DNA metabarcoding and morphological methods show complementary patterns in the metacommunity organization of lentic epiphytic diatoms

Alejandro Nistal-García[1], Pedro García-García[2], Jorge García-Girón[2], María Borrego-Ramos[3], Saúl Blanco[3], and Eloy Bécares[3]

[1]Universidad de Leon
[2]University of León
[3]Institute of Environment, Natural Resources and Biodiversity

September 17, 2020

Abstract

Diatoms are important organisms in aquatic ecosystems due to their position as primary producers and, therefore, analyzing their communities provides relevant information on ecosystem functioning. Diatoms have been historically identified based on morphological traits, which is time-consuming and require well-trained specialists. Nevertheless, DNA barcoding approach offers an alternative to overcome some limitations of the morphological approach. Unfortunately, however, only a few studies have compared beta diversity patterns for both DNA barcoding and morphological approaches. Here, we derive a new take on this issue and assess the ecological mechanisms underlying spatial variation in epiphytic diatom metacommunities using a comprehensive dataset from 22 Mediterranean ponds at different taxonomic resolutions. Our results suggest a relatively poor correspondence in the compositional variation between morphology–based and molecular–based approaches. We speculate that the incompleteness of the reference database and the bioinformatics processing are the biases most likely related to the molecular approach whereas the limited counting effort and the presence of cryptic species are presumably the major biases related to morphological approach. On the other hand, we found that both approaches were strongly related to the environmental template, suggesting that epiphytic diatom communities were mainly controlled by species sorting at regional extents. Overall, this work suggests that both molecular and morphological approaches provide complementary information on diatom metacommunity organization and emphasizes the importance of DNA barcoding to addressing empirical research questions of community ecology in freshwaters.

Introduction

Diatoms are photoautotrophic unicellular organisms present in almost every freshwater habitat (Zimmermann, Glöckner, Jahn, Enke, & Gemeinholzer, 2015). They are one of the main components in the nitrogen, phosphorus, silicon and carbon biochemical cycles, being responsible of at least the 25% of the global carbon dioxide fixation and the 20% of the global net primary production (Wilhelm et al., 2006; Zimmermann, Jahn, & Gemeinholzer, 2011; Zimmermann et al., 2015). In addition, benthic diatoms are very important ecological indicators as they are highly sensitive to environmental conditions and have relatively short generation times (Vasiljević et al., 2014; Visco, Apothéloz-Perret-Gentil, Cordonier, Esling, Pillet, & Pawlowski, 2015; Round, Crawford, & Mann, 1990). Consequently, analyzing their communities is important to provide an overview of water quality and allow the detection of environmental changes in freshwater assemblages (Kermarrec et al., 2014).

Researches efforts have historically been focused on understanding how ecological communities are assembled in space and time and how they vary at local scale (Bishop, Robertson, van Rensburg, & Parr, 2015; Leibold

et al., 2004). In this vein, the metacommunity concept, first defined by Leibold et al. (2004) as "a set of local communities linked by dispersal of multiple potentially interacting species", offers the possibility to examine the complex ecological mechanisms, since it takes into account the regional-scale processes in addition to local ones (Leibold et al., 2004). From the very beginning (Brown, Sokol, Skelton, & Tornwall, 2016), four not mutually exclusive paradigms -patch dynamics, species sorting, mass effect and neutral perspective- have been associated with metacommunity organization. Each of the four paradigms is characterized by the amount of weight they place on a combination of regional and local processes, disturbance, and the degree to which species are equivalent in their functional biology (Brown et al., 2016). More specifically, in neutral perspective, speciation, extinction, migration and immigration processes participate structuring ecological communities, while in patch dynamics there is a colonization-competition trade off. In species sorting perspective, environmental filtering and biotic interactions filter the occurrence of the species, whereas in mass effect paradigm, high rates of dispersal in addition to environmental factors are taking into account (Heino et al., 2015; Göthe, Angeler, Gottschalk, Löfgren, & Sandin, 2013). Taking in consideration a metacommunity perspective, the β-diversity concept (i.e. variation in community composition and structure among sites in a geographical area) is essential to understand ecosystem functioning, since it provides important information about the patterns of diversity and processes that modify the ecosystems (Bonecker et al., 2013; Florencio, Díaz-Paniagua, Gómez-Rodríguez, & Serrano, 2014). As a consequence, several studies examining diatoms β-diversity at various spatial scales (Green et al, 2004; Smucker & Vis, 2011; Leboucher et al., 2019; Rodríguez-Alcalá et al., 2019) have reported that spatial structure is responsible of a significant proportion of the community variance, which suggest that diatoms lack strict ubiquitous dispersal and thereby exhibit clear biogeographical patterns (Soininen, 2007).

Morphological identification of diatoms species is subjected to a bottleneck since the number of expert taxonomists is decreasing whereas the number of identified taxa is increasing rapidly (Pečnikar and Buzan, 2013). Traditional taxonomic identification relies in morphological traits and light or electron microscopy, so is time-consuming and demands specialized knowledge (Blanco, 2020). In addition, the presence of cryptic species and phenotypic plasticity found in some species may hinder classical species identification (Hadi et al., 2016). In this vein, several studies have recently revealed the presence of hidden diversity on diatoms (Trobajo et al., 2010; Rovira, Trobajo, Sato, Ibáñez, & Mann, 2015), suggesting that its diversity has like been underestimated (Mann & Vanormelingen, 2013). As a consequence, exclusive reliance on morphological traits may lead to ambiguous identification of taxa (Zimmermann et al., 2011; Kowalska, Pniewski, & Latała, 2019). To overcome the biases related with morphological identifications, alternative techniques such as DNA barcoding have been developed in recent years. The term DNA barcoding was first used by Arnot, Roper, & Bayoumi (1993) to refer the possibility of differentiate stocks and lineages of *Plasmodium falciparum* based on targeting tandem repeats of circumsporozoite gene. DNA barcoding relies in a short DNA fragment, which is sequenced and compared with a reference library, allowing the identification of all taxa independently of its life stage (Zimmermann et al., 2015; Rimet, Vasselon, Keszte, & Bouchez, 2018; Hebert, Cywinska, Ball, & deWaard, 2003). In addition, DNA metabarcoding combine DNA barcoding methods with high-throughput techniques (HTS) allowing the sequencing of millions of DNA fragments from many samples simultaneously (Rivera et al., 2017). Metabarcoding has been applied regularly to inventory taxonomic diversity of freshwater diatoms since the pioneering study of Kermarrec et al. (2013). Since then, several studies have highlighted the usefulness of metabarcoding approach in combination with morphological methods for diatom biomonitoring (Mora et al., 2019; Zimmermann et al., 2015). However, the incompleteness of reference database still remains as one of the main biases constraining the accuracy of molecular-based inventories.

Here, we used a comprehensive dataset of 22 ponds located in a Mediterranean landscape to (i) assess if morphological and metabarcoding approaches could be comparable methods in β-diversity studies of benthic diatoms, and (ii) test whether environmental filtering and dispersal limitation, predictably assemble diatom communities. Importantly, we also assessed the influence of taxonomic resolution (genera and species level) on the observed community-level patterns and processes. We focused our study exclusively on incidence values, avoiding the use of abundance values in order to prevent the biases related with the copy number

variation of the marker gene per cell (Vasselon, Domaizon, Rimet, Kahlert, & Bouchez, 2017). Following recent works (Rimet et al., 2018; Medlin, 2018; Kowalska et al., 2019), we hypothesized that morphological methods based on light microscopy and metabarcoding approach provide similar results when it comes to the β-diversity patterns of diatom metacommunities (**Hyphothesis 1**). On the other hand, we expected to find a relationship between environmental template and compositional variation, since freshwater diatoms have usually been found to be strongly related to environmental dissimilarity at regional extents (**Hyphotesis 2** ; Declerk, Coronel, Legendre, & Brendonck, 2011).

Materials and methods

Study area

The present study was performed on 22 permanent and temporary ponds located in a lowland area (900 m. a. s. l.) area (ca. 230 km^2) in the northwest part of Spain **(Figure 1).** The study sites were characterized by a marked variability in environmental conditions, such us morphometry, nutrient content and mineralization **(Table 1).** The predominant land uses in the study area are arable land and pasture. The climate is Mediterranean dry moderate with mean temperatures in summer and winter of 18 °C and 3.2 °C, respectively, and a mean precipitation in summer and winter of 84.5 mm and 173 mm, respectively (1976-2015; data provided by the Spanish Met Agency-AEMET; http://www.aemet.es). The majority of ponds are shallow (0.2-1.5 m), are fed by groundwater and rainfall, and they experience a high reduction in water volume during summer.

Diatom sampling

Samples of benthic biofilms were collected in the shoreline of 22 ponds in June 2018. Epiphytic diatom samples were collected from submerged stems of *Schoenoplectus lacustris* following the methodology proposed by Blanco, Ector, & Bécares, (2004). A scheme of the diatom sampling procedure and the subsequent processes is provided on**Figure 2.** A total of 10-12 stems were cut 10 cm below the water surface. Stems were placed in a plastic bottle with 500 ml of distilled water and shaken for 2 minutes to dislodge attached diatoms following previous studies (Zimba and Hopson, 1997; Riato, Leira, Della Bella, & Oberholster, 2018; Borrego-Ramos, Olenici, & Blanco, 2019). The suspension was split in two samples. One of them was fixed with 4% v/v formaldehyde and used for morphological identification, and the other one fixed with 70% v/v ethanol and used for molecular identification. All samples were transferred to the laboratory in a cool box. Samples for molecular identification were kept at -20 °C until DNA extraction.

Morphological analysis

Diatom frustules were cleaned with 30% v/v hydrogen peroxide and hydrochloric acid following standard methods (European Standard EN 13946 2003). Permanent slides were mounted using Naphrax (refractive index of 1.74). At least 400 frustules were identified and counted in each sample under light microscopy using an Olympus BX60 microscope, according to Krammer & Lange-Bertalot, 1986a; Krammer & Lange-Bertalot, 1986b, Kramer & Lange-Bertalot, 1991a; Kramer & Lange-Bertalot, 1991b and Lange-Bertalot et al., 2017.

Molecular analysis

DNA extraction and PCR amplification

All samples were centrifuged at 11,000 xg for 30 minutes to harvest diatom cells. Then, supernatant was discarded and pellet was resuspended in 200 µL of nuclease-free water. DNA was isolated using the Power Soil DNA Isolation Kit (Mo Bio Laboratories, Carlsbad, California, US) following the manufacturer instructions and its concentration was measured with Nanodrop 1000 (NanoDrop Technologies, Wilmington, Delaware, US). A 312 bp fragment of *rbc* L gene (ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit) was amplified by PCR using an equimolar mix of degenerate primers with overhang adapters for Illumina sequencing **(Table 2)** (Rivera et al., 2017). For all samples, we performed six PCR reactions on 50 µL of reaction mixture containing 10-20 ng/µL of extracted DNA, 2 U of Platinum II Taq Hot-Start DNA Polymerase (Invitrogen, Grand Island, New York, US), 10 µL of 5X Platinum II PCR Buffer, 0.5 µM of

each primer, 5 µL of dNTP mix (2mM each), 10 µL of Platinum GC Enhancer and 9.6 µL of nuclease-free water. PCR conditions included an initial denaturalization step at 94 °C for 4 min followed by 40 cycles of denaturalization at 94 °C for 30 s, annealing at 55 °C for 30 s and extension at 68 °C for 30 s, and a final extension step at 68 °C for 10 min. PCR products were visualized with ultraviolet light in a 1.5% agarose gel stained with ethidium bromide. Bands targeting the *rbc* L barcode gene were excised off from agarose gel and then DNA extracted with Clean-Easy Agarose Purification Kit (Canvax Biotech, Córdoba, Spain) following the manufacturer instructions, except for elution volume, which was 15 µL.

Library preparation and sequencing

rbc L libraries preparation and sequencing were carried out by Sistemas Genómicos S. L. (Paterna, Valencia, Spain). Purified DNA from the six PCR replicates was pooled in one sample and then quantified with Qubit 3.0 (Invitrogen, Life Technologies, Grand Island, New York, US). 5 ng of pooled PCR products from each sample were subjected to indexing PCR on 50 µL of reaction mixture containing KAPA HiFi Hot Start DNA Polymerase (Kapa Biosystems, Wilmington, Delaware, US) and a specific primers combination to enable multiplexing of all PCR products in the same sequencing run. PCR products were purified with Agencourt AMPure XP beads (Beckman-Coulter, Brea, California, US). Then, quality and quantity of purified amplicons was assessed using a 4200 TapeStation (Agilent Technologies, Santa Clara, California, US) with a High Sensitivity D1000 ScreenTape. Finally, *rbc* L libraries were pooled and paired-end sequenced (2x250 bp) on MiSeq 2500 instrument (Illumina, San Diego, California, US) using a Micro MiSeq Reagent Kit v2.

HTS data processing

The sequencing process provided 22 demultiplexed raw fastq files. For each file, the quality of raw data was assessed using FastQC tools (Andrews, 2010). Each set of paired ends reads were merged into a single sequence using the VSEARCH software (Rognes, Flouri, Nichols, Quince, & Mahé, 2016). The minimum length of the overlap region between the reads was at least 36 bp, and the overlap region should not include more than 3 mismatches and 5 ambiguous bases. Next, primers were trimmed off using the trim_oligos.pl software (Sáenz de Miera; not published). Then, poor quality sequencing reads were filtered out according to the following parameters: minimum length = 251 bp, maximum length = 272 bp and Phred quality score = 35 using VSEARCH (Rognes et al., 2016). Next, the identical sequences were dereplicated to obtain Independent Sequence Units (ISUs). Finally, the UCHIME algorithm of VSEARCH (Rognes et al., 2016) was used to remove chimeric sequences.

Phylogenetic analysis

Near-identical sequences were clustered in order to generate Operational Taxonomic Units (OTUs) using the CLUSTER command of VSEARCH (Rognes et al., 2016) with an identity cut-off of 97%. Then, OTUs with only single read (singletons) were removed. Taxonomy was assigned to OTU list using the R-Syst::diatom v7.1 reference database (Rimet et al., 2016) and the BLASTn algorithm (Altschul et al., 1997) with a minimum identity value of 85%. Next, those DNA sequences that did not belong to diatom phylum (Bacillariophyta) were filtered out. All remaining OTU sequences were scanned by EMBOSS Getorf for Open Reading Frames (ORFs) with more than 251 bp (Rice, Longden, & Bleasby, 2000). Then, read values of resulting OTUs were normalized among samples to the most abundant value (82,446 reads). Finally, OTUs with a normalized read value less than 0.005 % were filtered out according with Bokulich et al. (2013).

For the purpose of obtain an overview of the taxonomic assigned OTUs we computed two phylogenetic trees. In order to save computational time, we build a reference phylogenetic tree using the 708 reference sequences that matched with our OTU inventory. In addition, we computed another phylogenetic tree using the same reference sequences and the 3138 taxonomy-assigned OTU sequences. We use MAFFT v7 program (Katoh and Standley, 2013) with the default settings to align all DNA sequences. The Maximum Likelihood (ML) phylogenies were constructed using the tool RAxML (Randomized Accelerated Maximum Likelihood) implemented on the CIPRES Portal (Miller, Pfeiffer, & Schwartz, 2010) using the GTRCATI (Generalized Time Reversible Model + optimization of substitution rates + optimization of site-specific evolutionary rates)

as model of evolution and 1000 replicates for the bootstrap analysis. Phylogenetic trees were visualized with FigTree v1.4.3 (Rambaut, 2016).

Molecular and morphological data comparison

With the aim of visualizing the taxonomic correspondence provided by morphological and molecular data at genus and species levels, we built Venn diagrams using the interactive tool Venny v.2.1 (Oliveros, 2015).

Statistical analyses

Non–metric multidimensional scaling (nMDS) using the Jaccard coefficient was conducted on both morphological and molecular diatom data sets (here, we ran separate analyses based on species– and genus–level resolution data). We used the R package vegan (Oksanen et al., 2019) to compute the Jaccard dissimilarity matrix and run the nMDS routines. Since the Jaccard index is sensitive to sample size, and since metabarcoding data often produce samples of heterogeneous sizes (Ohlmann et al., 2018), we partitioned the Jaccard dissimilarity matrix into its true turnover component (Baselga, 2010) with the package betapart (Baselga et al., 2018). For the sake of comparison, we also ran these and the following analyses using the spatial turnover component of β-diversity.

In order to explore the overall degree of correspondence between the ordinal results obtained for morphological and molecular data sets, Procrustean rotation analysis and the subsequent PROcrustean randomization TEST (PROTEST) were applied (Gower, 1971, Jackson, 1995). PROTEST generates from a permutation–based procedure a correlation–like statistic referred to as "correlation in a symmetric Procrustean rotation" or m_{1-2} and a p value that measures the significance of the concordance established by the Procrustean superimposition approach (Jackson, 1995). In addition, we related the direction of the movement between the base and the end of the procrustean arrow and the length of these arrows to the distribution of the morphology– and molecular–based diatom assemblages to identify the taxa that contributed most to the dissimilarity between ordinations (García-Girón, Fernández-Aláez, Fernández-Aláez, & Luis, 2018a; García-Girón, Fernández-Aláez, & Fernández-Aláez, 2018b).

Each of the resulting dissimilarity matrices (i.e. based on species– and genus–level resolution data separately, and considering the Jaccard coefficient and its true turnover component) was used in a number of statistical tests. First, we used the BIOENV analysis (Clarke and Ainsworth, 1993) to produce "the best" environmental distance matrix and identify associations between environmental features and community composition. In brief, this method is based on standardized environmental variables and tests all combinations of environmental features, providing information as to which combination shows the strongest correlation between compositional variation and environmental distance matrix. Next, we ran a number of Mantel tests and partial Mantel tests (Mantel, 1967; Legendre and Legendre, 2012) to examine the relationships between community dissimilarities and environmental or spatial distances, i.e. we tested for distance decay in community similarity along environmental or spatial gradients. The matrices of spatial distances contained pairwise geographical distances calculated from latitude and longitude, whereas "the best" subset of standardized environmental variables from the BIOENV were used to compute the environmental matrices based on the Euclidean distances between study sites. Mantel tests and BIOENV analysis were based on Spearman's rank correlation coefficient and 999 permutations for obtaining values.

We used Mantel correlograms (Oden and Sokal, 1986) to test if pairwise dissimilarities were spatially autocorrelated within each distance class. The distance classes were determined by Sturge's rule (Legendre and Legendre, 2012) and values were based on 199 permutation with Holm correction for multiple testing (Holm, 1979).

Finally, we ran distance–based redundancy analysis (db–RDA) on each community dissimilarity matrix to examine community–environment relationships in more detail (Legendre and Anderson, 1999). This method builds on redundancy analysis, but can be based on any dissimilarity or distance matrix (Legendre and Legendre, 2012). To retain comparability among BIOENV, Mantel tests and db–RDA, we used the environmental variables selected by the BIOENV routine as the predictor variables in the db–RDAs. Using db–RDA,

we tested for the amount of variation explained (R^2), overall significance of the ordination solutions, and the marginal significance of each environmental variable included in the model.

All analyses were performed in R version 3.6.0 (R Development Core Team 2018).

Results

Morphological analysis

We identified by light microscopy a total of 40 genera and 98 species across all study samples. The number of species and genera per sample ranged between 5-21 and 4-17 respectively, with an average of 14 and 10 respectively. Moreover, 5 species were identified in most samples -*Achnanthidium minutissimum* (Kützing) Czarnecki, *Eunotia bilunaris* (Ehrenberg) Schaarschmidt, *Fragilaria tenera* (W.Smith) Lange-Bertalot, *Ulnaria acus* (Kützing) Aboal, and*Gomphonema exilissimum* (Grun.) Lange-Bertalot & Reichardt-.

HTS data and phylogenetic analysis

The Illumina Miseq sequencing run generated a total number of 4,150,073 reads for all samples. After applied quality filters, dereplication and chimera removal processes we obtained 2,595,039 DNA sequences. The clustering process of sequences at 97% identity level resulted in a total of 15,234 OTUs for all samples. After removing singletons, OTUs with nonsense codons and not belonging to Bacillariophyta phylum we obtained 7834 OTUs. Finally, after removing OTUs with a normalized read value less than 0,005% we obtained 4707 OTUs. The number of OTUs per sample ranged between 390 and 980, with an average of 634 OTUs per sample. Taxonomic assignation of OTUs was positive for 3138 OTUs, which were assigned to 219 species and 90 genera. The number of genera and species per sample ranged between 31-56 and 58-107 respectively, with an average of 42 and 80 respectively. Twenty-two taxa were present in all molecular-analyzed samples, of which three species (*Ulnaria acus* ,*Eunotia bilunaris* and *Achnanthidium minutissimum*) and two genera (*Gomphonema* sp. and *Fragilaria* sp.) were also present in the most morphologically-analyzed samples. A total of 1569 OTUs could not be assigned to R-Syst::diatom reference database and remained unclassified.

According to our phylogeny constructed with 708 reference sequences, we observed several sequences not placed correctly. This result was more noticeable in the phylogeny constructed with the same reference sequences and 3138 taxonomy-assigned OTUs, where some reference sequences were placed out of their corresponding taxonomy-assigned OTUs. Accession to *rbcL* sequences alignments and phylogenetic trees is detailed at Data accessibility section.

Patters and processes in the metacommunity structuring of benthic diatoms

The morphological analysis recovered 40 genera, in contrast to the 90 genera recovered with metabarcoding. We found 98 species based on morphological identification and 219 species based on metabarcoding. The comparison between morphological and molecular inventories through Venn diagrams showed considerable differences between both methods at genus and species level. We found 10 genera detected only by light microscopy (of which 8 of them had reference sequences in the database), 60 genera detected only by metabarcoding and 30 genera (30%) detected by both methods **(Figure 3 a).** At species level, we found 55 species detected only by light microscopy (of which 27 of them had reference sequences in the database), 176 species detected only by metabarcoding and 43 species (15.7%) detected by both methods **(Figure 3 b).**

Stress values in nMDS ordinations for morphological (species–level nMDS =0.22, genus–level nMDS =0.23) and molecular (species–level nMDS =0.24, genus–level nMDS =0.22) data suggested a reasonable fit (Clarke, 1993). The Procrustes rotation analysis using nMDS scores**(Figure 4)** showed that most of the study sites displayed a relatively low degree of similarity, indicating a poor correspondence in the compositional variation between morphology–and molecular–based metacommunities (here, PROTEST for species–level resolution data,m_{1-2}=0.29 and p=0.33; and PROTEST for genus–level resolution data, m_{1-2} =0.30 and a p=0.28). The relatively high procrustean residuals (red arrows in **Figure 4**) re–emphasised this mismatch between the ordinal results of the test datasets at different taxonomic resolutions. Importantly, the direction of the movement and the length of the arrows in the procrustean plots were associated with the distribution of both

morphology– and molecular–based assemblages. In this vein, the low correspondence found for both species– and genus–level data was partially caused by e.g. *Navicula soehrensis* , *Cyclostephanos invisitatus* , *Cymbella subhelvetica* , *Navicula notha* , *Tabellaria fenestrata* ,*Luticola goeppertiana* and *Amphora indistincta* , which were present exclusively in the morphology–based samples. Moreover, molecular–based assemblages included a number of taxa that were absent in morphological identifications such as, *Nitzschia fruticosa* ,*Eunotia arcus* , *Attheya septentrionalis* , *Haslea pseudostrearia* , *Lucanicum sp* ., *Leptocylindrus sp* . and*Pseudictyota* sp.

Species–level data based on morphological identifications showed the strongest environment relationship (p[?]0.01) of all different study approaches, and the BIOENV routine selected conductivity, fluorides, total phosphorus (TP) and total suspended solids (TSS) for the best environmental distance matrix. Ammonium and TSS related to species–level data based on DNA metabarcoding of the entire assemblage, whereas fluorides and TSS structured the genus–level data based on morphological and molecular approaches, respectively **(Table 3)** . According to the non–ranked Mantel tests and partial non–ranked Mantel tests, correlations between dissimilarity and distance matrices showed that only environmental distances were significantly correlated with biological –Jaccard– dissimilarities, whereas geographical distances were never significantly correlated with compositional variation in diatom metacommunities **(Table 3)** . Distance–based redundancy analysis (db–RDA) showed that half of the environmental variables selected by BIOENV were significantly related to variation in community composition, but that only a rather small amount of compositional variation (in terms of adjusted R^2values; here,R^2<0.3) could be explained by these environmental variables **(Table 4)** . Of the four dissimilarity matrices subjected to db–RDAs, species–level data based on morphological identifications were best explained by the environmental variables and the genus–level data based on DNA metabarcoding the worst.

We used Mantel correlograms to examine if there was significant spatial autocorrelation at any distance class. In this regard, we detected only very weak, but no significant spatial autocorrelation for morphological and molecular data at different taxonomic resolutions **(Figure 5)** . Perhaps more importantly, re–running the analyses with the spatial turnover component of the Jaccard index did not alter our main results

(Supplemental Information Appendix S1; Figure S1 and S2 and Table S1 and S2).

Discussion

Diatoms are important organisms to understand aquatic ecosystem functioning since they play an important position as producers (Rimet et al., 2018). Unfortunately, our knowledge of diatom biodiversity is still limited given the great number of estimated extant species (Mann and Vanormelingen, 2013). However, DNA metabarcoding provides a powerful tool to examine unknown diatom diversity and expand our knowledge about their distribution patterns.

Contrary to our expectations (**Hyphothesis 1**), we observed a relatively poor correspondence between both morphology-based and molecular-based approaches. We found, however, that both methods provided similar information when it comes to the underlying processes determining geographical variation in diatom communities, thereby supporting our second hypothesis (**Hypothesis 2**). There are several potential explanations for the relatively low congruence found between the two approaches:

1. Choice of DNA marker. The *rbc* L gene is a common marker used for metabarcoding and phylogenetic studies (Keck, Vasselon, Rimet, Bouchez, & Kahlert, 2018). This gene coding for a protein, so alignment is simple, insertions or deletions are extremely rare and compared with ribosomal or mitochondrial markers, the likelihood of amplifying non-specific products is reduced (Soltis and Soltis, 1998; Evans, Wortle, & Mann, 2007). Moreover, the *rbc* L gene seems to separate taxa better than 18S rDNA gene at species level (Kermarrec et al., 2014). However, *rbc* L marker gene does not work for species lacking a functional plastid (obligatory heterotrophs) such as*Nitzschia alba* (Kowalska et al., 2019). In addition, the short length 312-bp *rbc* L barcode gene is readily PCR amplifiable, which makes the analysis easier. However, the using of a short sequence (<500 bp) for barcoding may constrain the taxonomic and phylogenetic assignment, as the information content in the sequence is limited (Medlin, 2018; Tedersoo, Tooming-Klunderud & Anslan, 2018). In this regard, Keck et al. (2018) compared

the placement accuracy of the 312-bp gene fragment with the full-length of the *rbc* L gene in their phylogeny, and observed that approximately the 45% of the species were placed exactly at full-length gene. Similarly, our phylogeny constructed with 708 reference sequences and 3138 taxonomy-assigned OTUs shown that several reference sequences were placed far of their corresponding taxonomy-assigned OTUs. The correct placement of sequences on a phylogeny depends on several factors as the choice of marker gene, the length of amplicons and the presence of closely related taxa in the reference phylogeny (Keck et al. (2018). Tedersoo et al. (2018) have highlighted the importance of length sequence in metabarcoding, emphasizing that longer amplicons increase the accuracy of identification at the species level. We speculate that using a combination of two or more DNA barcode regions or others marker genes (e. g. Second Internal Transcriber Spacer) could be more suitable for unambiguous species identification, especially to distinguish closely related species (Moniz and Kaczmarska, 2009).

2. The PCR reaction used to amplify the barcode region can be inhibited by contaminants and produce chimeric DNA molecules (Hugerth and Andersson, 2017). Moreover, several organisms may be underestimated if their DNA template does not hybridize with the designed primers.
3. On the other hand, the completeness of the reference database is a key factor that strongly limits the taxonomy assignment of OTUs. In this vein, a large number of diatom taxa morphologically identified (31 species and 8 genera) could not be detected by metabarcoding approach due to the lack of reference sequences in the R-Syst::diatom database. Thus, some species detected only by light microscopy (e. g.*Cocconeis euglypta, C. pediculus, Stauroneis producta and S. gracilis)* differed from those detected by metabarcoding approach (*C. cupulifera, C. mascarenica, S. anceps and S. gracilior*). Following Jahn, Zetzsche, Reinhardt, & Gemeinholzer (2007), we further hypothesize that taxa with sequences absent in the reference database could be compensated by taxa of the same genus that have sequences available in the reference database or by a taxon not expected in the studied ponds. This hypothesis could explain the relatively minor discrepancies observed between both inventories at genus level resolution.
4. The bioinformatics processing might have also played an important role in the discrepancies observed between morphological and molecular inventories. Typically, DNA sequences obtained in a high-throughput sequencing run are filtered and clustered, based on a distance matrix at a specified threshold, into Operational Taxonomic Units (OTUs) to reduce the PCR and sequencing errors and the polymorphism present in the barcode region (Chen, Zhang, Cheng, Zhang, & Zhao, 2013). The clustering process is mainly affected by the clustering method and the threshold value used for sequence similarity (Chen et al., 2013). Often, sequences are clustered at 97% similarity, however different taxa could have less distance between their barcodes (Hugerth and Andersson, 2017). By contrast, using of high sequence similarity threshold value increase the number of unclassified OTUs and the PCR and sequencing errors (Tapolczai et al., 2018). Nevertheless, a common identity threshold for assigns taxonomy to all diatom taxa does not appear exist yet due to the heterogenous evolution rate of the*rbc* L gene and the speciation process (Kermarrec et al., 2014). In addition, relation between OTUs and biological species is not straightforward (Ryberg, 2015; Balint et al., 2016).

Interestingly, we observed the presence of some marine species in our molecular inventory, e.g. *Thalassiosira profunda* and*Thalassiosira mediterranea* (Percopo, Siano, Cerino, Sarno, & Zinigone, 2011; Hasle 1990). The sequences assigned to such species were placed far of their respective reference sequences in our phylogeny, which could reflect an inaccurate taxonomic assignment. However, the taxonomic assignment at genus level of such sequences could be correct since thalassiosiroids feature prominently in freshwater ecosystems, rivaling their freshwater diversity with the marine ones (Alverson, 2014). On the other side, microscopy method has a lower capacity to detect rare species than metabarcoding (Rimet et al., 2018), whereas that molecular-based approach allows detecting all species that could be detected by this method, covering the full range of species richness. However, we hypothesize that using a higher similarity threshold value for taxonomic assignment or using simultaneously other marker genes could be more suitable to assign unequivocally taxonomy to such DNA sequences.

1. On the other hand, several species (cryptic) may be morphologically identical but have genetic differences (Zimmermann et al., 2015). Several molecular studies (Mann and Vanormelingen, 2013; An,

Choi, Lee, Lee, & Noh, 2018) have suggested that diatom biodiversity has been underestimated. For example, in our study we identified morphologically only 12 infrageneric taxa belonging to*Nitzschia* genus, whereas by metabarcoding approach were detected 24 taxa. This fact could be related with the cryptic diversity observed within the morphologically identified*Nitzschia palea* species complex (Trobajo et al., 2010). Likewise, genetically distinct entities have been observed within morphologically identified species in *Cyclotell* a,*Eunotia* , *Gomphonema* , *Hantzschia* , *Navicula* ,*Pinnularia* and *Sellaphora* (Rovira et al., 2015). On the other side, the intraspecific and intragenomic polymorphism present in the barcode region can overestimate the species richness, since members of a single taxon possess several genotypes at the barcode region and may clustered into different OTUs (Mora et al., 2019). In addition, individuals of the same species from different geographic populations may possess different barcode sequences (Medlin, 2018).

2. Other factors, as the presence of extracellular DNA, can affect the composition of molecular inventories. Thus, extracellular DNA from diatom species may be detected in a sample even if their cells are not physically present, adding extra taxa to the molecular inventory (Kermarrec et al., 2014; Rimet et al., 2018). Moreover, our morphological identifications were based on the observation of live material only. Thereby, some taxa founded in our molecular inventories (e.g. *Attheya septentrionalis)* may hardly be identified by microscopic methods since they are weakly silicified (Stachura-Suchoples, Enke, Schlie, Schaub, Karsten & Jahn, 2015). Finally, the high number of synonyms present on diatoms taxonomy may hinder the comparison of morphological and molecular inventories (Hillebrand, Watermann, Karez & Berninger, 2001).

In spite of all biases inherent to both morphological and metabarcoding methods, compositional variation of diatom communities was positively correlated with the environmental template, thereby emphasizing that diatom communities were mainly controlled by niche-based mechanisms (e.g. species sorting) and confirming our second hypothesis (**Hypothesis 2**). Similar results have been reported by other studies on diatom communities (Verleyen et al., 2009; Gothe et al., 2013; Jamoneau, Passy, Soininem, Leboucher, & Tison-Rosebery, 2017), in which the environmental factors dominated the spatial and biological processes on structuring benthic algal communities. Moreover, in our study similar environmental variables (e.g. total suspended solids) were correlated in both inventories with diatom composition variation, which could be related with the same sampled substrate (*S. lacustris*), since diatom species may exhibit a tight environmental tolerance and strong preferences for particular substrata (Soininen, 2007; Cantonati & Spitale, 2009). Host macrophytes are important elements supplying nutrients to epiphytic diatoms, especially in oligotrophic and mesotrophic waters (Letakova, Frankova, & Pouličková, 2018). In our study, morphological and molecular inventories were related with nutrients (e.g. total phosphorus and ammonium), which is expectable since nutrients (particularly phosphorus) are important for diatoms primary productivity and growth (Pan, Stevenson, Hill, Herlihy, & Collins, 1996). Ammonia influence importantly the diatom composition and may be a limiting nutrient in primary productivity (Natarajan, 1970). Similarly, fluoride can improve or inhibit the growth of diatoms depending of its concentration, exposure time and diatom species (Camargo, 2003). Moreover, conductivity was related with morphological inventory at species level resolution, which is foreseeable since diatoms are very sensitive to ionic content and composition, and consequently, they are often used to monitor conductivity fluctuations (Potapova and Charles, 2003). Finally, both morphological and molecular inventories were related with total suspended solids variable, which may influence diatom assemblages by processes as light decreasing, nutrient adsorption and algae aggregation (Hoshikawa et al., 2019).

Microorganisms, and particularly diatoms, have historically been considered to be ubiquitously distributed due their small size and huge population densities, and their communities mainly controlled by local environmental factors (Soininen, 2007; Hillebrand et al., 2001). Nevertheless, this distribution pattern has been challenged by several studies (Heino et al., 2010; Soininen, 2007; Blanco, Olenici, & Ortega, 2020), suggesting that variation in community structure cannot be explained by environmental factors alone, and thereby questioning the strict ubiquitous dispersal of diatom communities. We found no significant correlation between compositional variation of diatom assemblages and spatial distance, which may be explained by the relatively small extent of our study area. The effect of spatial distance may be more important at

large spatial extents, while environmental factors may be more important at reduced extents (Alahuhta & Heino, 2013; Declerk et al., 2011). However, in stochastic and highly heterogeneous systems as temporary ponds, environmental control may not necessarily be strong (Heino et al., 2015). Hence, other factors not assessed in our study, such as biotic interactions, may also be important to structure diatom communities (Göthe et al., 2013). Nevertheless, we are confident that we included an environmental template frequently known to influence the composition of diatom communities (Pan et al., 1996; Potapova and Charles, 2003). Moreover, the environmental template we included varied extensively across ponds, thereby leading potential for species sorting.

In summary, our study showed that both molecular and morphological methods were influenced by several biases inherent in its own methodology. The main biases related to molecular approach were probably the incompleteness of the reference database and the bioinformatics processing, which highlight the need of expand the reference database to include all genotypes of occurring taxa and the need of reach a consensus about the bioinformatics processing in order to favor the comparison between studies. In addition, establishing robust species identification thresholds and using a combination of two or more DNA barcode regions could be suitable for unambiguous species identification, especially in those cases where a single marker gene shows low variability. On the other side, the limited counting effort of morphological approach and the presence of cryptic species were presumably the main biases related with the morphological approach. Our results showed that both approaches were related with the environmental template, suggesting that Mediterranean epiphytic diatom communities are mainly controlled by niche-based mechanisms at regional extents. However, we have not found a significant correlation between compositional variation of diatom assemblages and spatial distance, probably explained by the regional spatial extent studied. In conclusion, our work shows that both molecular and morphological approaches provide complementary information on each other and highlighted the importance of metabarcoding approach to infer the composition of epiphytic diatom assemblages, especially when completeness of the reference databases improves and bioinformatics biases are overcome.

Acknowledgments

This study was supported by the project METAPONDS (CGL2017-84176R), grant by the Spanish Ministry of Economy and Industry and by the project BT-2019, grant by the Biodiversity Foundation and the Spanish Ministry for Ecological Transition and Demographic Challenge. We also appreciate the support of the Supercomputing Centre of Castilla y León (SCAYLE).

We would like to express our gratitude to the professor Dr. Luís Enrique Sáenz de Miera Carnicer for bioinformatics advising and the Limnology and environmental biotechnology group (University of León) for sample collection.

This paper is dedicated to the memory of the professor Dra. Margarita Fernández Aláez, who recently passed away.

References

Alahuhta, J. & Heino, J. (2013). Spatial extent, regional specificity and metacommunity structuring in lake macrophytes. *Journal of Biogeography* , 40(8), 1572-1582. doi: Alahuhta, J., & Heino, J. (2013). Spatial extent, regional specificity and metacommunity structuring in lake macrophytes. *Journal of Biogeography* , 40(8), 1572–1582. doi: 10.1111/jbi.12089

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology* , 215(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2

Alverson, A. J. (2014). Timing marine–freshwater transitions in the diatom order Thalassiosirales. *Paleobiology* , *40* (1), 91–101. doi: 10.1666/12055

An, S. M., Choi, D. H., Lee, H., Lee, J. H., & Noh, J. H. (2018). Next-generation sequencing reveals the diversity of benthic diatoms in tidal flats. *Algae* , *33* (2), 167–180. doi: 10.4490/algae.2018.33.4.3

Andrews, S. (2010). FASTQC: A quality control tool for high throughput sequence data. Available online at:*http://www.bioinformatics.babraham.ac.uk/projects/fastqc*

Arnot, D. E., Roper, C., & Bayoumi, R. A. L. (1993). Digital codes from hypervariable tandemly repeated DNA sequences in the Plasmodium falciparum circumsporozoite gene can genetically barcode isolates.*Molecular and Biochemical Parasitology* , *61* (1), 15–24. doi: 10.1016/0166-6851(93)90154-P

Bálint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., ... Tedersoo, L. (2016). Millions of reads, thousands of taxa: Microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews* , *40* (5), 686–700. doi: 10.1093/femsre/fuw017

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography* , *19* (1), 134–143. doi: 10.1111/j.1466-8238.2009.00490.x

Baselga, A., Orme, O., Villeger, S., De Bortoli, J., Leprieur, F., Logez, M., & Henriques-Silva, R. (2018). Betapart: partitioning beta diversity into turnover and nestedness components. R package version 1.5.1. Available online at:*https://cran.r-project.org/web/packages/betapart/index.html*

Bishop, T. R., Robertson, M. P., van Rensburg, B. J., & Parr, C. L. (2015). Contrasting species and functional beta diversity in montane ant assemblages. *Journal of Biogeography* , *42* (9), 1776–1786. doi: 10.1111/jbi.12537

Blanco, S. (2020). Diatom taxonomy and identification keys. In G. Cristobal, S. Blanco & G. Bueno (Eds.), *Modern trends in diatom identification: Fundamentals and Applications* (pp. 25-38). Cham, Switzerland: Springer Nature Switzerland AG.

Blanco, S., Ector, L., & Bécares, E. (2004). Epiphytic diatoms as water quality indicators in Spanish shallow lakes. *Vie et Milieu* ,*54* (2–3), 71–79

Blanco, S., Olenici, A., Ortega, F. (2020). Identifying environmental drivers of benthic diatom diversity: the case of Mediterranean mountain ponds. *PeerJ* , 8, e8825. doi.org: 10.7717/peerj.8825

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., ... Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.*Nature methods* , 10(1), 57-59. doi: 10.1038/nmeth.2276

Bonecker, C. C., Simões, N. R., Minte-Vera, C. V., Lansac-Tôha, F. A., Velho, L. F. M., & Agostinho, Â. A. (2013). Temporal changes in zooplankton species diversity in response to environmental changes in an alluvial valley. *Limnologica* , *43* (2), 114–121. doi: 10.1016/j.limno.2012.07.007

Borrego-Ramos, M., Olenici, A., & Blanco, S. (2019). Are dead stems suitable substrata for diatom-based monitoring in mediterranean shallow ponds? *Fundamental and Applied Limnology* , *192* (3), 215–224. doi: 10.1127/fal/2019/1163

Brown, B. L., Sokol, E. R., Skelton, J., & Tornwall, B. (2016). Making sense of metacommunities: dispelling the mythology of a metacommunity typology. *Oecologia* , 183, 643-652. doi: 10.1007/s00442-016-3792-1

Camargo, J. A. (2003). Fluoride toxicity to aquatic organisms: A review.*Chemosphere* , *50* (3), 251–264. doi: 10.1016/S0045-6535(02)00498-8

Cantonati, M. & Spitale, D. (2009). The role of environmental variables in structuring epiphytic and epilithic diatom assemblages in springs and streams of the Dolomiti Bellunesi National Park (south-eastern Alps).*Fundamental and Applied Limnology / Archiv für Hydrobiologie* , 174(2), 117-133. doi: 10.1127/1863-9135/2009/0174-0117

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., & Zhao, H. (2013). A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs.*PLoS ONE* , *8* (8). doi: 10.1371/journal.pone.0070837

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* , *18* (1), 117–143. doi: 10.1111/j.1442-9993.1993.tb00438.x

Clarke, K. R., & Ainsworth, M. (1993). A method of linking multivariate community structure to environmental variables. *Marine Ecology Progress Series* , *92* (3), 205–219. doi: 10.3354/meps092205

Declerck, S. A. J., Coronel, J. S., Legendre, P., & Brendonck, L. (2011). Scale dependency of processes structuring metacommunities of cladocerans in temporary pools of High-Andes wetlands. *Ecography* ,*34* (2), 296–305. doi: 10.1111/j.1600-0587.2010.06462.x

European Standard. (2003). Water quality-Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers (EN 13946:2003). Retrieved from: https://standards.iteh.ai/catalog/standards/cen/8e62f4b7-732a-4197-9efa-681b339b3740/en-13946-2003

Evans, K. M., Wortley, A. H., & Mann, D. G. (2007). An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in Sellaphora (Bacillariophyta). *Protist* , *158* (3), 349–364. doi: 10.1016/j.protis.2007.04.001

Fišer Pečnikar, Ž., & Buzan, E. V. (2013). 20 years since the introduction of DNA barcoding: from theory to application. *Journal of applied genetics* , 55(1), 43–52. doi.org/10.1007/s13353-013-0180-y

Florencio, M., Díaz-Paniagua, C., Gómez-Rodríguez, C., & Serrano, L. (2014). Biodiversity patterns in a macroinvertebrate community of a temporary pond network. *Insect Conservation and Diversity* ,*7* (1), 4–21. doi: 10.1111/icad.12029

García-Girón, J., Fernández-Aláez, C., Fernández-Aláez, M., & Luis, B. (2018a). Subfossil Cladocera from surface sediment reflect contemporary assemblages and their environmental controls in Iberian flatland ponds.*Ecological Indicators* , *87* (December 2017), 33–42. doi: 10.1016/j.ecolind.2017.12.007

García-Girón, J., Fernández-Aláez, M., & Fernández-Aláez, C. (2018b). Relationships between contemporary and subfossil macrophyte assemblages in Mediterranean ponds. *Marine and Freshwater Research* ,*69* (9), 1408–1417. doi: 10.1071/MF18023

Göthe, E., Angeler, D. G., Gottschalk, S., Löfgren, S., & Sandin, L. (2013). The influence of environmental, biotic and spatial factors on diatom metacommunity structure in swedish headwater streams. *PLoS ONE* , *8* (8). doi: 10.1371/journal.pone.0072237

Gower, J. C. (1971). Statistical methods of comparing different multivariate analyses of the same data. In F. R. Hodson, D. G. Kendall & P. Tăutu (Eds.), *Mathematics in the Archaeological and Historical Sciences* (pp. 138-149). Edinburgh, UK: Edinburgh University Press.

Green, J. L., Holmes, A. J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M., ... Beattie, A. J. (2004). Spatial scaling of microbial eukaryote diversity. *Natur* e, 432, 747-750. doi: 10.1038/nature03034

Hadi, S. I. I. A., Santana, H., Brunale, P. P. M., Gomes, T. G., Oliveira, M. D., Matthiensen, A., ... Brasil, B. S. A. F. (2016). DNA barcoding green microalgae isolated from neotropical inland waters.*PLoS ONE* , *11* (2), 1–18. doi: 10.1371/journal.pone.0149284

Hasle, G. R. (1990). The planktonic marine diatom Thalassiosira mediterranea (synonym Thalassiosira stellaris). *Diatom Research* ,*5* (2), 415–418. doi: 10.1080/0269249X.1990.9705132

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* , *270* (1512), 313–321. doi: 10.1098/rspb.2002.2218

Heino, J., Bini, L. M., Karjalainen, S. M., Mykrä, H., Soininen, J., Vieira, L. C. G., & Diniz-Filho, J. A. F. (2010). Geographical patterns of micro-organismal community structure: are diatoms ubiquitously distributed across boreal streams? *Oikos* , 119(1), 129-137. doi: 10.1111/j.1600-0706.2009.17778.x

Heino, J., Melo, A. S., Siqueira, T., Soininen, J., Valanko, S., & Bini, L. M. (2015). Metacommunity organisation, spatial extent and dispersal in aquatic systems: Patterns, processes and prospects. *Freshwater Biology* , *60* (5), 845–869. doi: 10.1111/fwb.12533

Hillebrand, H., Watermann, F., Karez, R., & Berninger, U.G. (2001). Differences in species richness patterns between unicellular and multicellular organisms. *Oecologia* , 126(1), 114-124. doi: 10.1007/s004420000492

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* , *6* (2), 65–70. doi: 10.2307/4615733

Hoshikawa, K., Fujihara, Y., Siev, S., Arai, S., Nakamura, T., Fujii, H., ... Yoshimura, C. (2019). Characterization of total suspended solid dynamics in a large shallow lake using long-term daily satellite images. *Hydrological Processes* , *33* (21), 2745–2758. doi: 10.1002/hyp.13525

Hugerth, L. W., & Andersson, A. F. (2017). Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology* , *8* , 1–22. doi: 10.3389/fmicb.2017.01561

Jackson, D. A. (1995). PROTEST: A PROcrustacean Randomization TEST of community environment concordance. *Ecoscience* , *2* (3), 297–303. doi: 10.1080/11956860.1995.11682297

Jahn, R., Zetzsche, H., Reinhardt, R., & Gemeinholzer, B. (2007). Diatoms barcoding: A pilot study on an environmental sample. *Proceedings of the 1st Central European Diatom Meeting* . Berlin. Retrieved from https://www.researchgate.net/publication/269164895_Diatoms_and_DNA_barcoding_a_pilot_-study_on_an_environmental_sample

Jamoneau, A., Passy, S. I., Soininen, J., Leboucher, T., & Tison-Rosebery, J. (2017). Beta diversity of diatom species and ecological guilds: Response to environmental and spatial mechanisms along the stream watercourse. *Freshwater Biology* , *63* (1), 62–73. doi: 10.1111/fwb.12980

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* , *30* (4), 772–780. doi: 10.1093/molbev/mst010

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

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J. M., Humbert, J. F., & Bouchez, A. (2014). A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science* , *33* (1), 349–363. doi: 10.1086/675079

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J. F., & Bouchez, A. (2013). Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: A test for freshwater diatoms. *Molecular Ecology Resources* , *13* (4), 607–619. doi: 10.1111/1755-0998.12105

Kowalska, Z., Pniewski, F., & Latała, A. (2019). DNA barcoding – A new device in phycologist's toolbox. *Ecohydrology and Hydrobiology* ,*19* (3), 417–427. doi: 10.1016/j.ecohyd.2019.01.002

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

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

Krammer, K. & Lange-Bertalot, H. (1991a). Bacillariophyceae 3. Teil: Centrales, Fragilariaceae, Eunoticeae. In: Ettl, H., Gerloff, J., Heynig, H., Mollenhauer, D. (Eds.), *Sübwasserflora von Mitteleuropa* 2/3 (pp. 1–576). Stuttgart, Germany: Gustav Fischer Verlag.

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

Lange-Bertalot, H., Hofmann, G., Werum, M., & Cantonati, M. (2017).*Freshwater benthic diatoms of central Europe: over 800 common species used in ecological assessment. English edition with updates taxonomy* and *added species* . Schmitten-Oberreifenberg, Germany. Koltz Botanical Books.

Leboucher, T., Budnick, W. R., Passy, S. I., Boutry, S., Jamoneau, A. Soininen, J., ... Tison-Rosebery, J. (2019). Diatom β-diversity in streams increases with spatial scale and decreases with nutrient enrichment across regional to sub-continental scales. *Journal of Biogeography* , 46(4), 734-744. doi: 10.1111/jbi.13517

Legendre, P., & Anderson, M.J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* , 69(1), 1-24. doi:10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2

Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (3rd ed.). Oxford, UK: Elsevier.

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., ... Gonzalez, A. (2004). The metacommunity concept: A framework for multi-scale community ecology. *Ecology Letters* ,*7* (7), 601–613. doi: 10.1111/j.1461-0248.2004.00608.x

Letáková, M., Fránková, M., & Poulíčková, A. (2018). Ecology and Applications of Freshwater Epiphytic Diatoms — Review.*Cryptogamie, Algologie* , *39* (1), 3–22. doi: 10.7872/crya/v39.iss1.2018.3

Mann, D. G., & Vanormelingen, P. (2013). An inordinate fondness? the number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology* , *60* (4), 414–420. doi: 10.1111/jeu.12047

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* , 27(2), 209-220.

Medlin, L. K. (2018). Mini review: Diatom species as seen through a molecular window. *Revista Brasileira de Botanica* , *41* (2), 457–469. doi: 10.1007/s40415-018-0444-1

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In*Gateway Computing Environments Workshop (GCE)* , pp. 1-8.

Moniz, M. B. J., & Kaczmarska, I. (2009). Barcoding diatoms: Is there a good marker? *Molecular Ecology Resources* , *9* (SUPPL. 1), 65–74. doi: 10.1111/j.1755-0998.2009.02633.x

Mora, D., Abarca, N., Proft, S., Grau, J. H., Enke, N., Carmona, J., ... Zimmermann, J. (2019). Morphology and metabarcoding: A test with stream diatoms from Mexico highlights the complementarity of identification methods. *Freshwater Science* , *38* (3), 448–464. doi: 10.1086/704827

Natarajan, K. V. (1970). Toxicity of ammonia to marine diatoms.*Journal of the water pollution control federation* , 42(5), 184-190

[dataset] Nistal-García, A., García-García, P., García-Girón, J., Borrego-Ramos, M., Blanco, S., & Bécares, E.; 2020; DNA metabarcoding and morphological methods show complementary patterns in the metacommunity organization of lentic epiphytic diatom; Figshare; https://doi.org/10.6084/m9.figshare.12933017

Oden, N. L., & Sokal, R. R. (1986). Directional autocorrelation: An extension of spatial correlograms to two dimensions. *Systematic Zoology* , *35* (4), 608–617. doi: 10.2307/2413120

Ohlmann, M., Mazel, F., Chalmandrier, L., Bec, S., Coissac, E., Gielly, L., ... Thuiller, W. (2018). Mapping the imprint of biotic interactions on β-diversity. *Ecology Letters* , *21* (11), 1660–1669. doi: 10.1111/ele.13143

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2019). Vegan: Community Ecology Package. R package ver. 2.5.6. Available online at:*https://cran.rproject.org/web/packages/vegan/index.html*

Oliveros, J. C. (2015). Venny. An interactive tool for comparing lists with Venn's diagrams. Available online at:*https://bioinfogp.cnb.csic.es/tools/venny/index.html*

Pan, Y., Stevenson, R. J., Hill, B. H., Herlihy, A. T., & Collins, G. B. (1996). Using diatoms as indicators of ecological conditions in lotic systems: A regional assessment. *Journal of the North American Benthological Society* , *15* (4), 481–495. doi: 10.2307/1467800

Percopo, I., Siano, R., Cerino, F., Sarno, D., & Zingone, A. (2011). Phytoplankton diversity during the spring bloom in the northwestern Mediterranean Sea. *Botanica Marina* , *54* (3), 243–267. doi: 10.1515/BOT.2011.033

Potapova, M., & Charles, D. F. (2003). Distribution of benthic diatoms in U.S. rivers in relation to conductivity and ionic composition.*Freshwater Biology* , *48* (8), 1311–1328. doi: 10.1046/j.1365-2427.2003.01080.x

R Development Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing,

Rambaut, A. (2016). FigTree. Version 1.4.3. Available at:*http://tree.bio.ed.ac.uk/software/figtree/.*

Riato, L., Leira, M., Della Bella, V., & Oberholster, P. J. (2018). Development of a diatom-based multimetric index for acid mine drainage impacted depressional wetlands. *Science of the Total Environment* ,*612* , 214–222. doi: 10.1016/j.scitotenv.2017.08.181

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* ,*16* (6), 276–277. doi: 10.1016/S0168-9525(00)02024-2

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

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. *Organisms Diversity and Evolution* , *18* (1), 51–62. doi: 10.1007/s13127-018-0359-5

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., & Rimet, F. (2017). Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia* ,*807* (1), 37–51. doi: 10.1007/s10750-017-3381-2

Rodriguez-Alcalá, O., Blanco, S., García-Girón, J., Jeppesen, E., Irvine, K., Nõges, T., ... Bécares, E. (2020). Large-scale geographical and environmental drivers of shalllow lake diatom metacommunities across Europe. *Science of the Total Environment,*707. doi: 10.1016/j.scitotenv.2019.135887

Rognes T, Flouri T, Nichols B, Quince C., & Mahé F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* , 4, e2584. doi: 10.7717/peerj.2584

Round, F. E., Crawford, R. M., & Mann, D. G. (1990). *The diatoms: Biology & Morphology of the genera* . Cambridge: Cambridge University Press.

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. doi: 10.1111/jeu.12240

Ryberg, M. (2015). Molecular operational taxonomic units as approximations of species in the light of evolutionary models and empirical data from Fungi. *Molecular Ecology* , *24* (23), 5770–5777. doi: 10.1111/mec.13444

Smucker, N. & Vis, M. L. (2011). Spatial factors contribute to benthic diatom structure in streams across spatial scales: Considerations for biomonitoring. *Ecological Indicators* , 11(5), 1191-1203. doi: 10.1016/j.ecolind.2010.12.022

Soininen, J. (2007). Environmental and spatial control of freshwater diatoms—a review. *Diatom Research* , *22* (2), 473–490. doi: 10.1080/0269249X.2007.9705724

Soltis, D. E., & Soltis, P. S. (1998). Choosing an Approach and an Appropriate Gene for Phylogenetic Analysis. In D. E. Soltis, P. S. Soltis, & J. J. Doyle (Eds.), *Molecular systematics of plants II: DNA sequencing* (pp. 7-8). New York, USA: Kluwer Academic Publishers.

Stachura-Suchoples, K., Enke, N., Schlie, C., Schaub, I., Karsten, U., & Jahn, R. (2015). Contribution towards a morphological and molecular taxonomic reference library of benthic marine diatoms from two Artic fjords on Svalbard (Norway). *Polar Biology* , 39, 1933–1956. doi: 10.1007/s00300-015-1683-2

Tapolczai, K., Vasselon, V., Bouchez, A., Stenger-Kovács, C., Padisák, J., & Rimet, F. (2018). The impact of OTU sequence similarity threshold on diatom-based bioassessment: A case study of the rivers of Mayotte (France, Indian Ocean). *Ecology and Evolution* , *9* (1), 166–179. doi: 10.1002/ece3.4701

Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2018). PacBio metabarcoding of Fungi and others eukaryotes: errors, biases and perspectives. *New Phytologist* , 217(3), 1370-1385. doi: 10.1111/nph.14776

Trobajo, R., Mann, D. G., Clavero, E., Evans, K. M., Vanormelingen, P., & McGregor, R. C. (2010). The use of partial cox1, rbcL and LSU rDNA sequences for phylogenetics and species identification within the Nitzschia palea species complex (Bacillariophyceae). *European Journal of Phycology* , *45* (4), 413–425. doi: 10.1080/09670262.2010.498586

Vasiljević, B., Krizmanić, J., Ilić, M., Marković, V., Tomović, J., Zorić, K., & Paunović, M. (2014). Water quality assessment based on diatom indices – small hilly streams case study. *Water Research and Management* , 4(2), 31-25.

Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., & Bouchez, A. (2017). Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? *Freshwater Science* , *36* (1), 162–177. doi: 10.1086/690649

Verleyen, E., Vyverman, W., Sterken, M., Hodgson, D. A., De Wever, A., Juggins, S., ... Sabbe, K. (2009). The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. *Oikos* , *118* (8), 1239–1249. doi: 10.1111/j.1600-0706.2009.17575.x

Visco, J. A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., & Pawlowski, J. (2015). Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data.*Environmental Science and Technology* , *49* (13), 7597–7605. doi: 10.1021/es506158m

Wilhelm, C., Büchel, C., Fisahn, J., Goss, R., Jakob, T., LaRoche, J., ... Kroth, P. G. (2006). The regulation of carbon and nutrient assimilation in diatoms is significantly different from green algae.*Protist* , *157* (2), 91–124. doi: 10.1016/j.protis.2006.02.003

Zimba, P. V., & Hopson, M. S. (1997). Quantification of epiphyte removal efficiency from submersed aquatic plants. *Aquatic Botany* ,*58* (2), 173–179. doi: 10.1016/S0304-3770(97)00002-8

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., & Gemeinholzer, B. (2015). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources* ,*15* (3), 526–542. doi: 10.1111/1755-0998.12336

Zimmermann, J., Jahn, R., & Gemeinholzer, B. (2011). Barcoding diatoms: Evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organisms Diversity and Evolution* ,*11* (3), 173–192. doi: 10.1007/s13127-011-0050-6

Data accessibility

The data that support the findings of this study are openly available from online repositories. All Illumina MiSeq raw reads are available on the NCBI Sequence Read Archive with the accession numbers SAMN14535185-SAMN14535228, under the Bioproject number PRJNA623014. All rbcL sequence alignments and tree files

are found on Figshare (https://doi.org/10.6084/m9.figshare.12933017). Data are under embargo until publication and any further data required are available from the corresponding author upon reasonable request.

Author contributions

The study was conceived by E. B. and S. B. Experiments and morphological identifications were performed by A. N. G., P. G. G. and M. B. R. Data analysis and its interpretation were performed by J. G. G. and P. G. G. The manuscript was written by A. N. G. with significant contributions from all authors. All authors reviewed the manuscript and gave final approval for publication.

Hosted file

`Table_1.docx` available at https://authorea.com/users/359712/articles/481547-dna-metabarcoding-and-morphological-methods-show-complementary-patterns-in-the-metacommunity-organization-of-lentic-epiphytic-diatoms

Hosted file

`Table_2.docx` available at https://authorea.com/users/359712/articles/481547-dna-metabarcoding-and-morphological-methods-show-complementary-patterns-in-the-metacommunity-organization-of-lentic-epiphytic-diatoms

Hosted file

`Table_3.docx` available at https://authorea.com/users/359712/articles/481547-dna-metabarcoding-and-morphological-methods-show-complementary-patterns-in-the-metacommunity-organization-of-lentic-epiphytic-diatoms

Hosted file

`Table_4.docx` available at https://authorea.com/users/359712/articles/481547-dna-metabarcoding-and-morphological-methods-show-complementary-patterns-in-the-metacommunity-organization-of-lentic-epiphytic-diatoms

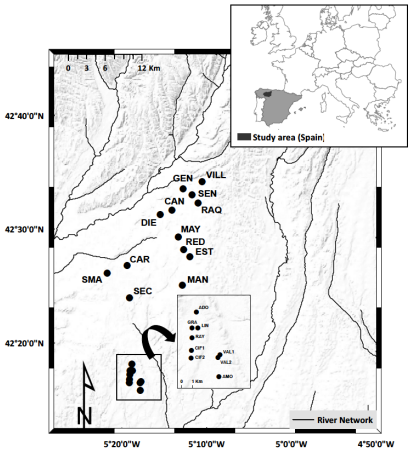

Figure 1. Location of the sampling area (black area in the map of Spain) and location of the 22 ponds within the sampling area.

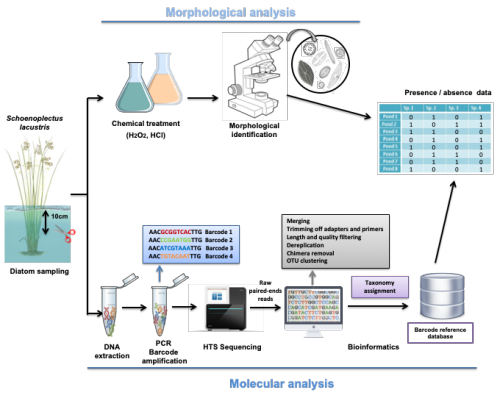

Figure 2. Graphical scheme of the diatom sampling procedure and the workflow illustrating the morphological and molecular methods used in this paper.

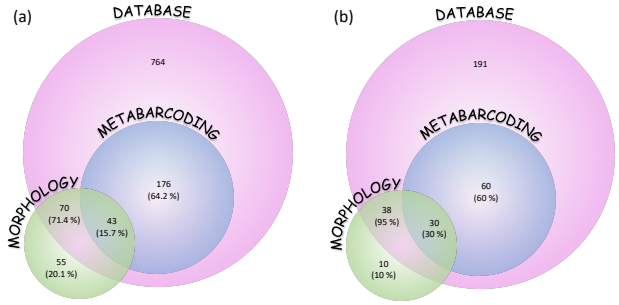

Figure 3. Venn diagrams showing the number of species **(a)** and genera **(b)** represented in the reference database (pink), detected by morphological approach (green) and by metabarcoding method (blue). Values in percentage are represented in brackets.

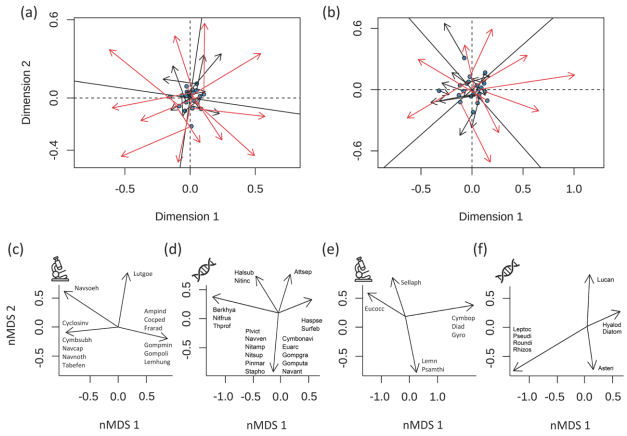

Figure 4. Procrustean superimposition plots generated from the ordinal results of morphology– and molecular–based nMDS for **(a)** species– and **(b)** genus–level data. Analyses were run for each approach based on the Jaccard coefficient. Blue circles represent scores from molecular–based samples and the end of the arrows the morphology–based assemblages. The distance between the two is the Procrustean residual. Samples with residual values greater than ca. the 50th percentile are marked with red arrows. **(c, d, e, f)** nMDS plots of epiphytic diatom compositional variation for **(c, e)** morphology– and **(d, f)** molecular–based data at different taxonomic resolutions (here, **(c, d)** species– and **(e, f)** genus–level resolutions). Inset are only the taxa with the highest axes (nMDS 1, nMDS 2) scores (here, score values greater than ca. the 90th percentile). Ampind, *Amphora indistincta*; Asteri, *Asterionella*; Attsep, *Attheya septentrionalis*; Berkhya, *Berkeleya hyalina*; Cocped, *Cocconeis pediculus*; Cyclosinv, *Cyclostephanos invisitatus*; Cymbonavi, *Cymbopleura naviculiformis*; Cymbop, *Cymbopleura*; Cymbsubh, *Cymbella subhelvetica*; Diad, *Diadesmis* Diatom, *Diatoma*; Euarc, *Eunotia arcus*; Eucocc, *Eucocconeis*; Frarad, *Fragilaria radians*; Gompgra, *Gomphonema gracile*; Gompmin, *Gomphonema minusculum*; Gompoli, *Gomphonella olivacea*; Gyro, *Gyrosigma* Halsub, *Halamphora subtropica*; Haspse, *Haslea pseudostrearia*; Hyalod, *Hyalodiscus*; Lemhung, *Lemnicola hungarica*; Lemn, *Lemnicola*; Leptoc, Leptocylindrus; Lucan, *Lucanicum*; Lutgoe, *Luticola goeppertiana*; Navant, *Navicula antonii*; Navcap, *Navicula capitatoradiata*; Navnoth, *Navicula nota*; Navsoeh, *Navicula soehrensis*; Navven, *Navicula veneta*; Nitamp, *Nitzschia amphibia*; Nitfrus, *Nitzschia fruticosa*; Nitinc, *Nitzschia inconspicua*; Nitsup, *Nitzschia supralitorea*; Pinacr, *Pinnularia acrosphaeria*; Pinmar, *Pinnularia marchica*; Plvic, *Planothidium victori*; Psamthi, *Psammothidium*; Pseudi, *Pseudictyota*; Rhizos, *Rhizosolenia*; Roundi, *Roundia*; Sellaph, *Sellaphora*; Stapho, *Stauroneis phoenicenteron*; Surfeb, *Surirella febigeri*; Tabefen, *Tabellaria fenestrata;* Thprof, *Thalassiosira profunda*.

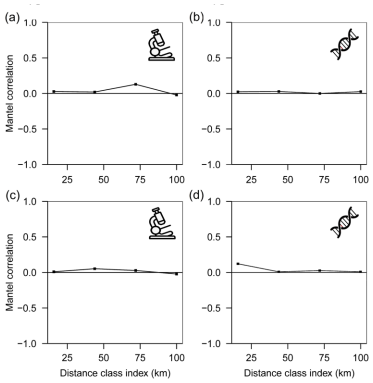

Figure 5. Mantel correlograms showing the spatial structures in community differences between sites. No significant spatial autocorrelation was found after Holm correction for multiple testing. Separate correlograms were drawn for **(a, c)** morphology– and **(b, d)** molecular–based data sets at different taxonomic resolutions (here, **(a, b)** species– and **(c, d)** genus–level resolutions).