diatoms. For this we chose two groups that are ecologically important and have been shown to be key for the WFD (Pérez-Burillo et al., 2020): *Achnanthidium minutissimum* sensu lato (said by Potapova and Hamilton, 2007, to be "one of the most frequently occurring diatoms in freshwater benthic samples globally") and *Fistulifera saprophila*. Both are very small-celled species that are difficult to treat morphologically. In addition, we selected *Nitzschia inconspicua*, because Sanger sequencing has already demonstrated a complex pattern of genetic and physiological variation within it (Rovira et al., 2015), and *N. soratensis*, which is so similar to *N. inconspicua* in the light microscope that identifying the two species and determining their ecological separation is highly challenging (Kelly et al., 2015). More specifically, we asked 1) do genetic variants (ASVs) within a species complex have similar geographical distributions within the study area? 2) Do ASVs within a species complex have the same ecological preferences or do they differ? 3) If there are differences in the ecological preferences of genetic variants within a species, do these correlate with their phylogeny?

To answer these questions, we used a large molecular dataset extracted from environmental samples collected in several river biomonitoring campaigns in contiguous areas of France and Catalonia (NE Spain). For evaluating ecological preferences and the spatial distributions of ASVs, we performed Threshold Indicator Taxa Analyses (TITAN) and Boosted Regression Trees (BRT) analyses, since both methods have been successfully applied in

morphological and metabarcoding studies addressing stressor-response and species distribution models (Lanzén et al., 2020; Smucker et al., 2020; Soininen et al., 2018, Wagenhoff et al., 2017).

2. Material and Methods

2.1 Study site and diatom sampling

The dataset used in this study consisted of 610 benthic diatom samples collected from both Catalan and French biomonitoring networks. Samples were originally taken as a part of the 2017 Catalan biomonitoring programme and two French monitoring campaigns held in 2016 and 2017. The hydrographic area of Catalonia is divided into internal and interregional hydrographic basins. The internal basins comprise a total of eleven main rivers, the basins of the rivers Llobregat and Ter being the most extensive, and the interregional basins cover the Catalan sections of the rivers Ebro, Garona (Garonne in French) and Xúquer. The French monitoring network area corresponds to seven main basins (Adour–Garonne, Artois–Picardie, Loire–Bretagne, Rhin–Meuse, Rhône–Méditerranée, Corse, and Seine-Normandie) of which the largest belong to the rivers Loire, Rhône, Seine and Garonne (Supplementary Fig. 1).

All Catalan sites were sampled for periphyton between April and July of 2017 following standard procedures (CEN, 2014). French sites were sampled between February and December and between February and October, for the campaigns held in 2016 and 2017 respectively, and followed French NFT 90 354 (AFNOR, 2007) and European (CEN, 2014) standards. At each site, diatoms were collected from at least five stones by brushing their upper surfaces using a toothbrush. The resulting samples were preserved by adding ≥ 90% ethanol (to a final concentration of 70%) and used for DNA metabarcoding analysis following the recommendations of the technical report of the European Committee for Standardization (CEN, 2018).

126

127 *2.2 Physicochemical and biotic parameters*

128 Physicochemical parameters that constituted the environmental dataset used for French and
129 Catalan river sites were obtained from the online "Naïades"
130 (http://www.naiades.eaufrance.fr/) and "SDIM" (http://aca-web.gencat.cat/sdim21/) water
131 quality datasets. Environmental parameters selected in this study were ammonium (NH_4^+;
132 mg/L), bicarbonates (HCO_3^-; mg/L), calcium (mg/L), total organic carbon (TOC; mg/L),
133 conductivity (µS/cm), nitrates (NO_3^-; mg/L), orthophosphates (PO_4^{3-}; mg/L), pH, sulphates
134 (SO_4^{2-}; mg/L), water temperature (°C) and altitude (m) (Table 1). The measures selected for
135 these parameters corresponded to the mean of all the records available for a period of 80 days
136 preceding and 10 days following the biological sampling. The diatom indices IBD ("Indice
137 Biologique Diatomées") and IPS ("Indice de Polluosensibilité spécifique") were retrieved
138 respectively for French and Catalan rivers sites analysed.

139

140 *2.3 DNA extraction, PCR amplification and high-throughput sequencing (HTS)*

141 The procedures for DNA extraction, PCR amplification and HTS for French and Catalan rivers
142 are described in Rivera et al. (2020) and Pérez-Burillo et al. (2020), respectively. Briefly, DNA
143 extraction of French samples from the 2016 campaign was performed using GenElute TM-LPA
144 protocol, while the Macheray–Nagel NucleoSpin® soil kit (MN-Soil) protocol was followed for
145 DNA extraction of Catalan and French samples from the 2017 campaigns. A short *rbcL* region of
146 312 bp constituted the DNA marker and this was amplified by PCR using an equimolar mix of
147 the modified versions of the primers Diat_rbcL_708F (forward) and R3 (reverse) given by
148 Vasselon et al. (2017b). Four Illumina Miseq runs were performed for sequencing separately
149 the French (3 runs) and Catalan (1 run) samples. In order to prepare the HTS libraries using a 2-
150 step PCR strategy, half of P5 and P7 Illumina adapters were included to the 5' end of the

151 forward and reverse primers respectively. Adapter sequences used were
152 CTTTCCCTACACGACGCTCTTCCGATCT (P5) and GGAGTTCAGACGTGTGCTCTTCCGATCT (P7) for
153 French samples and TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG (P5) and
154 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA (P7) for Catalan samples.

155 PCR1 reactions for each DNA sample were performed in triplicate using 1 µL of the
156 extracted DNA in a final volume of 25 µL. Conditions and the reaction mix of the PCR1 followed
157 the procedure described in Vasselon et al. (2017b) For each sample, the three PCR1 replicates
158 were pooled and sent for sequencing to "Plateforme Génome Transcriptome" (PGTB,
159 Bordeaux, France) or "GenoToul Genomics and Transcriptomics" (GeT-PlaGe, Auzeville,
160 France), where the PCR1 products were purified and used as template for a second round of
161 PCR (PCR2), with Illumina tailed primers targeting the half of P5 and P7 adapters. Finally, all
162 generated amplicons were dual indexed and pooled for sequencing on an Illumina MiSeq
163 platform using the V3 and V2 paired-end sequencing kits (250 bp × 2) for the French and
164 Catalan samples respectively.

165 The influence of DNA extraction methods on the diatom inventory produced by DNA
166 metabarcoding has been evaluated by Vasselon et al. (2017a). In this study, the authors
167 evaluated 5 different extraction methods, including the two methods used in our study. They
168 found some slight differences in relative abundance between methods for some particular
169 species, but the differences in community composition caused were far less than the
170 differences attributable to habitat. Importantly for the current analyses, the slight differences
171 did not affect species richness ("Regardless of taxonomic level (OTU or species), the taxonomic
172 composition of the community represented in the extracts was not affected by DNA extraction
173 methods…": Vasselon et al. 2017a). Furthermore, (1) the taxa contributing more than >1% of
174 the dissimilarities between diatom communities obtained with the GenElute and MN-Soil
175 protocols did not include either *Achnanthidium* or *Fistulifera* (Vasselon et al. 2017a, Table 3),
176 and (2) all our comparisons were made among ASVs belonging to the same species complex

177 and hence of diatoms with very similar physical characteristics (frustule shape, size and
178 robustness). Thus, we can expect that the different extraction methods used for the French
179 2016 and 2017 datasets will not have greatly affected either the presence/absence of ASVs
180 (especially since a high threshold of abundance was set for inclusion in the analyses) or relative
181 abundances across the combined dataset. Comparisons across wider ranges of species (e.g. all
182 *Achnanthidium*, all *Navicula*, etc) might have been more seriously affected.

183 *2.4. Bioinformatic analysis*

184 The sequencing facilities performed the demultiplexing of all the samples providing two fastq
185 files per sample, one corresponding to forward reads (R1) and one to reverse reads (R2). All
186 the demultiplexed Miseq reads were treated together using the R package *DADA2*, following
187 the method described by Callahan et al. (2016). Primers were removed from the R1 and R2
188 reads using cutadapt (Martin, 2011). The resulting R1 and R2 reads were truncated to 200 and
189 170 nucleotides respectively, based on to their quality profile (median quality score < 30) and
190 those reads with ambiguities or an expected error (maxEE) higher than 2 were discarded. The
191 DADA2 denoising algorithm was applied to determine an error rates model in order to infer
192 amplicon sequence variants (ASVs); ASVs detected as chimeras were discarded using the
193 function "removeBimeraDenovo" implemented in DADA2. Finally, the taxonomic affiliation of
194 the ASVs was determined using the database "A ready-to-use database for DADA2:
195 Diat.barcode_rbcL_312bp_DADA2" (Chonova et al., 2020), which is derived from the curated
196 diatom reference library Diat.barcode v9 (Rimet et al., 2019, available at
197 https://www6.inra.fr/carrtel-collection_eng/Barcoding-database and at
198 https://data.inrae.fr/file.xhtml?persistentId=doi:10.15454/TOMBYZ/IEGUXB&version=10.0),
199 and the naïve Bayesian classifier method (Wang et al., 2007); 50% was set as the minimum
200 confidence threshold (the default in DADA2). In this study we focused on ASVs that were
201 assigned by the pipeline to *Nitzschia inconspicua*, *N. soratensis*, *Achnanthidium minutissimum*
202 and *Fistulifera saprophila*. Of these, we retained for subsequent analyses only those with ≥

1000 reads and occurring in ≥ 2 samples with environmental data available, in order to remove rare ASVs and residual sequencing artifacts. The ASVs were numbered according to the rank order of their abundance; so, for example, *A. minutissimum* ASV6 was the sixth most abundant sequence in the whole dataset.

2.5. Phylogenetic analyses

Phylogenetic analyses were performed in order to 1) elucidate the phylogeny of the different ASVs obtained from *Nitzschia inconspicua, N. soratensis, Achnanthidium minutissimum* and *Fistulifera saprophila*, and 2) assess the taxonomic assignation obtained after executing the bioinformatics analyses by examining the phylogenetic relatedness between the ASVs and curated reference sequences from Diat.barcode v9 (together with some other, more recent sequences present in GenBank: https://ncbi.nlm.nih.gov/). For this purpose, maximum likelihood trees were constructed using ASVs and the reference sequences. A first tree included reference sequences and ASVs classified into *N. inconspicua* and *N. soratensis* species, while a second and a third used those ASVs and reference sequences classified into *A. minutissimum* and *F. saprophila* respectively. All three analyses were performed using raxmlGUI with the GRT-Gamma model, with 1000 replicates for the bootstrap analyses. Reference sequences and ASVs used for building each of the three trees were previously aligned using the Muscle alignment algorithm (Edgar, 2004) in MegaX software (Kumar et al., 2018). All the three trees calculated were drawn using iTOL (https://itol.embl.de) (Letunic et al., 2019).

2.6. Statistical analyses

2.6.1 Spatial variables

In order to study spatial distribution patterns of ASVs, Moran's eigenvector maps (MEMs) were used on sampling sites' latitude and longitude to generate explanatory variables that represent spatial patterns at different scale and can be used in canonical analysis. (Dray et al.,2006). MEMs are produced by the diagonalization of a spatial weighting matrix, which is obtained as the Hadamard product of a connectivity matrix by a similarity matrix. The connectivity matrix was based on Gabriel's graph geometrical connection scheme due to the non-regular distribution of the sampling sites (Legendre and Legendre, 2012). The R package *adespatial* (Dray et al., 2020) was used for calculating MEMs.

2.6.2 Redundancy analyses

ASV abundance data were Hellinger transformed and all environmental variables except pH were standardized following $X_{st} = (X - \mu)/SD$. Variance inflation factors (VIFs) were calculated to check the presence of collinearities among environmental variables and those variables with VIF >10 were removed to avoid the impact of collinearity. Forward selection with two stopping criteria (alpha significance level and adjusted coefficient of multiple determination, Blanchet et al., 2008) was applied separately on environmental and MEMs sets of variables. Two redundancy analyses (RDA) models were performed in order to analyse separately the relationships between the selected environmental and spatial variables (MEMs) and the ASVs. R packages *adespatial* (Dray et al., 2020) and *vegan* (Oksanen et al., 2020) were used for performing forward selection and RDA models respectively.

2.6.3 TITAN analyses

Threshold indicator taxa analyses (TITAN) were conducted in order to characterize ASV-specific responses for each environmental variable. TITAN handles multiple response variables (ASVs) but only one explanatory variable (i.e. environmental variables) at each analysis and it detects

252 change points, which are the values of the environmental gradient where the greatest change
253 in taxon abundance and frequency occurs within the observed samples. TITAN standardizes
254 the magnitude of responses as z scores in order to facilitate cross-taxa comparison. Z scores
255 reflect the type of response, positive (+ z scores) or negative (- z scores), of a particular taxon
256 (ASV in this case) along the environmental gradient and the sum of the z scores (sum z) gives
257 information about the assemblage responses, either negative (sum –z) or positive (sum +z),
258 along the gradient, the maximum z score occurring at the point at which change in assemblage
259 composition is greatest (Baker & King, 2010). We conducted TITAN analyses for each of the
260 environmental parameters and using ASV relative abundance. Number of permutations was
261 set to 250, number of bootstrap replicates used was 500, the minimum number of
262 observations required on each side of a candidate change point was 5 and the TITAN filtering
263 metrics of uncertainty "purity" and "reliability", used to separate reliable responders from
264 stochastic noise along the gradient, were set to 0.95. Z scores obtained for those ASVs whose
265 responses fulfilled purity and reliability criteria for at least 4 environmental variables were
266 hierarchically clustered and visualized through heatmaps in order distinguish groups of ASVs
267 with similar response patterns for environmental data. For that, Euclidean distance and ward.
268 D functions (Ward, 1963) were used to compute dissimilarity distance and hierarchical
269 clustering respectively. On the other hand, Kruskal–Wallis (Hollander and Wolfe, 1973) tests
270 with post hoc Dunn's test (Dunn, 1964) were performed to determine environmental data
271 statistically significant ($p < 0.05$) among the sites where species and ecological groupings
272 occurred. We used the implementation available in the R packages *TITAN2* (Baker et al., 2019),
273 *gplots* (Warnes et al., 2020), *stats* (R core team, 2020) and *dunn.test* (Dinno, 2017) to conduct
274 the TITAN analyses, heatmaps, Kruskal–Wallis test and Dunn's test respectively.

275

276 2.6.4 Boosted regression tress

277 Relationships of the groups of ASVs, defined after TITAN analysis, with environmental variables
278 were additionally evaluated using boosted regression trees (BRT). BRT is a machine learning
279 model that uses a boosting algorithm to combine large numbers of decision trees for
280 improving model accuracy (Elith et al., 2008). BRT handles multiple explanatory variables
281 (environmental variables) but only one response variable (groups of ASVs in our case). It
282 estimates the relative importance of environmental variables on the basis of the number of
283 times that a variable is selected and the extent to which it improves the model (Friedman,
284 2001). Partial dependence plots generated by BRT show the marginal effect of each predictor
285 on the response variable while accounting for the average effects of the other variables used
286 in the model. Thus, these plots are useful for comparing the relationship and influence of each
287 explanatory variable on the response variable. BRT analyses were conducted using the
288 Bernoullli family of presence/absence ASVs reads, a bag fraction of 0.5, a learning rate of 0.001
289 and a tree complexity of 3. BRT models were evaluated using a 10-fold cross validation
290 procedure (i.e. 90% of data is used for training and 10% for validation). The *dismo* (Hijmans et
291 al. 2020) R package was used to perform BRT analyses.

292

293 **3. Results**

294 3.1 Metabarcoding data

295 30,251,272 reads were obtained by Miseq Illumina sequencing of a total of 610 samples from
296 Catalan and French rivers. After quality filtering steps 25,452,802 reads and 6403 ASVs were
297 obtained, of which 148, 83, 29 and 14 were classified into *Achnanthidium minutissimum*,
298 *Fistulifera saprophila*, *Nitzschia inconspicua* and *N. soratensis*, respectively. After filtering to
299 remove ASVs having < 1000 reads and occurring in < 2 samples with environmental data, the
300 molecular inventory of the four species consisted of 531 samples and a total of 75 ASVs, of
301 which 45, 18, 9 and 3 belonged, respectively, to *Achnanthidium minutissimum*, *Fistulifera*

saprophila, *Nitzschia inconspicua* and *N. soratensis* (Supplementary table 1; Supplementary data). We checked the 75 ASVs using MegaX and with reference to a matrix of available *rbcL* sequences of diatoms. There was no evidence of sequencing artifacts, i.e. no indels, nor stop codons, nor implausible amino-acids such as substitutions that have no parallel in other diatoms or involve changes in the type of amino-acid (polar vs non-polar vs basic vs acidic).

3.2 Geographical distribution of ASVs within the study area

Most of the 20 most abundant ASVs from *A. minutissimum* (all with >10,000 reads in the dataset) were widely distributed in both Catalan and French rivers, as shown in Fig. 1 and Supplementary Fig. 2. However, despite their apparent ability to colonize sites across the whole geographical region surveyed, individual ASVs seemed to show contagious distributions (aka clumped distributions). For example, ASV70 dominated the *A. minutissimum* assemblage in Mediterranean river sites from the south-central and north-east of Catalonia, but was much less important than ASVs 6 and 7 over most parts of France, being an important ASV there only in scattered sites, e.g. in Normandy and along the Loire (Fig. 1); ASV119 was an important component at three sites located in the Pyrenees, but occurred also in the Jura and Alps regions of eastern France (Fig. 1). Some of the 20 most abundant ASVs appeared to be restricted to one or other country. Thus, ASVs 153 and 219 were only detected in Catalan rivers (Fig. 1; Supplementary Fig. 2), and ASV269 only in French rivers (Supplementary Fig. 2). Interestingly, one of the ASVs restricted to Catalonia, ASV219, was closely related – 1 bp difference in the 263 bp alignment– to ASV7, which dominated the *A. minutissimum* complement in the same central area of Catalonia where ASV219 occurred; likewise, ASV153 was generally found alongside or replacing ASV6 in central Catalan sites (Fig 1; Supplementary Fig. 2), the two ASVs again differing by only 1 bp. Out of the 25 less abundant ASVs of *A.*

minutissimum considered here, 11 were common in both countries, 3 ASVs were only recorded in Catalan rivers and 11 were only in French rivers (Supplementary Table 1).

In *Fistulifera* too, some ASVs showed wide patterns of distribution. Eleven out of the total of 18 *F. saprophila* ASVs were detected in both Catalan and French rivers and most of them were recorded at numerous sites across the study area (Supplementary Figs. 3 and 4; Supplementary Table 1), Nevertheless, contagion was obvious. For example, ASVs 16 and 74 dominated in Catalonia, though they also occurred in scattered sites in France, mostly in eastern parts, while ASV43 showed no obvious pattern in Catalonia, but was restricted to the upper regions of the Rhone catchment in France (Supplementary Fig. 3). Three ASVs were recorded only in French rivers and four in Catalan rivers (Supplementary Table 1). ASVs exclusively recorded in French rivers (ASVs 187, 198 and 233) were much more abundant than those only recorded in Catalan rivers (ASVs 643, 823 and 983) and were widely distributed within France (Supplementary Figs. 3 and 4; Supplementary Table 1). For example, while there was a noticeable concentration of ASV233 in the Garonne catchment of SE France, this ASV also occurred in scattered locations in almost all the other major French basins surveyed (Supplementary Fig. 3). One of the ASVs occurring only in Catalonia, ASV643, was restricted to just two sites in the NE, where it formed c. 50 to more than 75% of the *Fistulifera* reads (Supplementary Fig. 4). Conversely, *Fistulifera* ASVs 823 and 983 occurred in 13 and 17 sites, respectively, but were always rare (Supplementary Fig. 4; Supplementary Table 1). *Fistulifera* ASV234 tended to dominate any assemblage where it was present and rarely occurred with any of the other common ASVs (Supplementary Fig. 3).

In the case of *N. inconspicua*, 7 out of the 9 ASVs analysed were detected in both Catalan and French rivers, most of them (ASVs 53, 56, 113, 273, 463) being broadly distributed (Supplementary Fig. 5; Supplementary Table 1). The remaining 2 ASVs were detected in only one country, but were represented there by 10 (ASV572 in France) and 8 sites (ASV615 in Catalonia) respectively, and were not restricted to a single catchment (Supplementary Fig. 5;

Supplementary Table 1); indeed, ASV572 spanned the whole of France, from the extreme north to almost the most southerly site sampled (Supplementary Fig. 5). In France, ASVs 53 and 113 tended not to co-occur; for example, ASV53 picked out the course of the Loire but ASV113 the Garonne. In Catalonia, the pattern seemed to differ, the two cooccurring quite frequently (Supplementary Fig. 5). One of the rarer ASVs, ASV463, though only found at 14 sites (Supplementary Table 1), nevertheless occurred over a wide range, from the southern part of Catalonia to the Jura mountains in eastern France (Supplementary Fig. 5). The 3 ASVs identified as *N. soratensis* had a higher occurrence in French rivers, though all of them were also identified in Catalan rivers (Supplementary Fig. 6; Supplementary Table 1). However, in Catalonia ASV288 was found only at one site in the extreme north. There was no obvious pattern in the geographical distribution of ASV94 and ASV117 in France (Supplementary Fig. 6).

3.3 ASVs identified in our study area in relation to previously sequenced haplotypes

The ASVs detected in our dataset extended the known genetic diversity of each species, although the most abundant ASVs generally found a match among the Sanger sequences already available. For example, of the five most abundant *A. minutissimum* ASVs, only one (ASV70) represented a haplotype not in GenBank or Diat.barcode version 9 (Fig. 2) and this ASV differed by only 1 bp from the most similar haplotypes. Likewise, the three most abundant *F. saprophila* ASVs (ASVs 16, 43 and 48) all matched sequences already in Diat.barcode version 9 (Supplementary Fig. 7). Perhaps the most surprising 'newcomers' were ASV164 (24961 reads) and the clade of ASVs 156, 272 and 956 (together >42000 reads; Fig. 1), all within *A. minutissimum*: none of these seem to have been found before, despite their high abundance and wide distribution.

Among the matches between ASVs and Sanger sequences were several that were not surprising, given the source of the clones previously sequenced. For example, in *N.*

inconspicua, ASVs were found that were 100% similar to Sanger-sequenced clones isolated from southern Catalonia; these were ASVs 615, 113 and 273, which are identical to *N. inconspicua* genotypes G2, G3 and G4 of Rovira et al. (2015; see also Mann et al. 2021 for a four-gene analysis) (Supplementary Fig. 8). In other cases, however, the ASVs extended the range of a haplotype. For example, in the *A. minutissimum* complex, ASV253 was 100% identical to Sanger sequences from eastern N America and Portugal, and ASV77 and ASV256 were each identical to clones isolated from Siberia (Fig. 2). The Sanger sequences themselves provide a further example of a widely distributed haplotype, with 100% *rbcL* identity between clones isolated from Montana and Hawaii (GenBank accessions KJ658384 and KJ658385). Of the *F. saprophila* ASVs, ASV48 was 100% identical to sequences from Luxembourg and South Korea (Supplementary Fig. 7), while *N. inconspicua* ASV463 was identical to a *N. inconspicua* isolated from the tropical Indian Ocean island of Mayotte (Supplementary Fig. 8).

In contrast, some *rbcL* Sanger sequences in the four species were not represented in our HTS dataset. These included a clade of three *N. inconspicua* haplotypes from the islands of La Réunion and Mayotte, c. 1400 km apart in the Indian Ocean (from clones TCC474, 510 and 571; all were at least 10 bp different from any *inconspicua* ASV in our dataset); and two *N. inconspicua* haplotypes isolated from saline habitats in S Catalonia (G5 and G6, from 31–40 PSU and 5 PSU, respectively: Rovira et al. 2015) (Supplementary Fig. 8). In *Fistulifera* a 'tropical clade' of haplotypes from Mayotte and S Japan had no parallel in our dataset, nor *F. alcalina*, recently described from Florida, USA (Supplementary Fig. 7).

3.4 Redundancy analysis

Given the non-uniform distributions of the ASVs of all four species in the study area, we examined their occurrence and abundance in relation to environmental variables. Those selected by forward selection ($p < 0.05$) were altitude, calcium, conductivity, HCO_3^-, pH, PO_4^{3-},

SO_4^{2-}, TOC and water temperature. An RDA model that included these variables explained 15% of the constrained variance, the first two axes accounting respectively for 7.3% and 4.6% (Supplementary Fig 9). 69 MEMs were selected by forward selection and an RDA model that included these selected MEMs explained a total of 39% of the constrained variance, of which the first and second axes accounted for 11.3% and 10% respectively. This indicates an important degree of spatial structuring of the ASVs assemblages.

3.5 Responses to environmental data

3.5.1 *Achnanthidium minutissimum* (ADMI)

Z scores obtained by TITAN analyses performed on ASVs of *Achnanthidium minutissimum* were hierarchically clustered and visualized through a heatmap plot. Three main groups of ASVs ADMI EG1, ADMI EG2 and ADMI EG3 (= *A. minutissimum* Ecological Groupings 1, 2 and 3) could be distinguished on the basis of the magnitude (given by z score) and type (either positive or negative) of their responses (Fig. 3; Supplementary Table 2). ADMI EG1 constituted a group formed by 7 ASVs which, shared a positive response to altitude, calcium, conductivity, NH_4^+, pH, SO_4^{2-} and a negative response to water temperature (Fig. 3; Supplementary Table 2). In contrast to the positive response showed by ADMI EG1, the 7 ASVs that constituted ADMI EG3 group were characterized by an often negative response to altitude, calcium, conductivity, NH_4^+, pH, SO_4^{2-} and the response was especially strong for calcium and conductivity (Fig. 3; Supplementary Table 2). Assemblage changes points to calcium and conductivity differed between positive and negative responders (Supplementary Fig.10). Kruskal–Wallis and post-hoc Dunn's test indicated that the tree groupings were distributed at waters with significantly different levels of calcium, conductivity, pH, NH_4^+ and SO_4^{2-} (Table 3).

BRT analyses indicated that calcium importantly influenced the occurrence of ADMI EG3 and ADMI EG1 since for these groups it was the variable with the highest and second

highest relative importance respectively (Table 2). As in the TITAN analysis, partial dependence plots generated by BRT models indicated a positive relationship of ADMI EG1 with both calcium and conductivity but a negative relationship of ADMI EG3 with both variables (Fig. 4). These plots showed that the response to calcium and conductivity largely increased in ADMI EG1 group from 0 to 120 mg/L and from 0 to 700 µS/cm respectively but decreased in ADMI EG3 group from 35 to 55 mg/L and from 200 to 400 µS/cm respectively (Fig. 4). BRT models explained 47% and 44% of the total deviance and 30% and 27% of cross-validated deviance for ADMI EG1 and ADMI EG3 groups respectively.

In contrast to the ADMI EG1 and ADMI EG3 groups, the ADMI EG2 group, formed by 18 ASVs was characterized by a positive response to altitude and a negative response to NO_3^-, NH_4^+, PO_4^{3-}, TOC and water temperature (Fig. 3). The magnitude of response to altitude and TOC was especially strong in some ASVs (Fig. 3; Supplementary Table 2).

BRT models indicated that altitude, conductivity, TOC and water temperature were the four variables that most influenced the occurrence of ASVs in the ADMI EG2 group (Table 2). Partial dependence plots showed a positive relationship of the grouping with altitude and a negative with TOC and conductivity (Fig. 4). These plots depicted a large increase in the response to altitude from 0 to 200 m and a decrease in the response to TOC from 1.5 to 5 mg/L (Fig. 4). The BRT model based on the ADMI EG2 group explained 47% of deviance and 26% of cross-validated deviance.

3.5.2 *Fistulifera saprophila* (FSAP)

According to the heatmap based on TITAN z scores obtained for ASVs of *F. saprophila*, three ecological groupings were distinguished: FSAP EG1, FSAP EG2 and FSAP EG3 (Fig. 3). FSAP EG1 group was formed by 7 ASVs, most of them showing a positive response to calcium, conductivity, NO_3^-, NH_4^+, pH, SO_4^{2-}, PO_4^{3-} and TOC. Out of these variables, the strongest

452 responses (high z scores) in the group were to conductivity, NH_4^+ and PO_4^{3-} (Fig. 3;
453 Supplementary Table 2)

454 FSAP EG2 was comprised by 4 ASVs and they were characterized by a negative
455 response to altitude, calcium and conductivity and by a positive response to TOC and water
456 temperature (Fig. 3; Supplementary Table 2). In contrast to the responses shown by ASVs from
457 FSAP EG1 and FSAP EG2 groups, the ASVs from FSAP EG3 group were characterized by being
458 the only ASVs of *Fistulifera saprophila* that responded negatively to SO_4^{2-}, PO_4^{3-} and TOC (Fig.
459 3; Supplementary Table 2). With respect to SO_4^{2-} and TOC, assemblage change points differed
460 between positive and negative responders (Supplementary Fig. 11). Kruskal–Wallis and post-
461 hoc Dunn's test indicated that FSAP EG3 ASVs were distributed in river sites with statistically
462 different values of TOC, PO_4^{3-}, NO_3^- and diatom indexes (i.e. IPS and IBD) (Table 3).

463 BRT analyses indicated that altitude and TOC importantly influenced the occurrence of
464 FSAP EG3, since these variables were respectively the first and second variables with the
465 highest relative importance. PO_4^{3-} was the most important variable in FSAP EG1 models but
466 altitude in FSAP EG2 models (Table 2). Partial dependence plots showed that the response of
467 FSAP EG1 and G3 groups to TOC decreased from 1 mg/L to 4-4.5 mg/L TOC. After this gradient,
468 the response of FSAP EG1 to TOC largely increased from 4.5 mg/L to 5 mg/L whereas there was
469 not any response for FSAP EG3 after 4 mg/L TOC (Fig. 4).

470 These plots reflected a positive relationship of FSAP EG1 and FSAP EG2 with SO_4^{2-} and
471 a negative one of FSAP EG3 with SO_4^{2-}. This was observed in the increasing response of both
472 FSAP EG1 and EG2 (though intermittently in the former case) along the gradient between 0 to
473 500 mg/L and in the large decreasing response of FSAP EG3 from 0 to 20 mg/L (Fig. 4). Partial
474 dependence plots also indicated a negative relationship of ASVs from FSAP EG2 with altitude,
475 since the plot depicted a large decrease in the response from 200 to 600m (Fig. 4). In contrast,
476 the response increased from 0 to 400 m for the FSAP EG1 group, while in the case of FSAP EG3,

477 the response increased from 0 to 600 m and partially and gradually decreased from 700 to
478 1030 m (Fig. 4). BRT models explained 53.4%, 40.8%, 41.7% of the deviance for FSAP EG1, FSAP
479 EG2 and FSAP EG3 respectively, and 37.2%, 24.3% and 21.6% of cross-validated deviance for
480 FSAP EG1, FSAP EG2 and FSAP EG3 respectively.

481

482 3.5.3 *Nitzschia* species

483 Based on TITAN analysisof *Nitzschia inconspicua* (NINC) and *N. soratensis* (NSTS), two
484 ecological groupings were defined (Fig. 5). All the 5 ASVs in the first group corresponded to *N.*
485 *inconspicua* species and they were characterized by a marked positive response to NO_3^-, NH_4^+,
486 SO_4^{2-}, PO_4^{3-} and TOC and by a very strong positive response to conductivity. The second group
487 comprised all three ASVs from *N. soratensis*, which, unlike the *N. inconspicua* ASVs, showed a
488 negative response to calcium, conductivity, pH, SO_4^{2-}, the responses to the first two being
489 especially strong (Fig. 5; Supplementary Table 2). Sum z scores for calcium differed between
490 ASVs from NINC and NSTS (Supplementary Fig. 12). A Kruskal–Wallis test showed that both
491 species were distributed in waters with significant differences levels of calcium, conductivity,
492 NH_4^+, NO_3^-, pH, PO_4^{3-}, SO_4^{2-} and TOC (Table 3).

493 BRT models were performed separately for the group of ASVs from *N. inconspicua* and
494 the group of ASVs from *N. soratensis*. These models highlighted the importance of calcium for
495 explaining the distribution of ASVs from *N. soratensis*, since it was the variable with the highest
496 relative importance in the model (Table 2). In the case of the ASVs from *N. inconspicua*, the
497 two variables with the highest relative importance were conductivity and PO_4^{3-} respectively
498 (Table 2). Partial dependence plots (Fig. 6) indicated that the relationship of calcium with *N.*
499 *inconspicua* was positive but it was negative with *N. soratensis*. The models depicted an
500 increase in the response, though not continuously, from 10 mg/L to 150 mg/L for *N.*
501 *inconspicua* ASVs and a decrease from 50 to 70 mg/L for *N. soratensis* ASVs (Fig. 6). BRT

models explained 52.8% and 48.4% of deviance and 38.3% and 27.9% of cross-validated deviance in *N. inconspicua* and *N. soratensis* ASVs respectively.

3.6. Relationship between phylogeny and geographical–ecological groupings

Phylogenetic trees of the *A. minutissimum* complex showed very little correlation between the phylogeny and the ecological groupings (Fig. 2), although bootstrap for the tree nodes was low. Five out of the seven ASVs that comprised the ecological grouping ADMI EG3 were placed in the major Clade B and all the ASVs from the ADMI EG1 and ADMI EG2 groupings that passed TITAN uncertainty criteria, except for ASV156 and ASV164, were classified into a second major clade (Clade C). More specifically, all the ASVs from subclade d and all the ASVS except ASV 219 from the subclade h of the major clade C, belonged to the same ecological grouping, ADMI EG2. However, some important exceptions showed that preferences are not always clade-specific and must be determined at the ASV level: thus, ASVs 156 and 272 belong to the same clade and differ by just two base-pairs, but belong to different ecological groupings (2 and 3, respectively).

In the case of *F. saprophila* complex, the ASVs from the different ecological groupings were scattered across the phylogenetic tree, without following any clear pattern (supplementary Fig. 7).

4. Discussion

4.1 High diversity within species is captured by a short *rbcL* barcode

RbcL metabarcoding has been successfully applied for studying diatom species diversity (e.g., Rimet et al., 2018, Stoof-Leichsenring et al., 2020) and is especially useful for species that are difficult to identify based on their morphological characteristics, such as those studied here –

A. minutissimum, *F. saprophila*, *Nitzschia inconspicua* and *N. soratensis.* An extra dimension is given by the use of bioinformatics pipelines that generate amplicon sequence variants (ASVs) as opposed to OTUs, since it is possible not only to identify species but also to detect and quantify genetic diversity within them. Despite its short length, the 312-bp *rbcL* barcode we used revealed substantial genetic diversity within the species studied, even when analysis was restricted to the commoner ASVs, with ≥ 1000 reads and occurring in at least 2 samples with environmental data. These comprised 45 ASVs identified as belonging to the *A. minutissimum* complex and 18 of *F. saprophila*. However, it must be underlined that the total numbers of ASVs obtained for these two species were much higher: 148 for *A. minutissimum* and 76 for *F. saprophila* when ASVs having < 1000 reads and occurring in < 2 samples are also considered.

Interpretation of the low abundance ASVs is not straightforward, because both PCR and Illumina sequencing generate errors. Despite the variety of quality and filtering steps implemented in the various commonly used pipelines for HTS data analyses (Bailet et al., 2020), it is impossible to be sure in all cases which ASVs are real though rare genetic variants and which are artefactual. Clearly this can introduce a major bias in biodiversity studies (Turon et al. 2019; Tsuji et al., 2019). A partial solution in the case of *rbcL*, if no matching Sanger sequence is available, is to see whether the same ASV is present in different datasets generated in different Illumina runs. Another is to assess each ASV by reference to the amino-acids encoded: changes that are unlikely, based on amino-acid substitution matrices (e.g. BLOSUM-62: Styczynski et al. 2008) can be tentatively discarded as artifactual. In this study, the most common sequence of *A. minutissimum* that must be artifactual is ASV2237, the 72nd most abundant sequence assigned to the species and represented by 114 reads; this contains a stop codon and so cannot be functional. However, the least abundant *A. minutissimum* ASV analysed (ASV6401), represented by just one read in the whole dataset, had an amino-acid sequence identical to that of 8 of the 10 most abundant ASVs and cannot be discounted as an error. These results illustrate, therefore, the importance of assessing the validity of sequences

552 even after denoising; it is dangerous to rely only on the abundances, since moderately
553 abundant sequences may nevertheless be artefacts. Conversely, rare sequences or even
554 singletons (i.e. sequences detected with only 1 read) are not necessarily artefacts but can be
555 reliable, as noted in other studies (e.g. Alberdi et al. 2017).

556 The reliability of DNA metabarcoding studies also depends on successful taxonomic
557 assignation of the sequences generated and for this it is important to choose an appropriate
558 confidence threshold. This issue has already been addressed in some studies (Rivera et al.
559 2020; Zizka et al., 2020) and in particular, for the short region of 312pb of the *rbcL* marker,
560 non-strict confidence thresholds have been demonstrated for benthic diatom biomonitoring
561 purposes (Rivera et al., 2020). We chose to set a similarity threshold of 50% (the default in
562 DADA2) in order to catch the maximum number of ASVs assigned to the studied species
563 because there is a risk of losing important ecological information when real ASVs are discarded
564 from a dataset, as has been shown in the taxonomy-free approach developed by Tapolczai et
565 al. (2021). In our dataset, although some of the ASVs' taxonomic assignations had low
566 bootstrap support (i.e. the percentage of times that the sequence was classified into the same
567 taxonomy was low), phylogenetic analyses that included curated reference sequences
568 indicated that all the abundant ASVs used in this study were properly classified into the
569 relevant species complex. Our results indicate that it is advisable to use a non-strict similarity
570 threshold to capture high diversity, provided that other analyses can guarantee the reliability
571 of the taxonomic assignation.

572

573 4.2. Wide geographical distributions of ASVs suggest dispersal is not a major constraint

574 The spatial structuring of ASVs suggested by MEMs analyses is congruent with the fact that
575 different ASVs have different geographical distributions, which ultimately could imply dispersal
576 constraints or different environmental preferences, or both. Although individual ASVs tended

577 to be abundant only in particular regions, in most cases the most abundant ASVs were
578 nevertheless found across more or less the whole region surveyed: only a few abundant ASVs
579 were restricted to one or other of France and Catalonia. Furthermore, in several cases the
580 ASVs matched Sanger-sequenced clones isolated from locations far from the study area, even
581 on different continents. It seems therefore that the ASVs of the species studied here are
582 dispersed quite effectively. Hence, when a ASV of the four species is *not* found in the France–
583 Catalonia dataset, there is a *prima facie* case that the appropriate environmental conditions do
584 not occur there, or at least, not in rivers. Examples are the Indian Ocean clade of *N.*
585 *inconspicua* (TCC clones 474, 510 and 571) and the tropical clade of *Fistulifera.* The species
586 considered here could therefore be argued to conform to the ubiquitous dispersal hypothesis
587 (e.g. Finlay, 2002), like some previous examples that have been sampled extensively, including
588 *Sellaphora capitata* (Evans and Mann 2009) and *S. bisexualis* (Mann et al., 2009), in which
589 identical or extremely similar haplotypes enjoy very wide ranges, despite evidence from
590 microsatellite data (in *S. capitata*) of genetic differentiation between populations separated by
591 only some 10s of km (Vanormelingen et al., 2015). In *N. palea* too, particular haplotypes have
592 extremely wide distributions (Trobajo et al., 2010), even though overall there is evidence of a
593 positive relationship between genetic and geographical distances (Rimet et al. 2014),
594 suggesting that dispersal is not fully effective in preventing genetic divergence.

595 Extra factors that need to be taken into account in interpreting the spatial structuring
596 observed in some ASVs are i) spatial structuring of key environmental variables and ii) the
597 possibility that important variables were not measured. Spatial structuring of the environment
598 was particularly obvious in the case of calcium, conductivity and sulphates, whose levels were
599 generally higher in Catalan rivers than French ones (Table 1). This could partly explain why
600 ASVs characterized by a strong positive response to calcium and conductivity often
601 predominated in Catalan rivers or were restricted there (e.g. ASV153; ASV219; ASVs from *N.*
602 *inconspicua*), whereas ASVs that showed a strong negative response were often better

603 represented in France (e.g. ASV269; ASVs from *N. soratensis*). Unmeasured environmental
604 parameters - such as substrate composition, dissolved oxygen, turbidity, water flow, channel
605 width or metals concentration – may also be influential (cf. Castro et al., 2019; Dalu et al.,
606 2017; Keck et al., 2018a) accounting for the low amount of variance explained by the RDA
607 model built from environmental data.

608 Overall, our results support the idea that individuals can disperse over long distances
609 while stochastic events of colonization and extinction possibly combined with fine scale
610 environmental variation are likely to generate local patchiness, outlining the importance of
611 considering spatial scale when studying diatom biogeographical patterns (Keck et al. 2018a).

612

613 4.3 Ecological preferences differ among ASVs in *A. minutissimum* and *F. saprophila*.

614 Our findings evidence the existence of different ecological preferences among different
615 populations and lineages of both *A. minutissimum* and *F. saprophila*, and importantly, that
616 these preferences are correlated with variations in the short *rbcL* barcode. Clearly, base
617 substitutions in *rbcL* within species (most of which do not in fact affect the amino-acid
618 composition and structure of RuBisCO) are unrelated to the causes of ecotypic differentiation
619 in the four diatom species studied; they are instead useful markers that can be used in
620 metabarcoding datasets to explore the existence and distributions of ecotypes.

621 In both species we found that two of the ecological groupings of ASVs were clearly
622 separated by their opposite responses to calcium and conductivity, while in the case of *F.*
623 *saprophila* a third ecological grouping (FSAP EG3) showed a preference for waters with low
624 organic pollution. It might be argued that the type of response shown by this grouping
625 corresponds better, within the genus *Fistulifera*, to *F. pelliculosa*, since this species is
626 considered to occur from oligo to mesotrophic habitats (Lange-Bertalot et al., 2017). The
627 morphology of FSAP EG3 cells is of course unknown. However, the two ASVs from this

628 grouping (i.e. ASV234 and ASV655) have probably been reliably assigned to *F. saprophila* since
629 phylogenetic analyses positioned these ASVs (which are not close relatives of each other)
630 within clades defined by curated reference sequences of *F. saprophila* (Supplementary Fig. 7).
631 We therefore treat the EG3 ASVs as belonging to *F. saprophila*. However, their ecological
632 preferences contrast starkly with the ecology often assumed for the species. Thus, Lange-
633 Bertalot et al. (2017) wrote that *F. saprophila* exhibits "large populations in heavily degraded,
634 highly eutrophic habitats with strong organic pollution up to polysaprobic conditions … It is …
635 one of the most pollution-tolerant diatoms." A similar assessment was made by Gevrey et al.
636 (2004) and the IPS sensitivity value assigned by OMNIDIA (v5.5; Lecointe et al., 1993) is low
637 (IPSS=2). On the other hand, Lange-Bertalot et al. also noted that *F. saprophila* "can also be
638 found in moderately polluted water although in smaller numbers" and Zgrundo et al. (2013)
639 commented that the species is "a widely distributed taxon with broad ecological tolerances".
640 Our data suggest that, if there is a 'broad tolerance', it may be because the species comprises
641 variants with contrasting requirements and tolerances, not because all *F. saprophila* can grow
642 across a wide range of water types. There are implications for biomonitoring, since the same
643 indicator values cannot be assigned to all the genetic varieties and metabarcoding assessments
644 should take this into account. The well-known tolerance of *F. saprophila* to a wide salinity
645 range, eutrophic conditions, and heavily degraded and organically polluted waters (Zgrundo et
646 al., 2013, Lange-Bertalot et al., 2017, Pniewski et al., 2010) must surely reflect the preferences
647 of the EG1 and EG2 groupings, not the EG3 ASVs. Moreover, the contrasting responses of the
648 EG1 and EG2 *rbcL* ASVs to conductivity suggest that the wide range of salinities recorded for
649 the species (e.g. Zgrundo et al., 2013) is also somewhat misleading, primarily reflecting
650 genotypic diversity rather than phenotypic plasticity.

651 In a lesser way, deviation from the 'expected' ecology was also observed in *A.*
652 *minutissimum*. Whereas one grouping of ASVs (ADMI EG3) was particularly restricted to low
653 nutrient concentrations (i.e. PO_4^{3-}, SO_4^{2-}, NH_4^+ and NO_3^-), as might be expected from the

654 characterization of *A. minutissimum* as an indicator of nutrient-poor, good quality waters (e.g.
655 Potapova and Charles, 2007, especially Appendix A), the other two groupings of ASVs tolerated
656 higher nutrient levels and would explain extension of the species complex into more nutrient-
657 rich waters, creating the impression of a broad ecological tolerance – hence the
658 characterization by Lange-Bertalot et al. (2017) "ecological amplitude apparently very wide"
659 (see also Potapova and Hamilton, 2007; Snoeijs and Balashova, 1998; Round, 2004). The idea
660 that *A. minutissimum* is a heterogeneous collection of lineages with different ecological
661 preferences is not new. For example, Potapova and Hamilton (2007) were able to distinguish
662 morphotypes within *A. minutissimum* and to associate them to some extent with different
663 preferences for conductivity, pH and nutrients. However, the morphological differences
664 between these variants (and between some of those documented by Pinseel et al., 2017), are
665 very subtle and distinguishing them in LM-based assessments is arguably impractical. The
666 metabarcoding approach not only aids identification but also allows vastly greater sampling of
667 *A. minutissimum* across natural communities.

668 Thus, our results for *A. minutissimum* and *F. saprophila* tell the same story, that while
669 overall the two species (i.e. all ASVs assigned to each of *A. minutissimum* and *F. saprophila*
670 taken together) have a very broad ecological tolerance, individual genetic variants (ASVs) do
671 not, and the perceived ecological preferences – and indicator value – of the species will differ
672 according to the types and relative abundances of the different ASVs present.

673

674 4.4 Ecological groupings of ASVs do not correspond well to phylogenetic groupings

675 The preferences we obtained for the ASVs are based on correlations between their relative
676 abundances in different samples and the environmental characteristics at the sites where the
677 samples were obtained, exactly as has been done previously with microscopical cell counts to
678 determine the preferences of morphologically defined species. These correlations likely reflect

679 adaptations of the ASVs to different ecological conditions and ASVs that are closely related
680 phylogenetically might be expected to share similar adaptations and belong to the same
681 ecological grouping (Keck et al., 2016a, b, 2018b). Overall, we did not find very strong evidence
682 of a correlation between ecological and phylogenetic group, though there were some trend
683 that could be observed in some cases. For instance, in *A. minutissimum*, the more distantly
684 related ASVs generally belonged to the groupings that differed most (i.e. ADMI EG1 and ADMI
685 EG3). And in *Fistulifera*, ASV74, which tolerated a high conductivity level (c. 9.000 µS/cm), was
686 closely related to a sequence (HQ337547) from clone CCMP543, isolated from a brackish pond
687 (in Massachusetts USA; this clone is often kept in fully marine medium), and clone TCC809,
688 isolated from the River Arão estuary in Portugal (Rimet et al., 2019). However, the *F.*
689 *saprophila* ASV recorded in the highest conductivity site in our dataset (c. 13.000 µS/cm) was
690 ASV445, which is not closely related to ASV74 and belongs to a clade whose other members
691 were recorded from freshwaters.

692

693 4.5. *Nitzschia inconspicua* and *N. soratensis* differ in their ecology but ASVs in each species
694 showed very similar preferences

695 Phylogenetic analyses show that *Nitzschia inconspicua* and *N. soratensis* are not close relatives
696 (Mann et al. 2021) but in the light microscope they are barely separable (Trobajo et al. 2013).
697 However, the value of differentiating between them in ecological and biomonitoring studies
698 has already been shown (Trobajo et al. 2013 and Kelly et al. 2015) and is further confirmed
699 here. Calcium and conductivity were the environmental parameters that most influenced the
700 occurrence of these species according to our data and the preference of *N. soratensis* for low
701 calcium and conductivity (see also Kelly et al. 2015) might explain why this species was
702 widespread in French rivers but scarcely detected in the Catalan ones.

In relation to ecological preferences, we found no differentiation between the ASVs in *N. inconspicua* or *N. soratensis*, in contrast to *A. minutissimum* and *F. saprophila*. For *inconspicua* this was surprising because Rovira et al. (2015) showed that this 'species' is paraphyletic and comprises several very distantly related lineages. Furthermore, their experimental work showed different salinity responses among *inconspicua* genotypes (Rovira et al., 2015). However, the absence of the 'Indian Ocean' haplotypes from French and Catalan rivers (section 3.3) may suggest ecological differentiation from the European ASVs and hence that the structure of the *N. inconspicua* complex is not unlike that in *A. minutissimum* and *F. saprophila*, containing populations adapted to different ecological conditions. This can only be studied using molecular markers via a metabarcoding approach.

Conclusions

Our results show how intraspecific and cryptic diversity can be assessed and understood through the application of DNA metabarcoding, leading to improvements in the knowledge of dispersion patterns, phylogeny and ecological preferences of species and infraspecific variants (see also De Luca et al., 2021; Rivera et al., 2018; Wattier et al., 2020; Zizka et al., 2020). This approach is particularly appropriate for species or species complexes that are difficult to distinguish on the basis of morphological characteristics and whose preferences are therefore still not well-defined. There are many further examples in diatoms that would benefit greatly from this approach, such as the *Cocconeis placentula* complex (Lange-Bertalot et al., 2017) and *Planothidium* species (Jahn et al., 2017).

In relation to the questions we posed for this study, it is clear that genetic variants within *Achnanthidium minutissimum* and *Fistulifera saprophila* are not distributed evenly across the study area and it seems that this is at least partly due to differences in their ecological preferences. Our data indicate that the broad ecological tolerances and wide

distributions claimed for some diatom species may well be the result of a continuum of overlapping preferences among individual genetic variants, which can only be discriminated using molecular markers. Importantly, however, there was little or no agreement between ecological and phylogenetic groupings in *A. minutissimum* and *F. saprophila*, which shows that, at least here, it is necessary to work at the lowest "taxonomic" level possible – ASVs – because it cannot be assumed that clades of species and infraspecific variants share the same ecological preferences and distributions.

Acknowledgements

We are very grateful to the Catalan Water Agency (ACA) for managing and organizing the river survey and the following consultancies for taking DNA samples for us: Sorelló, Estudis del Medi Aquàtic; CERM,Centre d'Estudis dels Rius Mediterranis -Universitat de Vic; GESNA Estudis Ambientals; and Hidrologia i Qualitat de l'Aigua. We also thank the OFB (Office Français de la Biodiversité), the French Water Agencies and the DREAL (Direction Régionale de l'Environnement, de l'Aménagement et du Logement) who made possible the study in France; and two anonymous reviewers for very constructive comments on the manuscript.

The authors also acknowledge support from the CERCA Programme/ Generalitat de Catalunya. J. Pérez-Burillo acknowledges IRTA and 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 was also facilitated by COST Action DNAqua-Net (CA15219), supported by the COST (European Cooperation in Science and Technology) program.

751

752 References

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

755 Alberdi, A., Aizpurua, O., Gilbert, M., & Bohmann, K. (2017). Scrutinizing key steps for reliable
756 metabarcoding of environmental samples. *Methods in Ecology and Evolution*, 9(1),
757 134–147. https://doi.org/10.1111/2041-210X.12849

758 Armbrust, E.V. (2009). The life of diatoms in the world's oceans. *Nature* 459, 185–192.
759 https://doi.org/10.1038/nature08057.

760 Bailet, B., Apothéloz-Perret-Gentil, L., Baričević, A., Chonova, T., Franc, A., Frigerio, J.-M., Kelly,
761 M., Mora, D., Pfannkuchen, M., Proft, S., Ramon, M., Vasselon, V., Zimmermann, J. &
762 Kahlert, M. (2020). Diatom DNA metabarcoding for ecological assessment: Comparison
763 among bioinformatics pipelines used in six European countries reveals the need for
764 standardization. *Science of the Total Environment*, 745, 140948.
765 https://doi.org/10.1016/j.scitotenv.2020.140948.

766 Baker, M. E., & King, R. S. (2010). A new method for detecting and interpreting biodiversity and
767 ecological community thresholds. *Methods in Ecology and Evolution*, 1(1), 25–37.
768 https://doi.org/10.1111/j.2041-210X.2009.00007.x

769 Baker, M. E., King, R. S., & Kahle., D. (2019). TITAN2: Threshold Indicator Taxa Analysis. R
770 package, version 2.4. https://CRAN.R-project.org/package=TITAN2

771 Blanchet, F. G., Legendre, P., & Borcard, D. (2008). Forward selection of explanatory variables.
772 *Ecology*, 89(9), 2623 –2632. https://doi.org/10.1890/07-0986.1

773 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P.
774 (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nature*
775 *Methods*, 13, 581–583. https://doi.org/10.1038/nmeth.3869

776 Castro, E., Siqueira, T., Melo, A. S., Bini, L. M., Landeiro, V. L., & Schneck, F. (2019).
777 Compositional uniqueness of diatoms and insects in subtropical streams is weakly
778 correlated with riffle position and environmental uniqueness. *Hydrobiologia,* 842, 219–
779 232. https://doi.org/10.1007/s10750-019-04037-8

780 CEN (2014). CEN_EN 13946: Water Quality - Guidance for the Routine Sampling and
781 Preparation of Benthic Diatoms From Rivers and Lakes. pp. 1–22.

782 CEN (2018). CEN/TR 17245: Water Quality –Technical Report for the Routine Sampling of
783 Benthic Diatoms From Rivers and Lakes Adapted for Metabarcoding Analysis. CEN/ TC
784 230/WG23 – Aquatic Macrophyte and Algae. pp. 1–8.

785 Chonova, T., Keck, F., Bouchez, A., & Rimet, F. (2020). "A ready-to-use database for DADA2:
786 Diat.barcode_rbcL_263bp_DADA2 based on Diat.barcode v9". Portail Data INRAE, V2.
787 https://doi.org/10.15454/QBLSXP

788 Dalu, T., Wasserman, R. J., Magoro, M. L., Mwedzi, T., Froneman, P. W., & Weyl, O. L. F. (2017).
789 Variation partitioning of benthic diatom community matrices: effects of multiple
790 variables on benthic diatom communities in an Austral temperate river system. *Science*
791 *of the Total Environment*, 601, 73–82. https://doi.org/10.1016/j.scitotenv.2017.05.162

792 Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer,
793 S., Bista, I., Lodge, D. M., Vere, N., Pfrender, M. E., & Bernatchez, L. (2017).
794 Environmental DNA metabarcoding: Transforming how we survey animal and plant

communities. *Molecular Ecology,* 26(21), 5872–5895. https://doi.org/10.1111/mec.14350

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 journal*. https://doi.org/10.1038/s41396-021-00895-0

Dinno, A. (2017). Dunn's Test of Multiple Comparisons Using Rank Sums. R package, version 1.3.5. https://cran.r-project.org/web/packages/dunn.test/dunn.test.pdf

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N. & Wagner, H.H. (2020). Adespatial: Multivariate Multiscale Spatial Analysis. R package, version 0.3-8. https://CRAN.R-project.org/package=adespatial

Dray, S., Legendre, P., & Peres-Neto, P.R. (2006). Spatial modelling: a comprehensive frame-work for principal coordinate analysis of neighbours matrices (PCNM). *Ecological Modelling*, 196(3–4), 483-493. https://doi.org/10.1016/j.ecolmodel.2006.02.015

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3), 241–252.

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

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

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

Evans, K.M. & Mann, D.G. (2009). A proposed protocol for nomenclaturally effective DNA barcoding of microalgae. *Phycologia* 48, 70–74. https://doi.org/10.2216/08-70.1

Finlay, B. J., Monaghan, E. B., & Maberly, S. C. (2002). Hypothesis: The Rate and Scale of Dispersal of Freshwater Diatom Species is a Function of their Global Abundance. *Protist*, 153(3), 261–273. https://doi.org/10.1078/1434-4610-00103

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gevrey, M., Rimet, F., Park, Y. S., Giraudel, J.-L., Ector, L., & Lek, S. (2004). Water quality assessment using diatom assemblages and advanced modelling techniques. *Freshwater Biology*, 49(2), 208–220. https://doi.org/10.1046/j.1365-2426.2003.01174.x

Hijmans, R.J., Phillips, S., Leathwick, J., & Elith, J. (2020). dismo: Species Distribution Modeling. R package version 1.3-3. https://CRAN.R-project.org/package=dismo

Hollander, M., & Wolfe, D.A. (1973). *Nonparametric statistical methods*, 2nd ed. New York, NY, USA, Wiley.

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 Research*, 32(1), 75–107. https://doi.org/10.1080/0269249X.2017.1312548.

839 Keck, F., Bouchez, A., Franc, A., & Rimet, F. (2016a). Linking phylogenetic similarity and
840 pollution sensitivity to develop ecological assessment methods: a test with river
841 diatoms. Journal of Applied Ecology, 53(3), 856–864. https://doi.org/10.1111/1365-
842 2664.12624

843 Keck, F., Rimet, F., Franc, A., & Bouchez, A. (2016b). Phylogenetic signal in diatom ecology:
844 perspectives for aquatic ecosystems biomonitoring. *Ecological Applications*, 26(3),
845 861–872. https://doi.org/10.1890/14-1966

846 Keck, F., Franc, A., & Kahlert, M. (2018a). Disentangling the processes driving the
847 biogeographyof freshwater diatoms: a multiscale approach. *Journal of Biogeography*,
848 45, 1582–1592. https://doi.org/10.1111/jbi.13239.

849 Keck, F., Vasselon, V., Rimet, F., Bouchez, A., & Kahlert., M. (2018b). Boosting DNA
850 metabarcoding for biomonitoring with phylogenetic estimation of operational
851 taxonomic units' ecological profiles. Molecular Ecology Resources, 18(6), 1299–1309.
852 https://doi.org/10.1111/1755-0998.12919

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

857 Kelly, M. G., Trobajo, R., Rovira, L., & Mann, D .G. (2015). Characterizing the niches of two very
858 similar *Nitzschia* species and implications for ecological assessment, *Diatom Research*,
859 30(1), 27–33. https://doi.org/10.1080/0269249X.2014.951398

860 Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary
861 Genetics Analysis across computing platforms. *Molecular Biology and Evolution*, 35(6),
862 1547–1549. https://doi.org/10.1093/molbev/msy096

863 Lange-Bertalot, H., Hofmann, G., Werum, M., & Cantonati, M. (2017).*Freshwater benthic*
864 *diatoms of Central Europe: over 800 common species used in ecological assessment.*
865 *English Edition with updated taxonomy and added species.* English Edition with
866 updated taxonomy and added species. Schmitten-Oberreifenberg, Koeltz Botanical
867 Books.

868 Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2020). A microbial *mandala* for
869 environmental monitoring: Predicting multiple impacts on estuarine prokaryote
870 communities of the Bay of Biscay. *Molecular Ecology,* 1– 19.
871 https://doi.org/10.1111/mec.15489

872 Lecointe, C., Coste, M., & Prygiel, J. (1993). OMNIDIA—software for taxonomy, calculation of
873 diatom indexes and inventories management. *Hydrobiologia*, 269, 509–513.
874 https://doi.org/10.1007/BF00028048

875 Legendre, P., & Legendre, L (2012). Chapter thirteen – Spatial Analysis. *Numerical ecology*, 3rd
876 ed. (pp. 785–858). Amsterdam, Elsevier Science BV.

877 Letunic I., & Bork P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new
878 developments. *Nucleic Acids Research*, 47(W1), W256–W259.
879 https://doi.org/10.1093/nar/gkz239.

880 Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone,
881 D., de Vargas, C., Bittner, L., Zingone, A., & Bowler, C. (2016). Insights into global
882 diatom distribution and diversity in the world's ocean. *Proceedings of the National*

Academy of Sciences of the USA, 113(111), E1516–E1525. https://doi.org/10.1073/pnas.1509523113

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

Mann, D. G., Evans, K. M., Chepurnov, V. A., & Nagai, S. (2009). Morphology and formal description of *Sellaphora bisexualis*, sp. nov. (Bacillariophyta). *Fottea*, 9(2), 199–209. https://doi.org/10.5507/fot.2009.021

Mann, D. G., Trobajo, R., Sato, S., Li, C., Witkowski, A., Rimet, F., Ashworth, M. P., Hollands, R. M., & Theriot, E. C. (2021). Ripe for reassessment: A synthesis of available molecular data for the speciose diatom family Bacillariaceae. *Molecular Phylogenetics and Evolution*, 158, 106985. https://doi.org/10.1016/j.ympev.2020.106985

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17, 10–12.

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. *Ecological Indicators*, 106, 105470. https://doi.org/10.1016/j.ecolind.2019.105470.

Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). Vegan: Community Ecology Package. R package, version 2.5-7. https://CRAN.R-project.org/package=vegan

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

Pinseel, E., Vanormelingen, P., Hamilton, P. B., Vyverman, W., Van de Vijver, B., & Kopalova, K. (2017). Molecular and morphological characterization of the *Achnanthidium minutissimum* complex (Bacillariophyta) in Petuniabukta (Spitsbergen, High Arctic) including the description of *A. digitatum* sp. nov. *European Journal of Phycology*, 52(3), 264–280. https://doi.org/10.1080/09670262.2017.1283540

Piredda, R., Claverie, J.-M., Decelle, J., De Vargas, C., Dunthorn, M., Edvardsen, B., Eikrem, W., Forster, D., Kooistra, W. H. C. F., Logares, R., Massana, R., Montresor, M., Not, F., Ogata, H., Pawlowski, J., Romac, S., Sarno, D., Stoeck, T., & Zingone, A. (2018). Diatom diversity through HTS-metabarcoding in coastal European seas. *Scientific Reports*, 8, 18059. https://doi.org/10.1038/s41598-018-36345-9

Pniewski, F., Friedl, T., & Latała, A. (2010). Identification of diatom isolates from the Gulf of Gdańsk: testing of species identifications using morphology, 18S rDNA sequencing and DNA barcodes of strains from the Culture Collection of Baltic Algae (CCBA). *Oceanological and Hydrobiological Studies*, *39*(3), 3-20. https://doi.org/10.2478/v10009-010-0031-7

Potapova, M., & Charles, D. F. (2007). Diatom metrics for monitoring eutrophication in rivers of the United States. *Ecological Indicators*, 7(1), 48–70. https://doi.org/10.1016/j.ecolind.2005.10.001

Potapova, M., & Hamilton, P. B. (2007). Morphological and ecological variation within the *Achnanthidium minutissimum* (Bacillariophyceae) species complex. *Journal of Phycology*, 43(3), 561–575. https://doi.org/10.1111/j.1529-8817.2007.00332.x

Poulíčková, A., Letákova, M., Hašler, P., Cox, E., & Duchoslav, M. (2017). Species complexes within epiphytic diatoms and their relevance for the bioindication of trophic status. *Science of the Total Environment*, 599-600, 820–833. https://doi.org/10.1016/j.scitotenv.2017.05.034

Poulíčková, A., Špačková, J., Kelly, M. G., Duchoslav, M., & Mann, D. G. (2008). Ecological variation within *Sellaphora* species complexes (Bacillaryophyceae): specialists or generalists? *Hydrobiologia*, 614, 373–386. https://doi.org/10.1007/s10750-008-9521-y

R Core Team, R. (2020). *A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*. Retrieved from https://www.R-project.org/

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., Mann, D. G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., & Bouchez, A. (2019). Diat.barcode, an open-access curated barcode library for diatoms. *Scientific Reports*, 9, 15116, 1–12. https://doi.org/10.1038/s41598-019-51500-6.

Rimet, F., Trobajo, R., Mann, D. G., Kermarrec, L., Franc, A., Domaizon, I., & Bouchez, A. (2014). When is sampling complete? The effects of geographical range and marker choice on perceived diversity in *Nitzschia palea* (Bacillariophyta). *Protist*, 165(3), 245–259. https://doi.org/10.1016/j.protis.2014.03.005

Rimet, F., Vasselon, V., A.-Keszte, B., & Bouchez, A. (2018). Do we similarly assess diversity with microscopy and high-throughput sequencing? Case of microalgae in lakes. *Organism Diversity & Evolution*, 18, 51–62. https://doi.org/10.1007/s13127-018-0359-5.

Rivera, S. F., Vasselon, V., Ballorain, K., Carpentier, A., Wetzel, C. E., Ector, L., Bouchez, A., & Rimet, F. (2018). DNA metabarcoding and microscopic analyses of sea turtles biofilms: Complementary to understand turtle behavior. *PLoS ONE*, 13(4): e0195770. https://doi.org/10.1371/journal.pone.0195770

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. *Ecological Indicators*, 109, 105775. https://doi.org/10.1016/j.ecolind.2019.105775

Round, F. E. (2004). pH scaling and diatom distribution. *Diatom*, 20, 9–12. https://doi.org/10.11464/diatom1985.20.0_9

Rovira, L., Trobajo, R., Sato, S., Ibáñez, C., & Mann, D. G. (2015). Genetic and physiological diversity in the diatom *Nitzschia inconspicua*. *Journal of Eukaryotic Microbiology*, 62(6), 815–832. https://doi.org/10.1111/jeu.12240

Rynearson, T. A., Newton, J. A., & Armbrust, E. V. (2006). Spring bloom development, genetic variation, and population succession in the planktonic diatom *Ditylum brightwelli*. *Limnology and Oceanography*. 51(3), 1249–1261. https://doi.org/10.4319/lo.2006.51.3.1249

Snoeijs, P., & Balashova, N. (1998). *Intercalibration and distribution of diatom species in the Baltic Sea*. Opulus Press, Uppsala.

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. *Ecological Applications*, 30(8), e02205. https://doi.org/10.1002/eap.2205

971 Smetacek, V. (1999). Diatoms and the ocean carbon cycle. *Protist*, 150(1), 25–32.
972 https://doi.org/10.1016/S1434-4610(99)70006-4

973 Soininen, J., Jamoneau, A., Rosebery, J., Leboucher, T., Wang, J., Kokociński, M., & Passy, S. I.
974 (2018). Stream diatoms exhibit weak niche conservation along global environmental
975 and climatic gradients. *Ecography*, 42(2), 346–353.
976 https://doi.org/10.1111/ecog.03828

977 Souffreau, C., Vanormelingen, P., Van de Vijver, B., Isheva, T., Verleyen, E., Sabbe, K. &
978 Vyverman, W. (2013). Molecular evidence for distinct antarctic lineages in the
979 cosmopolitan terrestrial diatoms *Pinnularia borealis* and *Hantzschia amphioxys*.
980 *Protist*, 164(1), 101–115. https://doi.org/10.1016/j.protis.2012.04.001

981 Stoof-Leichsenring, K. R., Pestryakova, L. A., Epp, L. S., Herzschuh, U. (2020). Phylogenetic
982 diversity and environment form assembly rules for Arctic diatom genera—A study on
983 recent and ancient sedimentary DNA. *Journal of Biogeography*. 47(5), 1166–1179.
984 https://doi.org/10.1111/jbi.13786

985 Styczynski, M. P., Jensen, K. L., Rigoutsos, I., & Stephanopoulos, G. (2008). BLOSUM62
986 miscalculations improve search performance. *Nature Biotechnology*, 26(3), 274–275.
987 https://doi.org/10.1038/nbt0308-274

988 Tapolczai, K., Selmeczy, G. G., Szabó, B., B-Béres, V., Keck, F., Bouchez, A., Rimet, F., & Padisák,
989 J. (2021). The potential of exact sequence variants (ESVs) to interpret and assess the
990 impact of agricultural pressure on stream diatom assemblages revealed by DNA
991 metabarcoding. *Ecological Indicators*, 122, 107322.
992 https://doi.org/10.1016/j.ecolind.2020.107322.

993 Trobajo, R., Mann, D. G., Clavero, E., Evans, K. M., Vanormelingen, P., & McGregor, R. C.
994 (2010). The use of partial *cox*1, *rbc*L and LSU rDNA sequences for phylogenetics and
995 species identification within the *Nitzschia palea* complex (Bacillariophyceae). *European*
996 *Journal of Phycology*, 45(4), 413–425. https://doi.org/10.1080/09670262.2010.498586

997 Trobajo, R., Rovira, L., Ector, L., Wetzel, C. E., Kelly, M., & Mann, D. G. (2013). Morphology and
998 identity of some ecologically important small *Nitzschia* species. *Diatom Research*,
999 28(1), 37–59. https://doi.org/10.1080/0269249X.2012.734531

1000 Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2019). Evaluating
1001 intraspecific genetic diversity using environmental DNA and denoising approach: A
1002 case study using tank water. *Environmental DNA*, 2(1), 42–52.
1003 https://doi.org/10.1002/edn3.44

1004 Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2019). From metabarcoding
1005 to metaphylogeography: separating the wheat from the chaff. *Ecological Applications*,
1006 30(2), e02036. https://doi.org/10.1002/eap.2036

1007 Vanormelingen, P., Evans, K. M., Mann, D. G., Lance, S., Debeer, A.-E., D'Hondt, S., Verstraete,
1008 T., De Meester, L., & Vyverman, W. (2015). Genotypic diversity and differentiation
1009 among populations of two benthic freshwater diatoms as revealed by microsatellites.
1010 *Molecular Ecology,* 24(17), 4433–4448. https://doi.org/10.1111/mec.13336

1011 Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., Bouchez, A. (2017a). Application of
1012 highthroughput sequencing (HTS) metabarcoding to diatom biomonitoring: do DNA
1013 extraction methods matter? Freshwater Science, 36, 162–177.
1014 https://doi.org/10.1086/690649

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

Wagenhoff, A., Liess, A., Pastor, A., Clapcott, J. E., Goodwin, E. O., & Young, R. G. (2017). Thresholds in ecosystem structural and functional responses to agricultural stressors can inform limit setting in streams. *Freshwater Science*, 36(1), 178–194. https://doi.org/10.1086/690233

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Associatio*n, 58, 236–244. 10.1080/01621459.1963.10500845.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., & Venables, B. (2020). Gplots: Various R Programming Tools for Plotting Data. R package, version 3.1.1. https://CRAN.R-project.org/package=gplots

Wattier, R., Mamos, T., Copilaş-Ciocianu, D., Jelić, M., Ollivier, A., Chaumot, A., Danger, M., Felten, V., Piscart, C., Žganec, K., Rewicz, T., Wysocka, A., Rigaud, T., & Grabowski, M. (2020). Continental-scale patterns of hyper-cryptic diversity within the freshwater model taxon *Gammarus fossarum* (Crustacea, Amphipoda). *Scientific Reports*, 10**,** 16536. https://doi.org/10.1038/s41598-020-73739-0

Zgrundo, A., Lemke, P., Pniewski, F., Cox, E.J., & Latala, A. (2013). Morphological and molecular phylogenetic studies on *Fistulifera saprophila*. *Diatom Research*, 28(4), 431–443. https://doi.org/10.1080/0269249X.2013.833136

Zizka, V. M. A., Weiss, M., & Leese, F. (2020). Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. *Metabarcoding and Metagenomics*, 4, e51925. https://doi.org/10.3897/mbmg.4.51925

Figures caption

Fig 1. Spatial distribution of the 10 most abundant ASVs from *Achnanthidium minutissimum* in French and Catalan rivers. Segments in each circle represent the proportion of *A. minutissimum* reads recorded in each sample site.

Fig 2. Maximum likelihood phylogenetic tree of *Achnanthidium minutissimum* ASVs obtained in this study and related reference sequences extracted from Diat.barcode v9 and GenBank. The tree was obtained using raxmlGUI and a GRT-Gamma model with 1000 replicates for the bootstrap analyses. The tree was drawn using iTOL. ASVs belonging to the different ecological groupings defined after TITAN analyses are represented: EG1 in red, EG2 in green and EG3 in blue. Black circles represent bootstrap support values of 50-100. Three major clades can be distinguished (A, B and C) and a number of subclades (Ca to Cl).

Fig 3. Heatmap dendrogram based on z score obtained by the different TITAN analyses performed on ASVs from a) *Achnanthidium minutissimum* and b) *Fistulifera saprophila*. Euclidean distance and ward.D functions were used to compute dissimilarity distance and hierarchical clustering respectively on ASVs z scores obtained for the different environmental variables. Only those ASVs with more than 3 responses that fulfilled purity and reliability criteria are represented. Red colour indicates positive responses while blue negative responses. Magnitude of response (z score) are given by the contrast of the colour; dark colours depict strong responses while light colours indicate weak responses. Chart in the upper-left corner indicates the correspondence between colour gradient and z-score.

Fig 4. Partial dependence plots generated by boosted regression trees analyses depicting the response of the ecological groupings of ASVs from *Achnanthidium minutissimum* to altitude (m), Calcium (mg/L), conductivity (µS/cm) and Total Organic Carbon (TOC, mg/L) and ecological groupings of ASVs from *Fistulifera saprophila* to Total Organic Carbon (TOC, mg/L), sulphates (mg/L) and altitude (m). The different groups of ASVs were defined after TITAN analyses. Y axis shows fitted function.

Fig 5. Heatmap based on z score obtained by the different TITAN analyses performed on ASVs from *Nitzschia inconspicua* (NINC) and *N. soratensis* (NSTS). Only those ASVs with more than 3 responses that fulfilled purity and reliability criteria are represented. Red colour indicates positive responses while blue negative responses. Magnitude of response (z score) are given by the contrast of the colour; dark colours depict strong responses while light colours indicate weak responses. Chart in the upper-left corner indicates the correspondence between colour gradient and z-score.

Fig 6. Partial dependence plots generated by boosted regression trees analyses depicting the response of ASVs from *Nitzschia inconspicua* and *Nitzschia soratensis* to Calcium (mg/L). Y axis represents fitted function.

Tables

Table 1. Physicochemical parameters information from the 531 river sites studied.

Variable	Number of sampling sites with available data	Number of sampling sites with available data (Catalan rivers)	Number of sampling sites with available data (French rivers)	Average ± standard deviation of number of records per sampling site within the 90-day period (Catalan rivers)	Average ± standard deviation of number of records per sampling site within the 90-day period (French rivers)	Range (average ± standard deviation) in Catalan rivers	Range (average ± standard deviation) in French rivers
Altitude (m)	531	148	383	NA	NA	3.89 - 1243.97 (303.75 ± 255.29)	0 – 1933 (255.7 ± 314.39)
Ammonium (mg/L)	513	136	377	1 ± 0.08	2.61 ± 2.03	0.1 - 15.33 (0.69 ± 2.21)	0.004 - 1.4 (0.07 ± 0.13)
Bicarbonates (mg/L)	200	35	165	1 ± 0.08	2.09 ± 1.25	25 – 182 (54.18 ± 39.11)	6.4 – 600 (184.37 ± 95.8)
Calcium (mg/L)	335	148	187	1 ± 0.08	2.3 ± 1.81	2.5 - 673.33 (116.05 ± 93.83)	0.7 – 333 (63.66 ± 43.35)
Conductivity (µS/cm)	336	136	200	1 ± 0.08	3.85 ± 3.25	99.5 - 13341.33 (1054.61 ± 1382.76)	25.67 - 2377.67 (341.23 ± 283.14)
Total organic carbon (mg/L)	514	136	378	1 ± 0.08	2.69 ± 2.33	0.5 - 10.65 (3.47 ± 1.85)	0.2 – 15 (2.46 ± 1.67)
Nitrates (mg/L)	502	123	379	1 ± 0.08	2.63 ± 2.12	2.5 - 76.45 (13.94 ± 13.55)	0.48 - 47.27 (7.49 ± 7.22)
Orthophosphates (mg/L)	515	136	379	1 ± 0.08	2.57 ± 1.90	0.1 - 9.73 (0.58 ± 1.09)	0.01 - 2.53 (0.16 ± 0.25)
pH	336	136	200	1 ± 0.08	3.84 ± 3.25	7.65 - 8.8 (8.19 ± 0.23)	6.3 - 8.6 (7.83 ± 0.42)
Sulphates (mg/L)	301	136	165	1 ± 0.08	2.09 ± 1.25	4 – 1500 (178.69 ± 217.83)	1 – 416 (39.92 ± 57.1)
Water temperature (ºC)	330	130	200	1 ± 0.00	4.26 ± 6.44	5 – 26 (12.45 ± 3.17)	7.13 - 24.5 (18.12 ± 3.99)

Table 2. Relative importance (%) of each environmental variable resulting from the boosted regression tree models (with 10-fold cross validation of data) performed for the different groups of ASVs of *Achnanthidium minutissimum* (ADMI), *Fistulifera saprophila* (FSAP), *Nitzschia inconspicua* (NINC) and *Nitzschia soratensis* (NSTS). Groups of ASVs were defined on the basis of TITAN analyses.

Variable	**ADMI G1**	**ADMI G2**	**ADMI G3**	**FSAP G1**	**FSAP G2**	**FSAP G3**	**NINC**	**NSTS**
Orthophosphates	17.90	8.35	9.96	23.44	8.81	2.93	19.69	14.14
Calcium	13.81	5.01	21.75	8.57	5.20	3.89	4.77	20.49
Conductivity	11.89	11.79	7.43	16.28	4.38	3.83	27.49	14.14
Nitrates	8.50	6.41	13.35	8.57	7.98	7.93	13.68	12.35
Altitude	10.09	13.94	9.43	5.30	32.33	29.66	7.00	7.38
pH	8.38	10.03	1.52	2.98	4.66	1.72	2.87	6.85
Water temperature	10.79	22.27	6.88	6.25	11.22	9.14	5.04	3.77
TOC	5.90	12.80	12.12	3.52	7.34	20.79	4.28	8.98
Bicarbonates	7.38	2.26	3.63	6.05	8.05	6.14	1.53	5.24
Ammonium	2.93	4.72	10.89	11.43	7.24	5.89	3.32	3.36
Sulphates	2.45	2.41	3.03	5.46	2.75	8.24	10.28	3.24

Table 3. Range, average and standard deviation environmental parameters analysed in the sites were different defined ecological groupings occurred. [a] and [b] indicate species and ecological groupings with statistically significant differences (Kruskal–Wallis for *Nitzschia inconspicua* and *N. soratensis* and post-hoc Dunn's test for groupings from *Achnanthidium minutissimum* and *Fistulifera saprophila*, $p < 0.05$).

Variable	ADMI G1	ADMI G2	ADMI G3	FSAP G1	FSAP G2	FSAP G3	NINC	NSTS
Orthophosphates	[a]0.01-3.35 (0.23±0.39)	[b]0.01-4.1(0.22±0.40)	[ab]0.01-2(0.13±0.22)	[a]0.01-9.73(0.50±0.95)	[a]0.01-4.3(0.28±0.54)	[a]0.01-4.3(0.22±0.53)	[a]0.01-9.73(0.48±0.24)	[a]0.01-4.10(0.24±0.42)
Calcium	[a]8.25-605 (108.09±76.54)	[a]1.90-477 (86.02±60.58)	[a]1.55-379 (44.12 ± 55.97)	[a]5.11-673.33(108.52±88.99)	[a]3.2-477(81.41±69.77)	4.1-266(85.29±47.23)	[a]8.25-673.33(111.02±87.92)	[a]0.7-333(51.71±52.66)
Conductivity	[a]71.2-9371 (842.55±874.67)	[a]30.67-2885(599.95±479.62)	[a]30-2885 (271.09±335.46)	[ab]83.01-13341.33(1077.06±1452.74)	[a]48-2738(550.23±452.36)	[b]30.67-2377.67(556.84±480.72)	[a]77.5-13341.33(914.78±1254.12)	[a]25.66-2377.67(341.69±346.34)
Nitrates	[a]0.47-61.20 (10.74±11.07)	[b]0.47-76.45 (8.86±9.68)	[ab]0.47-53.50 (6.31±7.67)	[a]0.95-76.45(11.88±11.23)	[a]0.5-61.2(9.36±8.57)	[a]0.5-36.90(6.98±7.21)	[a]0.95-76.45(11.54±10.81)	[a]0.5-53.50(6.94±7.27)
Altitude	[a]0-1476 (323±282.69)	[b]0-1933 (311.16 ±311.80)	[ab]0-1243.97 (193.38 ±220.99)	[a]0-1200 (277.53±228.95)	[a]0-1933(150.48±196.97)	[a]0-1589(409.20±338.60)	0-1042.78(183.22±184.16)	0-1243.97(189.91±235.5)
pH	[a]7.07-8.8 (8.16±0.27)	[a]6.90-8.8 (8.05±0.34)	[a]6.3-8.6 (7.77±0.50)	[ab]7.21-8.8(8.15±0.28)	[a]7-8.8(7.98±0.35)	[b]6.80-8.6(8.03±0.30)	[a]7.33-8.8(8.11±0.28)	[a]6.43-8.6(7.87±0.42)
Water temperature	[a](5-24.02)13.47±3.7	[a]5-24.35(15.03±4.52)	[a]6-23.9 (16.93±4.70)	[a]6-26(13.63±3.72)	[ab]6-24.5(17.29±4.42)	[b]5-23.7(14.18±4.11)	[a]5-26(15.76±4.64)	[a]6-24.35(17.28±4.56)
TOC	0.2-8.5(2.60±1.58)	[a]0.2-15(2.49±1.66)	[a]0.6-13.71 (2.86±1.75)	[a]0.2-15(3.25±2.00)	[b]0.2-15(3.03±1.56)	[ab]0.2-10.65(1.95±1.56)	[a]0.5-10.65(3.39±1.62)	[a]0.6-15(3.00±1.71)
Bicarbonates	[a]25-345(169.65±102.89)	[b]8.67-350(161.64±94.81)	[ab]6.4-600(117.61±127.64)	[a]21.25-384(136.91±102.59)	[b]12.75-350(139.77±93.16)	[ab]13.33-384(199.47±81.13)	25-600(148.66±109.90)	12.2-384(125.93±98.00)
Ammonium	[a]0.01-4.8(0.17 ±0.39)	[a]0.004-12.1(0.15±0.63)	[a]0.05-1.2(0.06 ±0.11)	[a]0.01-15.33(0.55±1.93)	[a]0.01-15.27(0.20±0.98)	[a]0.01-12.10(0.23±1.23)	[a]0.01-15.33(0.48±1.81)	[a]0.01-2.6(0.11±0.26)
Sulphates	[a]3.73-1500 (139.28±197.63)	[a]2.4-970 (96.11±137.11)	[a]1.2-538.5 (36.76±78.65)	[ab]4-1500(154.22±212.57)	[a]3.05-970(103.71±151.58)	[b]2.80-458(79.88±104.84)	[a]4-1500(160.54±209.45)	[a]1-135(32.54±25.85)
IPS	6.19-19.95(13.88±3.53)	6.19-19.95(13.75±3.60)	9.05-19.65(15.23±3.42)	[a]6.19-18.89(12.42±3.16)	[b]6.52-18.57(12±2.99)	[ab]8.01-19.95(14.88±3.57)	6.19-18.71(12.28±3.16)	6.52-18.71(13.01±3.93)
IBD	[ab]10.9-20(17.05±2.47)	[a]5.4-20(15.56±3.33)	[b]8.2-20(15.02±2.82)	[a]5.7-20(14.04+-3.33)	[b]5.4-20(13.62±2.81)	[ab]8.7-20(16.86±3.11)	5.4-19.1(12.78±2.86)	5.4-20(13.33±3.21)